

Shape-based Scenario Generation using Copulas

Michal Kaut*
michal.kaut@himolde.no

Stein W. Wallace*
stein.w.wallace@himolde.no

13 September 2006

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.

Keywords: stochastic programming, scenario-generation, copula, shape, multivariate structure.

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 modeling 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 et al. (2003), Høyland and Wallace (2001), Høyland et al. (2003), Pflug (2001), Römisch and Heitsch (2003), and Heitsch and Römisch (2005). For an overview, see Dupačová et al. (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 et al. (2003). While our experience shows that in many applications four moments provide a sufficient control over 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

*Molde University College, P.O. Box 2110, N-6402 Molde, Norway

1 random variables. It does not capture any non-linear dependencies, and it does not tell us anything
2 about the “shape” of the bivariate structure. In fact, using the Pearson correlation often implicitly
3 means assuming the elliptical shape of the normal distribution.

4 On the other hand, several recent studies—e.g. Hu (2006), Longin and Solnik (2001), Patton
5 (2002, 2004)—point out that some financial data are not elliptical, showing for example higher cor-
6 relations for downturns than for upturns (all markets tend to crash together). This is illustrated in
7 Figure 1, which shows a scatterplot of daily returns of US and UK small cap stocks, using data from
8 MSCI¹. To demonstrate that the asymmetry does not come from the marginal distributions, we present
9 also a plot of returns with margins transformed to the standard normal distribution.

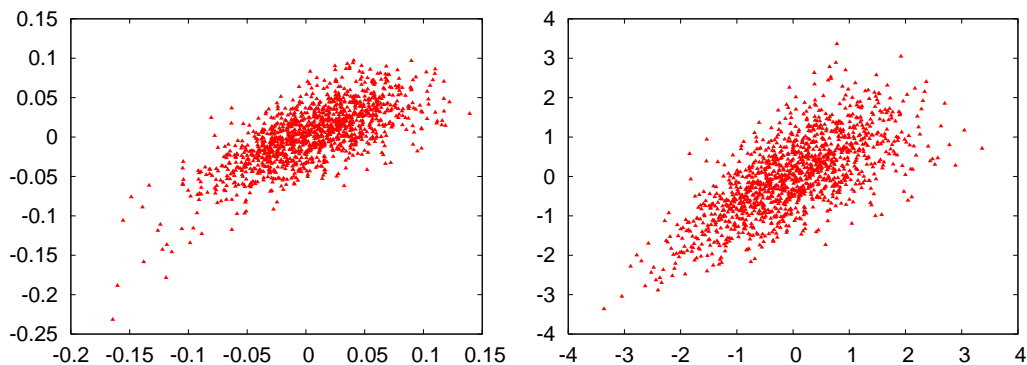


Figure 1: Scatter plot of daily returns of US small caps vs. UK small caps. The left figure shows the actual data, the right figure the data with margins transformed to standard normal distribution, to demonstrate that the asymmetry is not caused by the marginal distributions.

10 While we are not aware of studies from other areas, we find it likely that structures significantly
11 different from those of the normal distribution can be found in many practical settings.

12 In this paper, we propose a general framework that can, at least theoretically, generate scenarios
13 with any multivariate structure. In addition, we propose several methods that fit into the framework
14 and can be used in different cases.

15 The framework is based on *copulas*, a concept that has been used in statistics and finance for
16 some time—see for example Bouyé et al. (2000), Clemen and Reilly (1999), Rosenberg (2003)—yet
17 remains virtually unknown in the rest of the OR community. To our knowledge, copulas have not yet
18 been presented as a basis for scenario generation.

19 The rest of the paper is organised as follows: In the first section, we present the main results
20 from the copula theory and discuss what it can offer for the scenario-generation problem. In the next
21 section, we present the general framework, with more details coming in Section 3. The framework is
22 then exemplified in Section 4.

23 **1 Copulas and their place in scenario generation**

24 This section first presents the notion of a copula, then presents the main results from copula theory,
25 and finally shows what this means for scenario generation.

¹ Morgan Stanley Capital International Inc.; www.msci.com/equity/.

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 $C : [0, 1]^n \rightarrow [0, 1]$. *Sklar’s theorem* states that for any n -dimensional CDF F with marginal distribution functions F_1, \dots, F_n , there exist a copula C such that

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

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 $\mathbf{u} = (u_1, \dots, u_n) \in [0, 1]^n$,

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

where F_i^{-1} is the generalised 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 $G_i^{-1}(F_i(\tilde{X}_i))$ has CDF G , 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 $C(u_1, u_2) = u_1u_2$, 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 :

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

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

For more information about copulas, see for example Clemen and Reilly (1999), Nelsen (1998), Sklar (1959, 1996). In addition, substantial information can be found in the help file of Matlab[®] 7, in the section “Simulating Dependent Random Variables Using Copulas”.

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 the 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: we start by modelling/generating the copula, i.e. the multivariate

1 structure with uniform margins. This can be done, for example, by generating or sampling a distri-
2 bution with the desired structure without regard to the marginal distributions involved. Thereafter we
3 transform the margins to the standard uniform distribution to obtain the copula. Once we have the
4 copula, we transform the margins again to get the desired marginal distributions. This opens some
5 new possibilities for scenario generation, some of which are listed here:

6 **Combining different (standard) copulas and margins**

7 If we compare the normal distribution with t distributions (with sufficiently small number of d.o.f.),
8 the most obvious difference is in the tails of the marginal distributions. There is, however, also one
9 important difference between the two implied copulas, i.e. between the multivariate structures: the
10 t distribution exhibits a *tail dependence*, defined as follows: A bivariate random vector $(\tilde{X}_1, \tilde{X}_2)$ is
11 lower-tail dependent if its lower-tail dependence coefficient

$$12 \quad \lambda_L = \lim_{v \rightarrow 0^+} \mathbb{P}\{\tilde{X}_1 \leq F_{\tilde{X}_1}^{-1}(v) | \tilde{X}_2 \leq F_{\tilde{X}_2}^{-1}(v)\}$$

13 is strictly positive. Upper-tail dependence is defined analogously. The normal distribution is tail-
14 independent as long as the correlation is strictly smaller than one. This means that the extreme events
15 are always independent, so we won't get really extreme scenarios where everything goes awry.

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

20 **Introducing asymmetry**

21 Instead of the standard t copula, we can use a copula from one of the skew- t distributions. These
22 distributions allow for several types of asymmetric dependencies, the most important of which is the
23 possibility of having higher correlation on the down-turn than on the up-turn—an effect observed, for
24 instance, in some financial data.

25 Unfortunately, there are several different skewed versions of t distributions, each with different
26 strengths and weaknesses. For information about the most important ones, see for example Adcock
27 (2003), Azzalini and Capitanio (2003), Bauwens and Laurent (2002), Demarta and McNeil (2005),
28 Jondeau and Rockinger (2000), Jones (2001). In addition, there is the noncentral t distribution and
29 Pearson Type IV distribution. For information on the latter, see Heinrich (2004).

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

33 **Using principal components**

34 In many applications, it can be argued that there are too many random variables in the model, and the
35 dimension could (and should) be reduced by techniques like principal components analysis (PCA).
36 In addition to decreasing the dimension of the stochastic vector, the principal components are also
37 uncorrelated—and therefore, *in the case of normal distributions*, independent. This means that sce-
38 narios for the individual principle components can be generated independently, converting the multi-
39 variate scenario generation to a much easier univariate generation problem. (The univariate margins
40 can be combined into the multivariate vector in an all-against-all fashion, or by a random coupling of

1 the margins. With the former, the number of scenarios grows exponentially with the dimension of the
 2 random vector, often resulting in a need to use a scenario-reduction procedure afterwards, while the
 3 latter yields scenarios that are only approximately independent.)

4 For other than normal distributions, however, the principal components are only uncorrelated, so
 5 there is still some dependence structure to be captured. As an example, see Figure 2, where the two
 6 principle components are clearly not independent, despite having zero correlations. Yet, as long as
 7 we use correlations as the only description of the multivariate structure, we are not able to make the
 8 distinction between uncorrelated and independent random variables and therefore can not model the
 9 structure properly.

10 It is therefore easy to forget the distinction between uncorrelated and independent. So much so,
 11 that it is possible to find papers that either claim that principal components are independent—see, for
 12 example, DeMiguel and Mishra (2006),—or generate them as independent without explicitly saying
 13 so—as in Jamshidian and Zhu (1996). In addition, even when authors are aware of the problem they
 14 might still assume normality and treat the principal components as independent, most likely for the
 15 lack of better tools: Frauendorfer and Schürle (2005), Haldrup (2002), Topaloglou et al. (2002).

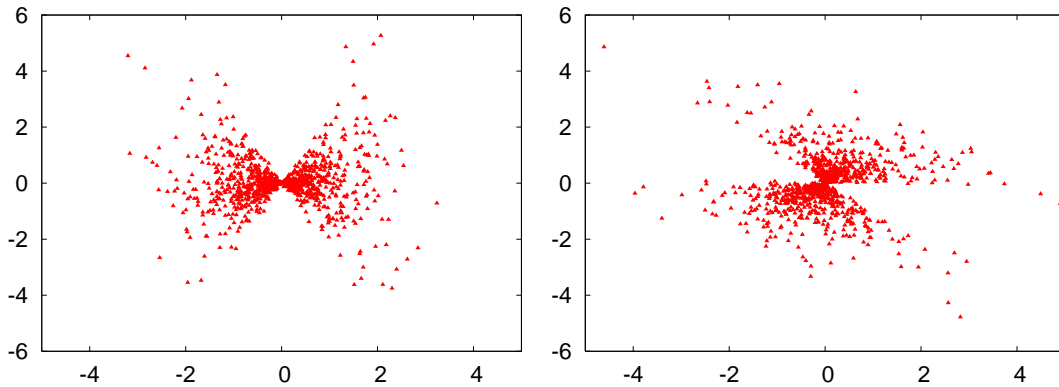


Figure 2: Bi-variate distribution with margins $\tilde{x}_1 = \tilde{\xi}_1$, $\tilde{x}_2 = \tilde{\xi}_1 \tilde{\xi}_2$, with $\tilde{\xi}_1, \tilde{\xi}_2 \sim N(0, 1)$, independent. The left figure shows a sample from the random vector $\tilde{\mathbf{x}} = (\tilde{x}_1, \tilde{x}_2)$, the right 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.

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

21 1.3 Stability and Optimality gap

22 The ultimate test of the quality of a scenario tree will be how well it fits the corresponding stochastic
 23 program. It is not the purpose of this paper to discuss that issue, but we would like to point out that
 24 there are several ways to think about quality of a scenario tree. Obvious possibilities are to compare
 25 the scenario tree directly to the underlying distribution, using metrics from probability theory—see for
 26 example Heitsch et al. (2006),—or comparing (optimal) values of the relevant optimization problem,
 27 hence using the optimization problem as a metric. In the latter case, the performance of the scenario-
 28 based solutions can be evaluated using either a simulator—as in Kaut and Wallace (2003),—or a

1 confidence interval obtained by solving several optimization problems, see Bayraktan and Morton
 2 (2005), Chiralaksanakul and Morton (2004), Linderoth et al. (2002).

3 **2 The algorithm**

4 In this section, we present the algorithm for scenario-generation, i.e. for generating a discrete multi-
 5 variate sample satisfying given properties/characteristics. In particular, we focus on the case when
 6 the distribution is given in the form of historical data—but the algorithm can be, possibly with some
 7 adjustments, used in other cases as well.

8 **2.1 Basic structure**

9 Assume that we start with historical data set $\mathbf{D} \in \mathbb{R}^{n \times n_D}$, where n is the number of random variables
 10 and n_D is the length of the data set. We want to “replicate” \mathbf{D} by n_S scenarios. i.e. by a discrete
 11 n -dimensional random vector $\tilde{\mathbf{X}}$ with n_S outcomes per variables. Typically, $n_S < n_D$ or even $n_S \ll n_D$.
 12 The algorithm/framework is then as follows:

13 1. Transform the data set \mathbf{D} to a new set \mathbf{C} with standard uniform margins by

$$14 \quad c_{ij} = F_i(d_{ij}), \quad i = 1 \dots n, \quad j = 1 \dots n_D,$$

15 where F_i is an estimate of the (univariate) distribution function of the i -th random variable. In
 16 our case, we use empirical CDFs from the data set, so columns of \mathbf{C} consists of the ranks of the
 17 columns of \mathbf{D} , scaled to the interval $(0, 1)$:

$$18 \quad c_{ij} = \frac{1}{n_D} (\text{rank}(u_{ij}, \mathbf{u}_i) - 0.5), \quad i = 1, \dots, n,$$

19 where $\text{rank}(x_s, \mathbf{x})$ is the rank (order) of value x_s in a vector \mathbf{x} , with

$$20 \quad 1 = \text{rank}(\min(\mathbf{x}), \mathbf{x}) \leq \text{rank}(x_s, \mathbf{x}) \leq \text{rank}(\max(\mathbf{x}), \mathbf{x}) = n_D.$$

21 In other words, $\mathbf{C} = (c_{ij})$ is a copula corresponding to the data set \mathbf{D} . We will refer to it as the
 22 “historical copula”.

23 2. Based on the historical copula \mathbf{C} , create the copula for the scenarios, i.e. a sample \mathbf{U} from an
 24 n -variate discrete random vector $\tilde{\mathbf{U}}$ with standard uniform margins and structure close to the
 25 one of \mathbf{C} . This can be done in several different ways, details will be discussed in Section 3:

- 26 • Sampling from the historical copula \mathbf{C} .
- 27 • Using some parametric family of copulas, with parameters estimated from the historical
 28 copula.
- 29 • Creating the structure by coupling of the ranks.

30 3. Once we have the “scenario copula” \mathbf{U} , we have to transform the margins from the uniform to
 31 the desired distribution. The options are:

- 32 • Using the empirical CDF from the historical data.
- 33 • Using some parametric family of distributions, with parameters estimated from the histor-
 34 ical data. The margins are then obtained by the inversion method.

- Compute the moments from the historical data and use some 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).

2.2 Details and comments

Controlling the correlations/covariances

While the change of the margins does not change any copula-based measure (like Spearman’s rank correlations), it will change the Pearson’s correlations. The difference is generally hard to predict as it depends on several factors.

For example, if we sample the copula from historical data and then transform the margins to the distributions equal to (or close to) the historical ones, the difference can be expected to be small—in fact, it will converge to zero as the sample size increases (provided we have enough data).

On the other hand, if we obtain the copula from one multivariate distribution (like t -distribution) and then transform the margins to another distribution (like normal), there will always be a difference in the correlations of the starting distribution and the final scenarios. The size of the difference will depend on the difference (in shape) between the initial and the final marginal distributions.

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.

Fixing margins of the copula sample

If we use sampling to get the copula for the scenarios in Step 2 of the algorithm, we face the usual pitfalls of the sampling approach: even if the procedure is unbiased in the limit, for small n_S it can be very unstable. Fortunately, in the case of copulas the margins follow the standard uniform distribution and can be adjusted accordingly.

In our case, we have stretched each margin so that the points fell into the centers of subintervals of length $1/n_S$:

$$u'_{is} = \frac{1}{n_S} (\text{rank}(u_{is}, \mathbf{u}_i) - 0.5).$$

Since the stretching constitutes a monotonous transformation of the margins, it does not change the copula.

A word of caution: while fixing the margins improves the stability, it also causes the margins to have the same values in all scenario trees (assuming the transformation in Step 3 is deterministic). The only difference is how the margins are connected to form the multivariate distribution—the copula. If this, for any reason, causes a problem to the modeller, but the sampled margins \mathbf{u}_i are too unstable to be used unchanged, it is possible to use weighted averages $\alpha \mathbf{u}'_i + (1 - \alpha) \mathbf{u}_i$ instead, with a suitable value for α .

1 Using moment-matching for the transformation of margins

2 If we use a cubic transformation to transform the margins to a distribution with specified first four
3 moments, it is important to realise that the cubic transformation is not necessarily strictly increasing,
4 so it can change the copula. In addition, the transformation may not even be possible in one step,
5 since the ‘distance’ of the target distribution from the uniform distribution may be too big.

6 To minimise this danger, we should first transform the margins to a distribution that is closer to
7 the target, using an (increasing) inverse CDF that does not change the copula. As a result, the cubic
8 transformation will be closer to identity, decreasing the possible distortion of the copula. Already
9 transforming from uniform to normal will help for most of the common distributions, but we could do
10 even better with some four-parameter distributions like skew- t , non-central t , or Pearson Type IV—see
11 Section 1.2 for a list of related papers.

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

13 If we require control of moments and correlations, we can use the scenarios obtained by the copula-
14 based method as a starting point for the moment-matching algorithm from Høyland et al. (2003). In
15 this context, the algorithm can be seen as a series of transformations to transform a starting sample to
16 a sample with distribution with specified first four moments and a correlation matrix. Two transfor-
17 mations are used to do this: a cubic transformation to correct the moments of the margins and a matrix
18 transformation using a Cholesky component of the correlation matrix to correct the correlations.

19 Compared to using the moment-matching algorithm only, using the copula-based approach to get
20 a starting point should improve the algorithm in the following ways:

- 21 • In the current implementation of the algorithm, we start with discretized normal variates that
22 are paired randomly and then transformed to the specified correlation matrix. Hence, there is no
23 control over the initial structure and even a simple sampling from the historical copula should
24 give better results.
- 25 • Just as the matrix transformation distorts the margins, the cubic transformation distorts the
26 copula structure, with size of the distortion increasing with the distance of the starting and the
27 target distributions. Therefore, even starting with approximate marginal distributions from the
28 copula-based method should be better than starting with the normal margins (obviously unless
29 the target distribution is normal).

30 Both these improvements should lead to better stability and/or better solutions of the optimization
31 model, for a given number of scenarios. The extent (and indeed the presence) of the improvement will
32 be tested in Section 4.

33 3 Methods for constructing the scenario copula

34 In this section, we present the methods used for the generation of the scenario copula, i.e. the scenario
35 distribution with uniform margins. For each method, we discuss where it could be used and present a
36 numerical example.

37 3.1 Sampling

38 The easiest option is to sample the values of \tilde{U} from the historical copula \mathbf{C} . It is also the only method
39 that guarantees the correct distribution in the limit (in this case, as n_S approaches n_D). Note that we

1 sample only the copula (the multivariate structure), as the marginal distributions are fixed later in
2 Step 3 of the algorithm. This is a major difference from the standard sampling method, where we do
3 not have any control over the marginal distributions (except, perhaps, for simple corrections of means
4 and variances).

5 **3.2 Using a standard copula**

6 Another option is using some standard copula: copulas, just like distributions, have many parametric
7 families with specialized methods for generation. Once we have decided for a particular copula, we
8 have to estimate its parameters from the historical copula \mathbf{C} and then use an appropriate method to
9 create a sample from the copula. The best source of information on copula families is probably Nelsen
10 (1998), other options include Bouyé et al. (2000), Hu (2006), Romano (2002).

11 In addition to the copula families, it is possible to use copulas from some standard distribution
12 like normal or t , or the skewed versions of t distributions mentioned in Section 1.2. In this case, we
13 generate a sample from the given distribution and then transform it to a copula in the same way as we
14 did with the data in Step 1 of the algorithm in Section 2.

15 Note that the transformation to copula removes all information of the marginal distributions, so
16 only the copula (structure) of the chosen distribution remains. This means, for example, that we do
17 not have to estimate the scale parameters, as they do not influence the copula. In other words, the
18 normal copula depends only on the correlations, the t copula in addition on the degrees of freedom,
19 and the skewed version of t in addition on the skewness parameter(s). Furthermore, the skewness
20 parameter(s) are used only to control the asymmetry of the skewed- t copulas, they have no relation to
21 the skewness of the final distribution (again, because the marginal distributions are removed by the
22 transformation to copula). This is illustrated in Figure 3 where the distribution remains skewed even
23 when the margins are transformed to the standard normal distributions. For comparison, we present
24 also a distribution obtained by combining the skewed- t margins with a standard normal copula. Note
25 that unlike the skewed- t distributions, the extreme values do not happen together when we use the
26 same margins with the normal copula. This is in concordance with the fact that the normal copula
27 does not exhibit tail dependence, as mentioned in Section 1.2. For more information on using copulas
28 of standard distributions, see for example Demarta and McNeil (2005), Romano (2002).

29 **3.3 Coupling of the ranks of the margins**

30 Any finite discrete copula can be seen as a set of couplings of the ranks of the margins: the first
31 element could, for example, consist of the second smallest value of the first margin and the 13th
32 smallest value of the second margin, etc. Hence, to generate a copula it is enough to generate a set
33 of couplings with given properties. This would lead to a “property-matching-type” of algorithm for
34 the copula. Development of such an algorithm is, however, out of the scope of this paper as is left for
35 future research.

36 **4 Case study – portofilo optimization with CVaR constraint**

37 In this section, we test several variants of the scenario-generation method on a portfolio optimization
38 model with a CVaR constraint. It is a one-period LP model, with positive variables (positions) that
39 sum up to one. The LP formulation of the CVaR constraint comes from Rockafellar and Uryasev
40 (2000) and Uryasev (2000).

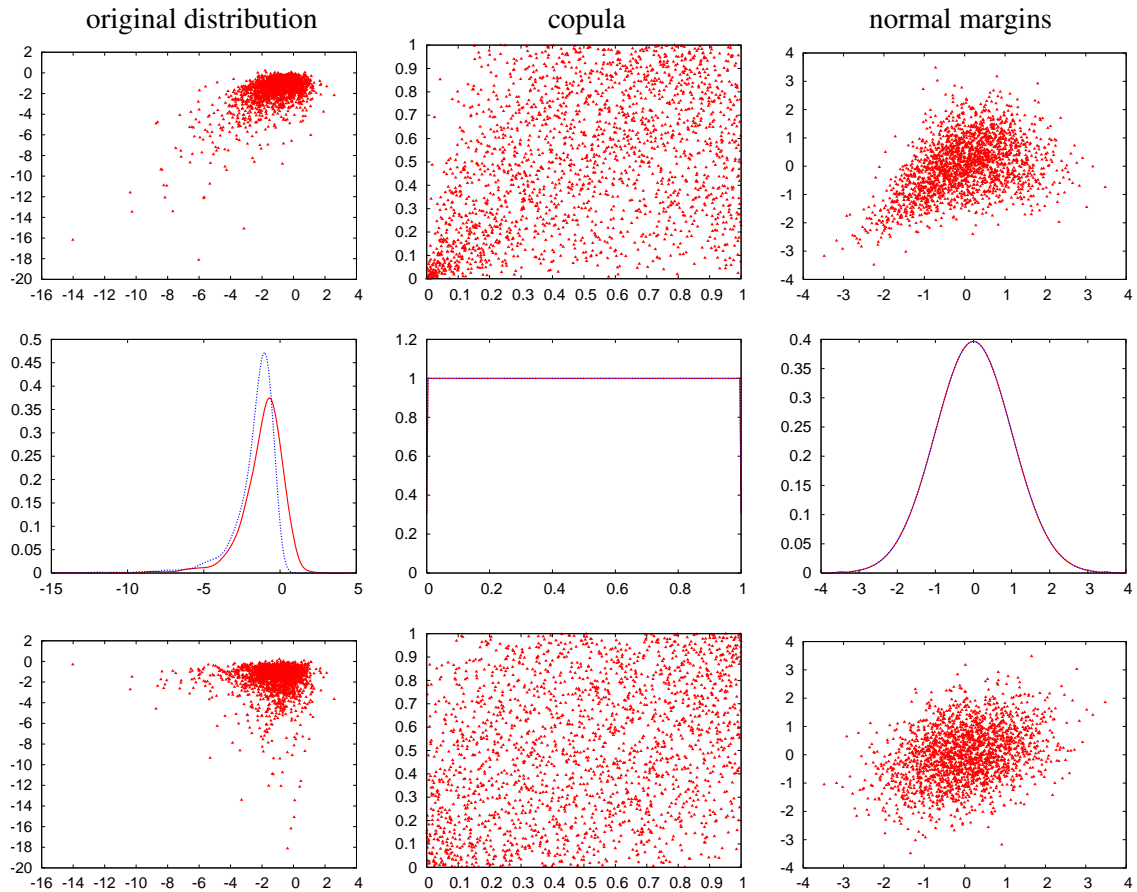


Figure 3: Skewed- t distribution and copula, using a skewed- t variant from Azzalini and Capitanio (2003) with 5 degrees of freedom and skewness parameters $(-0.5, -0.9)$. The first two rows show the two-dimensional scatter plots and marginal densities, respectively. The third row shows a distribution obtained by combining the skewed- t margins with a standard normal copula. In other words, the marginal distribution in the second row correspond both to the first and the third row. Note that 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, $U(0, 1)$ and $N(0, 1)$, respectively.

1 The CVaR model has been chosen because it can be expected to react to differences in the shape
2 of the distribution, particularly the shape of lower tail of the return distribution. Two sets of data
3 were used for the model: the main data set consists of daily prices of seven stock indices and three
4 government bonds, from 1987-07-09 to 2005-04-05 (4476 points). This data set was kindly provided
5 by Kjetil Høyland from DNB Nor, Oslo, Norway. The second data set consists of 1302 daily prices
6 of 10 stock indices, obtained from MSCI.

7 We have tested several version of the copula-based approach, as well as sampling and the moment-
8 based approach from Høyland et al. (2003). In addition, several post-processing (adjusting) methods
9 were tried to improve the scenarios.

10 All the copula-based approaches used sampling in Step 2 of the algorithm presented in Section 2,
11 and differed in the implementation of the transformation of the margins in Step 3 of the algorithm:

1 one option was to transform the margins using the inverse of the empirical CDFs from the data.²
 2 Alternatively, we can transform margins to some standard distribution with known CDF (in our case
 3 normal) and then use a post-process (in our case moment-matching) to get the correct margins. In
 4 addition, we tested using the fixed margins, as described in Section 2.2.

5 This gives the following scenario-generation methods:

- 6 **histRet** Sample from historical returns.
- 7 **mom+cor** Use the moment-matching alg. from Høyland et al. (2003) to match the first four moments
 8 of the margins plus the correlation matrix.
- 9 **histCopNorm** Sample from copula and transform margins to $N(0, 1)$.
- 10 **histCopFixNorm** The same, using fixed margins of the copula.
- 11 **histCopFixICdf** Sample from copula and transform margins using the inverse of the historical CDF.
- 12 **histCopFixICdf** The same, using fixed margins of the copula.

13 Note however that ‘histCopFixICdf’ is the same as ‘histRet’: first we get the sample from the his-
 14 torical returns and apply the empirical CDF to transform it to copula. Then we apply the inverse of
 15 the empirical CDF to get the correct distribution of margins, leaving us with the original sample. This
 16 method will therefore not be considered in the tests. Note also that ‘histCopFixICdf’ is different, as
 17 we change the copula sample before using the inverse CDF.

18 An obvious choice of a post-process is to adjust the margins to correct their means and variances—
 19 this can be done by a simple linear transformation. Alternatively, we can use the moment-matching
 20 method from Høyland et al. (2003) to correct the moments of the margins and the correlation matrix.³

21 This gives the following post-processing methods:

- 22 **none**
- 23 **meanVar** Correct means and variances of the margins.
- 24 **moments** Use the moment-matching method to correct the first four moments of the margins.
- 25 **mon+cor** Correct the correlation matrix, in addition to the moments.

26 Not all combinations of scenario-generation and post-process methods are possible and sensible,
 27 so only the following 13 methods were tested:

	initialization	post-process	comment
0	histRet	none	standard sampling
1	histRet	meanVar	sampling with correction of means and variances
2	histRet	moments	
3	histRet	mom+cor	
4	mom+cor	none	‘pure’ moment-matching, as done in Høyland et al. (2003)
5	histCopFixNorm	moments	
6	histCopFixNorm	mom+cor	
7	histCopFixICdf	moments	
8	histCopFixICdf	mom+cor	
9	histCopFixICdf	none	sampling from the copula, using fixed uniform margins
10	histCopFixICdf	meanVar	
11	histCopNorm	moments	
12	histCopNorm	mom+cor	

² In the actual implementation we have interpolated the inverse empirical CDF using cubic splines to get values at points different from the data points.

³ The algorithm allows specifying a starting point (distribution) for the iteration process.

1 4.1 The tests

2 For the main data set, we have tested stability with two different CVaR constraints, one close to the
3 minimum-risk value, and one more risky. For the MSCI data only one CVaR value was tested. Three
4 different sizes of scenario trees were used in each case: 50, 250, and 1000 scenarios.

5 In each case, one hundred scenario trees were generated, the model solved on them, and the
6 solution evaluated on the reference tree consisting of the whole data set. We could thus perform both
7 the in- and out-of-sample tests as described in Kaut and Wallace (2003), as well as checking the bias
8 introduced by the scenarios.

9 The CVaR model was written in the GNU MathProg language and solved by `glpsol`, both parts
10 of GLPK⁴. The other tests were implemented in GNU Octave⁵. Finally, GNUPlot was used to produce
11 the charts to visualize the results of the simulations.

12 4.2 The main result

13 Out of the thirteen tested methods, the one that performed consistently best in terms of both stability
14 and bias was ‘histCopFixICdf’ with ‘mom+cor’ post-process (method 8 above), so let us first describe
15 it in more details:

- 16 1. Sample values from the historical returns, exactly like in the standard sampling approach.
- 17 2. Get the copula of the sample, i.e. transform all the margins to the standard uniform distribution,
18 using the empirical CDFs.
- 19 3. For each margin, spread the values evenly on interval (0,1), as described in Section 2.2.
- 20 4. Transform the margins back to the original distributions, using the inverse of the historical
21 CDFs. *Already now we have a significantly improved sample, i.e. a method that is significantly
22 better than standard sampling.*
- 23 5. To further improve the match of the sample properties to the historical data, call the moment-
24 matching code with the current sample as a starting point.

25 4.3 Other observations

26 It is not possible to present the results of all the tests, so we have instead decided to present results for
27 the most “important” methods: sampling with correction of means and variances, moment-matching,
28 and the best copula-based method. See Figure 4 for stability plots of these three methods on trees with
29 50, 250, and 1000 scenarios.

30 The figure can also illustrate most of the following observations—even if the observations them-
31 selves are based on results of all the tests (all methods, all data sets, all sizes of scenario trees).

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

⁴ The GNU Linear Programming Kit, see <http://www.gnu.org/software/glpk/glpk.html>. GNU MathProg is an implementation of a subset of AMPL. Precompiled version of GLPK for Windows can be obtained from <http://gnuwin32.sourceforge.net/>.

⁵ GNU Octave is a high-level language mostly compatible with Matlab. See <http://www.octave.org/>.

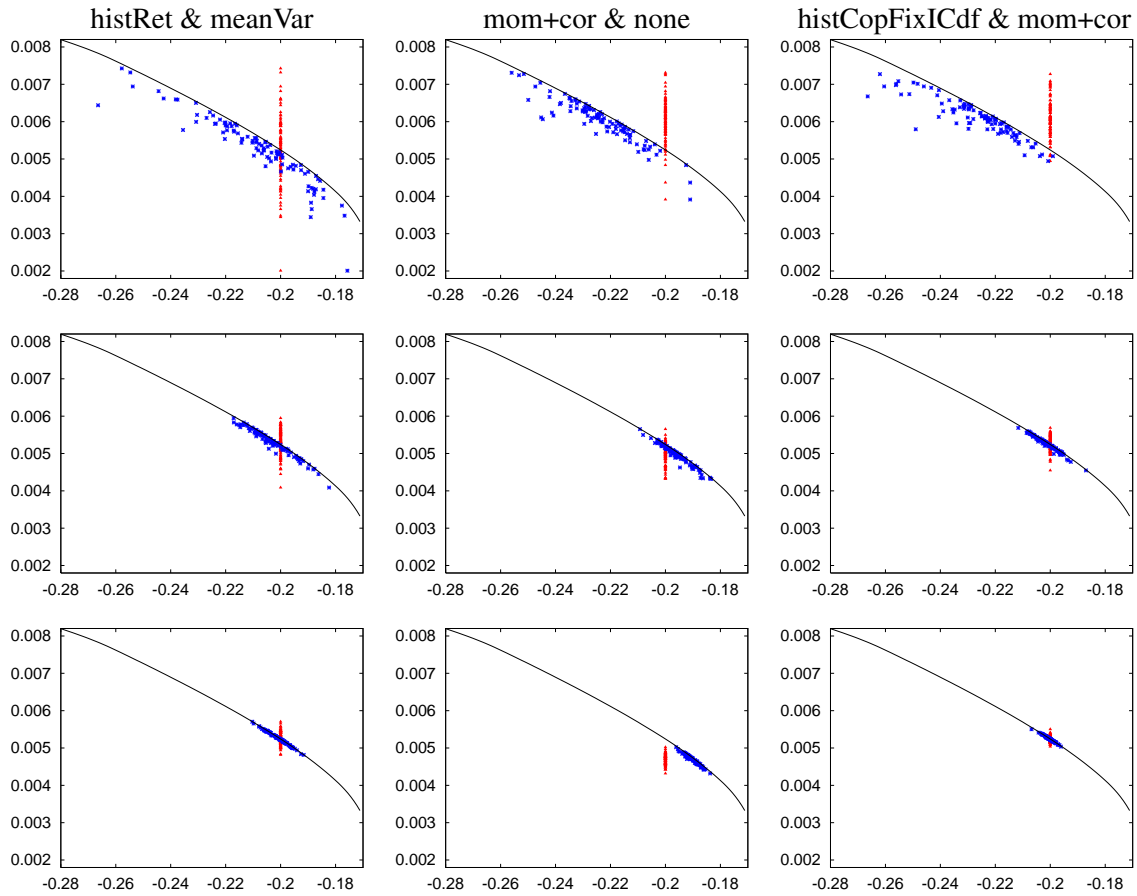


Figure 4: In-sample and Out-of-sample properties of three selected scenario generation methods, on trees with 50, 250, and 1000 scenarios. On the x -axis is CVaR, on the y -axis the objective function values. The in-sample values are scattered along a vertical line “CVaR=-0.2”, caused by the constraint 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 (out-of-sample) values.

- 1 • Matching the moments and correlations of the historical returns without any direct use of the
- 2 data leads to one of the most stable methods, but can introduce a bias to the results. This is due
- 3 to the fact that starting without any particular structure basically implies the elliptical structure
- 4 of the normal distribution. When the data has significantly different structure, this approach
- 5 leads to a bias in the results.
- 6 In the last row of plots in Fig. 4, we can see that the moment-matching for 1000 scenarios led
- 7 to smaller risk than required. However, in the case of CVaR constraint at -0.25 (instead of -0.2),
- 8 the moment-matching resulted in portfolio with a higher-than-required risk. This illustrates that
- 9 the bias caused by moment-matching is unpredictable, including the sign of the bias.
- 10 • Fixing the margins of the copulas to a fixed discretization improved the stability in most of the
- 11 tests (and did not make it worse in the rest). As it did not introduce/increase the bias of the
- 12 results as well, we conclude that the fixing of margins is a useful technique.
- 13 • Transforming the margins via the normal distribution (methods histCopNorm and histCop-

1 FixNorm) may in some cases decrease stability or introduce a bias to the results. This is, again,
2 not surprising as this method relies only on the four moments to set the margins, discarding thus
3 the additional information from the empirical CDF. In addition, the cubic transformation is not
4 guaranteed to be monotonous, so it can distort the copula, see Section 1.1.

- 5 • As expected, controlling the first four moments and correlations provides more stable results
6 than controlling only the moments, which in turn is more stable than controlling only means
7 and variances.⁶

8 Conclusions

9 In this article we have shown how to separate marginal distributions from the multivariate structure—
10 the copula—when generating scenarios. This way we can combine different approaches which, sep-
11 arately, may be good for only one of these factors. We show that in some situations the combined
12 approach retains both margins and copula, while in others we end up with approximations. By this
13 separation of margins and copula it is for example possible to sample from the underlying distribution
14 to obtain an approximation of the structure, while not having to rely on the same sample for margins.
15 The margins can then be set up with other methods that are better suited to handle them, but which
16 are possibly even unable to handle multivariate structure. Our example from portfolio management
17 indicates that such an approach is indeed a good idea.

18 Acknowledgements

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

24 References

- 25 C. J. Adcock. Asset pricing and portfolio selection based on the multivariate skew-student distribution.
26 Discussion Papers Series 2003-02, The Department of Economics, The University of Sheffield,
27 2003.
- 28 Adelchi Azzalini and Antonella Capitanio. Distributions generated by perturbation of symmetry with
29 emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B*
30 (*Statistical Methodology*), 65:367–389, 2003.

⁶ There was, however, a strange effect in the case of 50 scenarios with the main data set. There, the trees generated with the moment-matching post-process introduced a bias so the out-of-sample risk was consistently bigger than the in-sample one. When only means and variances were corrected, the bias disappeared, though the results were significantly less stable. This effect has not been observed on the MSCI dataset or on trees with 250 or 1000 scenarios.

At the moment, we do not know what causes this behaviour, but it could be connected to the way we compute the (higher) moments, using population-based rather than sample-based formulas.

- 1 Luc Bauwens and Sébastien Laurent. A new class of multivariate skew densities, with application
2 to GARCH models. *Computing in Economics and Finance 2002* 5, Society for Computational
3 Economics, 2002. available at <http://ideas.repec.org/p/sce/scecf2/5.html>.
- 4 Güzin Bayraksan and David P. Morton. Assessing solution quality in stochastic programs. *Stochastic
5 Programming E-Print Series*, <http://www.speps.info>, 2005.
- 6 Eric Bouyé, Valdo Durrleman, Ashkan Nikeghbali, Gaël Riboulet, and Thierry Roncalli. Copulas for
7 finance: A reading guide and some applications. working paper, Crédit Lyonnais, Paris, 2000.
- 8 Anukal Chiralaksanakul and David P. Morton. Assessing policy quality in multi-stage stochastic
9 programs. *Stochastic Programming E-Print Series*, <http://www.speps.info>, 2004.
- 10 Robert T. Clemen and Terence Reilly. Correlations and copulas for decision and risk analysis. *Man-
11 agement Science*, 45(2):208–224, February 1999.
- 12 S. Demarta and A. J. McNeil. The t copula and related copulas. *International Statistical Review*, 73
13 (1):111–129, 2005.
- 14 Victor DeMiguel and Nishant Mishra. A multistage stochastic programming approach to network
15 revenue management. Lbs working paper, London Business School, 2006.
- 16 Jitka Dupačová, Giorgio Consigli, and Stein W. Wallace. Scenarios for multistage stochastic pro-
17 grams. *Ann. Oper. Res.*, 100:25–53, 2000. ISSN 0254-5330.
- 18 Jitka Dupačová, Nicole Gröwe-Kuska, and Werner Römisch. Scenario reduction in stochastic pro-
19 gramming. *Mathematical Programming*, 95(3):493–511, 2003.
- 20 A. I. Fleishman. A method for simulating nonnormal distributions. *Psychometrika*, 43:521–532, 1978.
- 21 Karl Frauendorfer and Michael Schürle. Refinancing mortgages in switzerland. In Stein W. Wallace
22 and William T. Ziemba, editors, *Applications of Stochastic Programming*, chapter 23, pages 445–
23 469. SIAM Society for Industrial and Applied Mathematics, Philadelphia, 2005.
- 24 Roger Halldin. *Scenario Trees for Inflow Modelling in Stochastic Optimisation for Energy Planning*.
25 PhD thesis, Lund University, Sweden, 2002.
- 26 Joel Heinrich. A guide to the pearson type IV distribution. Technical Report Memo 6820, The Collider
27 Detector at Fermilab, Fermilab, Batavia, Illinois, 2004. Available at [http://www-cdf.fnal.gov/
28 publications/cdf6820_pearson4.pdf](http://www-cdf.fnal.gov/publications/cdf6820_pearson4.pdf).
- 29 H. Heitsch and W. Römisch. Scenario tree modelling for multistage stochastic programs. Techni-
30 cal Report Preprint 296, DFG Research Center MATHEON, “Mathematics for key technologies”,
31 Technische Universität Berlin, Germany, 2005.
- 32 H. Heitsch, W. Römisch, and C. Strugarek. Stability of multistage stochastic programs. Technical
33 Report Preprint 324, DFG Research Center Matheon ”Mathematics for key technologies”, 2006.
34 Accepted for publication in *SIAM Journal in Optimization*.
- 35 K. Høyland and S. W. Wallace. Generating scenario trees for multistage decision problems. *Man-
36 agement Science*, 47(2):295–307, 2001.

- 1 Kjetil Høyland, Michal Kaut, and Stein W. Wallace. A heuristic for moment-matching scenario gen-
2 eration. *Computational Optimization and Applications*, 24(2-3):169–185, 2003. ISSN 0926-6003.
- 3 Ling Hu. Dependence patterns across financial markets: A mixed copula approach. *Applied Financial*
4 *Economics*, 16(10):717–729, 2006.
- 5 Farshid Jamshidian and Yu Zhu. Scenario simulation: Theory and methodology. *Finance and Stochas-*
6 *tics*, 1(1):43–67, 1996.
- 7 Eric Jondeau and Georg Michael Rockinger. Conditional volatility, skewness and kurtosis: Existence
8 and persistence. Working Paper 77, Banque de France, Direction Generale des Etudes, 2000. Also
9 appeared as HEC (Ecole des Hautes Etudes Commerciales) Department of Finance Working Paper
10 No. 710/2000.
- 11 M. C. Jones. Multivariate t and beta distributions associated with the multivariate f distribution.
12 *Metrika*, 54:215–31, 2001.
- 13 Michal Kaut and Stein W. Wallace. Evaluation of scenario-generation methods for stochastic pro-
14 gramming. Stochastic Programming E-Print Series, <http://www.speps.info>, May 2003.
- 15 Michal Kaut, Stein W. Wallace, Hercules Vladimirou, and Stavros Zenios. Stability analysis of a
16 portfolio management model based on the conditional value-at-risk measure. Feb 2003.
- 17 J. T. Linderoth, A. Shapiro, and S. J. Wright. The empirical behavior of sampling methods for stochas-
18 tic programming. Optimization Online, <http://www.optimization-online.org>, 2002.
- 19 François Longin and Bruno Solnik. Extreme correlation of international equity markets. *The Journal*
20 *of Finance*, 56(2):649–676, 2001.
- 21 Roger B. Nelsen. *An Introduction to Copulas*. Springer-Verlag, New York, 1998.
- 22 Andrew Patton. Skewness, asymmetric dependence, and portfolios. In *PhD Thesis*, chapter 3. De-
23 partment of Economics, University of California, San Diego, 2002.
- 24 Andrew J. Patton. On the out-of-sample importance of skewness and asymmetric dependence for asset
25 allocation. *Journal of Financial Econometrics*, 2(1):130–168, 2004.
- 26 G. C. Pflug. Scenario tree generation for multiperiod financial optimization by optimal discretization.
27 *Mathematical Programming*, 89(2):251–271, 2001.
- 28 R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2
29 (3):21–41, 2000.
- 30 Claudio Romano. Calibrating and simulating functions: An application to the italian stock market.
31 working paper 12, Centro Interdipartimale sul Diritto e l’Economia dei Mercati, 2002.
- 32 W. Römisch and H. Heitsch. Scenario reduction algorithms in stochastic programming. *Computa-*
33 *tional Optimization and Applications*, 24(2-3):187–206, 2003.
- 34 Joshua V. Rosenberg. Non-parametric pricing of multivariate contingent claims. *The Journal of*
35 *Derivatives*, 10(3):9–26, 2003.

- 1 A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de*
2 *Statistique de l'Université de Paris*, 8:229–231, 1959.
- 3 A. Sklar. Random variables, distribution functions, and copulas – a personal look backward and for-
4 ward. In L. Rüschendorf, B. Schweizer, and M. Taylor, editors, *Distributions with Fixed Marginals*
5 *and Related Topics*, pages 1–14. Institute of Mathematical Statistics, Hayward, CA, 1996.
- 6 Nikolas Topaloglou, Hercules Vladimirou, and Stavros A. Zenios. CVaR models with selective hedg-
7 ing for international asset allocation. *Journal of Banking and Finance*, 26(7):1535–1561, 2002.
- 8 S. Uryasev. Conditional value-at-risk: Optimization algorithms and applications. *Financial Engineer-*
9 *ing News*, (14):1–5, 2000. URL <http://fenews.com/>.
- 10 S.W. Wallace and W.T. Ziemba, editors. *Applications of Stochastic Programming*. MPS-SIAM Series
11 on Optimization, Philadelphia, 2005.