Scenario Generation Employing Copulas

K. Sutiene and H. Pranevicius

Abstract—**Multistage stochastic programs are effective for solving long-term planning problems under uncertainty. Such programs are usually based on scenario generation model about future environment developments. In the present paper, the scenario model is developed for the case when enough data paths can be generated, but due to solvability of stochastic program the scenario tree has to be constructed. The proposed strategy is to generate multistage scenario tree from the set of individual scenarios by bundling scenarios based on cluster analysis. The K-means clustering approach is modified to capture the interstage dependencies. Such generation of scenario tree can be useful in cases when it is difficult to construct the adequate scenario tree from the stochastic differential equations or time-series models, and the sampled paths can be obtained by sampling or resampling techniques. While generating the initial fan of individual scenarios, the copula is employed for modeling the dependence between stochastic variables in a multivariate structure. This allows to model nonlinear dependencies between non-elliptically distributed stochastic variables. While investigating the copula effect on the scenario tree structure, we will try to answer the question: does the copula features are captured in the approximate representation of uncertainty in the form of scenario tree. The proposed scenario tree generation method is implemented on sampled data of discount bond yields. The Gaussian copula and Student's t-copula are employed while generating the set of individual scenarios in the multivariate structure.**

Index Terms—**Copula, K-means clustering, Multistage scenario tree construction, Stochastic programming.**

I. Introduction

The concept of scenarios is usually employed for the modeling of randomness in stochastic programming models [1], [2], in which data evolve over time and decisions have to be made independent upon knowing the actual paths that will occur. Such data are usually subject to uncertainty or some kind of risk. For instance, the random variables are the return values of each asset on an investment in portfolio management problems, and the investment decisions must be implemented before the asset performance can be observed. Each scenario can be viewed as one realization of an underlying multivariate stochastic data process. The modeling of randomness employees the set of available past data with the aim of building submodels for each individual stochastic parameter. These submodels are used to generate a set of scenarios that encapsulate the consistent depictions of pathways to possible futures based on assumptions about economic and technological developments. Thus, the factors driving risky events are approximated by a discrete set of scenarios, or sequence of events. This process is known as scenario generation. Scenarios can be generated using various methods, based on different principles: conditional sampling, sampling from given marginals and correlations, moment matching, path based methods, optimal discretization, as in [3]–[7].

A good approximation may involve a very large number of scenarios with probabilities. A better accuracy of uncertainties is described when scenarios are constructed via a simulated data path structure, also known as a scenario fan. But the number of scenarios is limited by the available computing power, together with a complexity of the decision model. To deal with this difficulty, we can reduce the dimension of the initial scenario set by constructing the multistage scenario tree out of it. The decision on the number of stages, time periods and the branching scheme is very important for a good representation of the uncertainty in the form of scenario tree, which is input in the multistage stochastic program. The detailed description of both scenario fan and scenario tree will be given in Section II.

In the present paper, we concentrate on the generation of scenario trees when the underlying stochastic parameters have been determined and the data paths of their realizations can be generated. The scenario tree can be constructed out of sampled paths by employing some classifying method, such as clustering analysis. While bundling the scenarios to the clusters, the interstage dependencies have to be captured. An approach similar to our work is introduced in the article [8], but without a detailed clustering algorithm. Due to this, the K-means clustering method is modified to treat properly the interstage dependencies and is implemented while constructing the scenario tree from simulated data paths.

Such generation of scenario tree can be useful in cases when it is difficult to construct the adequate scenario tree from the stochastic differential equations or time-series models, and the sampled paths can be obtained by sampling or resampling techniques. The proposed scenario tree construction algorithm allows incorporating a copula-based dependence measure [9], [10] to describe the dependence between stochastic variables in a multivariate structure. Due to assumptions of using the Pearson's correlation coefficient, the usefulness of such correlations is restricted. The main advantage of employing

Manuscript received March 22, 2007.

K. Sutiene is with the Business Informatics Department, Kaunas University of Technology, Studentu 56-301, Kaunas LT-51424, Lithuania (corresponding author to provide phone: 370-681-52842; fax: 370-37-451654; e-mail: kristina.sutiene@stud.ktu.lt).

H. Pranevicius is with the Business Informatics Department, Kaunas University of Technology, Studentu 56-301, Kaunas LT-51424, Lithuania (e-mail: hepran@if.ktu.lt).

copulas is that they allow to model the nonlinear dependencies between non-elliptically distributed stochastic variables. To our knowledge, the copulas still are not very popular in generating the scenario trees. According to this, we propose to approximate the multivariate stochastic process by a scenario fan with multivariate structure using copulas. Then, the scenario tree is constructed out from the sampled paths using the modified K-means clustering algorithm. Numerical experience is reported for constructing multivariate scenario trees of discount bond yields, employing two separate – Gaussian and Student's t – copulas.

The rest of the paper is organized as follows. The scenario generation model is introduced in Section II. The copula is incorporated while generating the scenario fan. Section III describes how the data paths can be transformed to scenario tree using the cluster analysis. The K-means clustering algorithm is modified to bundle the time-dependent data. Section IV demonstrates the numerical example of scenario tree generation based on discount bond yields data. Finally, some concluding remarks are given.

II. Copula's Place in Scenario Generation

In general, the scenario generation consists of following steps [11]:

1) Choosing the appropriate model to describe the stochastic parameters. For instance, Econometric models and Time Series (Autoregressive models, Moving Average models, Vector Auto Regressive models), Diffusion Processes (Wiener Processes).
2) Calibration of model parameters using historical data.
3) Generation of data paths from the chosen model. Using statistical approximation (Property Matching, Non parametric methods) or sampling (Random sampling, Bootstrapping), the data paths can be generated performing the discretization of the distribution.
4) Constructing the scenario tree with the desired properties.

The aim of scenario generation is to create a tree structure out of scenarios, which is input in stochastic model. In this Section we will consider the 1)–3) steps of scenario generation.

In multistage stochastic programs the underlying multivariate stochastic data process has to be discrete in time, i.e. $\mathbf{T} = \{0,\dots,T\}$. The points in time $t \in \mathbf{T}$ are called as stage index. The process $\xi = \{\xi_t\}_{t=0}^{T}$ is defined on some probability space $(\Omega, \mathcal{F}, \mathcal{P})$ with ξ_t taking values in some $\mathcal{R}^d$. For instance, these data may correspond to the observed return of d financial assets at different time moments t. In the stochastic programming model the observations and decisions are given as a sequence $x_0, (\xi_0, x_1), (\xi_1, x_2), \dots, (\xi_{T-1}, x_T)$, where $x = \{x_t\}_{t=0}^{T}$ is a decision process, measurable function of ξ. The constraints on a decision at each stage involve past observations and decisions. It means that decision x_t at t is measurable with respect to $\mathcal{F}_{t-1} \subseteq \mathcal{F}$. Following [8], the decision process is said to be nonanticipative. It means that the decision $x_t = x_t(x_{t-1}, \xi_{t-1})$ taken at any $t > 1$ does not depend on future realizations of stochastic parameters or on future decisions.

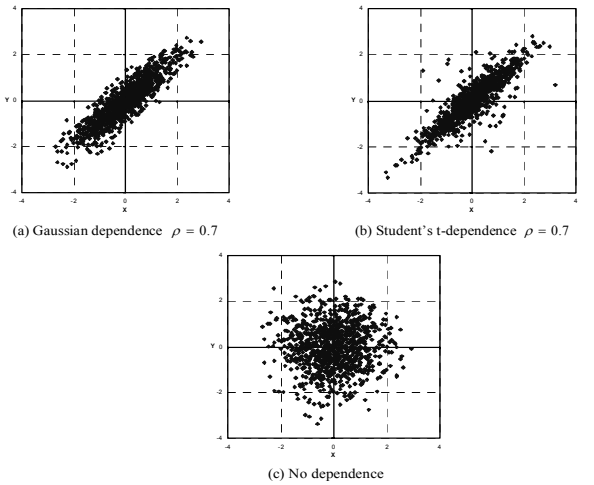

Fig. 1. Dependence structures

The d-dimensional probability distribution function of $\xi_t = (\xi_t^1, \dots, \xi_t^d)'$ at point $y = (y_1, \dots, y_d)'$ is denoted by $f(y)$, the d-dimensional cumulative distribution function is denoted by $F(y)$. The joint distribution F provides a complete information concerning the behavior of ξ_t. The marginal probability distribution function and cumulative distribution function of each element ξ_t^i at point y_i, $i = 1,\dots,d$ is denoted by $f_i(y_i)$ and $F_i(y_i)$, respectively. The primary aim of scenario generation is to represent the distribution f in a reasonable way. In stochastic programming the underlying probability distribution f is replaced by a discrete distribution P carried by a finite number of atoms $\xi^s = (\xi_0^s, \xi_1^s, \dots, \xi_T^s)$, $\xi_t^s = (\xi_t^{s,1}, \dots, \xi_t^{s,d})'$, $s = 1,\dots,S$ with probabilities $p_s = P(\xi^s)$, $p_s \geq 0$ and $\sum_{s=1}^{S} p_s = 1$. The atoms ξ^s, $s = 1,\dots,S$ of the distribution P are called as scenarios. Naturally, the historical data in conjunction with an assumed background model are used to generate the scenarios, applying suitable estimation, simulation and sampling procedures.

In this paper we concentrate on generation of scenarios representing the realizations of multivariate stochastic process whose components are correlated. We define such scenarios as intercorrelated scenarios, meaning that they correlate through the components of multivariate structure. Usually the modeling of dependent variables is performed employing the Pearson's correlation matrix to describe the multivariate structure. Many applications show that relationships among stochastic variables may be very complex, and linear dependence can't reflect these relationships adequately. The reason is that the Pearson's correlation coefficient does not capture any non-linear dependencies, and it is used in application assuming the elliptical shape of normal distribution. We include Fig. 1 as motivation for the ideas of this paper. It shows 1000 bivariate

realizations in three different cases of (X,Y). In all pictures, variables X and Y have standard normal marginal distributions: case (a) and case (b) depict bivariate structure of X and Y with linear correlation coefficient $\rho = 0.7$; case (c) displays the circular plot whereas there are no associations between random variables. However, in case (a) and in case (b) the dependence structure between X and Y is qualitatively quite different. It relates that in case (b) extreme values have a tendency to occur together. This example shows that the dependence between random variables cannot be distinguished on the grounds of correlation alone.

Due to the restrictions of using Pearson's correlation coefficient, we choose to consider the dependence relations of a monotonic nature: it indicates the tendency of two random variables to increase/decrease concomitantly (positive dependence) or contrariwise (negative dependence). If one believes that the dependence relationships among a pair of variables (or their suitable transformation) fulfill the definition of monotone dependence, then the modeling of dependence with Spearman's rank or Kendall's tau correlations is a reasonable way. During the discretization process of *d*-dimensional distribution function F, one can strengthen the dependence in different parts of distribution through the choice of copula. Let's define the copula itself.

Let F be the *d*-dimensional distribution function with margins $F_1, F_2, \ldots, F_d$. The *d*-dimensional copula is a *d*-dimensional distribution function restricted to $[0,1]^d$ with uniform $(0,1)$ marginals. For a given copula C and marginals $F_1, F_2, \ldots, F_d$, the *d*-dimensional distribution function F can be written in such way:

$$F(y_1, \ldots, y_d) = C(F_1(y_1), \ldots, F_d(y_d)). \quad (1)$$

It means that C couples the marginals $F_1, F_2, \ldots, F_d$ to the joint distribution function F. Conversely, for a given joint distribution function F with margins $F_1, F_2, \ldots, F_d$ there is always a copula C satisfying (1). It follows that

$$C(u_1, \ldots, u_d) = F\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right),\ u_i \in [0,1].$$

Copulas are thus multivariate uniform distributions which describe the dependence structure of random variables. From a practical point of view, the copula-based approach allows to select the appropriate marginal distributions for the components of a multivariate system freely, and then link through a suitable copula. That is, the dependence structure between stochastic variables can be modeled independently of marginal distributions. The estimation of copula from historical data can be found in article [10]. In Section IV the simulation algorithm for Gaussian and Student's t-copulas is given.

While approximating the multivariate stochastic distribution F employing copulas, the set of *d*-dimensional intercorrelated scenarios $\xi^s = (\xi_0^s, \xi_1^s, \ldots, \xi_T^s)$, $\xi_t^s = (\xi_t^{s,1}, \ldots, \xi_t^{s,d})'$, $s = 1, \ldots, S$ is generated. Assuming that all scenarios coincide at $t = 0$, i.e. $\xi_0^1 = \ldots = \xi_0^S$, the initial root node is formed, and thus the simulated data paths are called as a scenario fan (Fig. 2). The structure of simulated data paths can be divided into two stages. The first stage is usually represented by a single root node, and the values of random parameters during the first stage are known with certainty. Moving to the second stage, the structure branches into individual scenarios at time $t = 1$, as shown in the Fig. 2. If such scenario fan is used as input in multistage stochastic program, the model is of 2-stage problem, as all σ -fields $\mathfrak{F}_t$, $t = 1, \ldots, T$ coincide. The 2-stage multiperiod stochastic program has the following properties, as in [8]:

1) Decisions at all time instances $t = 0,1,\ldots,T$ are made at once and no further information is expected.
2) Except for the first stage no nonanticipativity constraints appear.

Depending on the considered problem, such properties can be regarded as disadvantages. Our aim is to create a multistage scenario tree which can be used for multistage models. Multistage formulation is characterized by its robustness, stability of solutions: similar subscenarios result in similar decisions. The multistage tree reflects the interstage dependency and decreases the number of nodes while comparing to the scenario fan. The structure of multistage tree at $t = 0$ is also described by a sole root node and by branching into a finite number of scenarios as it was in previous case. The nodes further down represent the events of the world which are conditional at second stage. The arcs linking nodes represent various realizations of random variables. This branching continues for $t < T$, resulting the multistage tree (Fig. 3).

The distinction between stages, which correspond to the

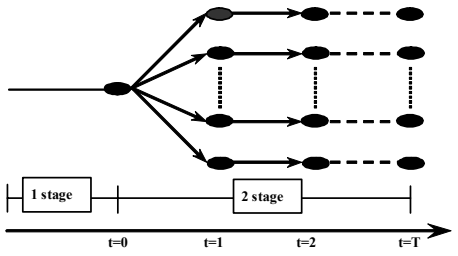

Fig. 2. Scenario fan

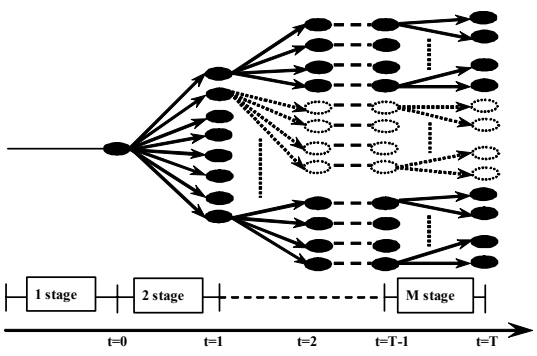

Fig. 3. Multistage scenario tree

decision moments, and time periods is essential, because in practical application it is important that the number of time periods would be greater than the corresponding nodes.

The algorithm of transforming the scenario fan to multistage scenario tree is described in the next section.

III. K-Means Clustering: Path to Tree

While constructing the multistage scenario tree from the scenario fan, the fan of individual scenarios is modified by bundling scenarios based on the cluster analysis. It is assumed that a set of individual scenarios for the entire time horizon is already generated. The idea of bundling the scenarios to the clusters is depicted in the Fig. 4. The scenario fan of 11 scenarios is schematically illustrated. At time $t = 0$ all these scenarios (which are the same) form the root of the tree. Next, two clusters are formed by the first iteration of some clustering algorithm. It results that we have six and five scenarios in each cluster. The centers of each cluster are computed, which represent the one-level nodes at time $t = 1$. Two black points denote the nodes corresponding to the conditional decisions. The formed clusters are then divided into subclusters in the next time period $t = 2$. We have four, two, three and two paths in each cluster, representing two-level nodes, since the centers are calculated. These nodes are denoted by four black points in the scenario tree. Such strategy of bundling scenarios to the clusters continues till the end of time horizon is reached. Joining the black points by line, the scenario tree structure is obtained.

The discussed technique allows to produce the tree with such characteristics:

1) The projection of random variable nearer the time horizon is less critical than those for the near future, because number of scenarios S grows smaller down the tree and the centers that represent the scenario cluster are calculated from a smaller sample size.
2) It allows to model extreme events because at every stage the simulated scenarios in all of the clusters are not discarded, and at the next stage all simulated scenarios in all of the clusters are used to calculate the centre of cluster.

In the first step of clustering scenarios we need to delineate the initial structure of scenario tree: the number of stages and the branching scheme. Some criteria for bundling the scenarios is chosen. The scenario fan usually consists of large number of scenarios, that's why the hierarchical methods can fail. We don't also require the method that in finding the clusters would be optimal by some measures. In the literature, the cluster methods usually are used for stable data. We should make some modifications in order to cluster the time dependent data. Let assume that K branches are desired from each scenario tree node. It means that K clusters will need to be formed. After such consideration, the K-means clustering algorithm [12] is chosen to construct the scenario tree from the set of simulated paths. Clustering consists in partitioning of a data set into subsets, so that the data in each cluster share the common attribute. This similarity is often defined by some distance measure. After a discussion of the kind of requirements we are using, we describe the modified K-means clustering algorithm.

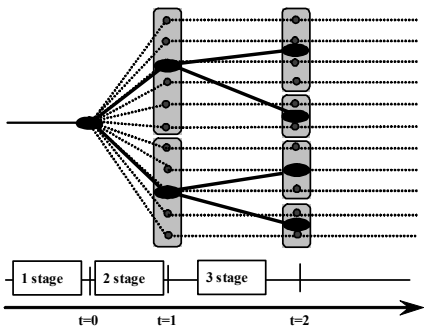

Fig. 4. Illustration of 3-stage tree construction

Given a fan of individual scenarios $\xi^s = \left(\xi_0^s, \xi_1^s, \ldots, \xi_T^s\right)$, $s = 1, \ldots, S$ and the number K of desired clusters $\widetilde{C}^1, \ldots, \widetilde{C}^K$, it is needed to find the cluster centers $\overline{\xi}^k$, $k = 1, \ldots, K$ such that the sum of the 2-norm distance squared between each scenario ξ^s and its nearest cluster center $\overline{\xi}^k$ is minimized:

$$\sum_{k=1}^{K} \sum_{\xi^s \in \widetilde{C}^k} \left\| \xi^s - \overline{\xi}^k \right\|_2^2 \to \min.$$

The modified K-means clustering algorithm is given as follows. At the beginning, the decision moments are set, corresponding to the stage index $t \in (1, \ldots, T)$. Then iterate:

Step 1: *Setting initial centers*. Let $\overline{\xi}^k$, $k = 1, \ldots, K$ be the cluster centers, which might be chosen to be the first K scenarios, since the scenarios are independently generated.

Step 2: *Cluster assignment*. For each scenario ξ^s, assign ξ^s to the cluster $\widetilde{C}^k$, such that center $\overline{\xi}^k$ is nearest to ξ^s in the 2-norm, which is modified to exploit the whole sequence of simulated data path:

$$d\left(\xi^s, \overline{\xi}^k\right) = \sum_{i=0}^{T} \left\| \xi_i^s - \overline{\xi}_i^k \right\|_2.$$

Step 3: *Cluster update*. Compute $\overline{\xi}^k$ as the mean of all scenarios assigned to the cluster $\widetilde{C}^k$:

$$\overline{\xi}^k = \mathbb{E}\left\{\xi^s\right\}_{\xi^s \in \widetilde{C}^k}.$$

This formula can be replaced by other estimate, such as median, mode or else.

Step 4: *Repeat.* Go to Step 2 until convergence, i.e. no scenario move group.

Step 5: *Calculation of probabilities.* Probability of $\overline{\xi}^k$ is equal the sum of probabilities of the individual scenarios ξ^s, belonging to the relevant cluster $\widetilde{C}^k$.

Step 6: *Modification.* Modify $\xi^s = \left(\xi_0^s, \xi_1^s, \ldots, \xi_T^s\right)$ by replacing ξ_t^s with $\overline{\xi}^k$ if $\xi_t^s \in \widetilde{C}^k$.

Step 7: *Repeat*. Go to Step 1 if next stage index exists. The clustering procedure starts over using each of clusters formed in current iteration.

This produces a separation of scenarios into groups. The given algorithm lets to treat properly the interstage dependencies, exploiting the whole sequence of simulated scenario path. At the end, the scenario tree is constructed, consisting of nodes $\overline{\xi}_k$ with their probabilities and the branching scheme.

IV. Computational Experiment

The scenario tree generation approach is applied to construct scenario trees out of sampled scenarios provided by Hibbert, Mowbray and Turnbull (HMT) stochastic asset model [13]. We use this model to generate the data, which consist of a finite number of scenarios, representing realizations of discount (zero-coupon) bond yields.

In HMT model presented here, the underlying movements in inflation and real interest rates generate the process for nominal interest rates. The model produces the term structure that has closed-form solutions for bond prices so that the entire term structure for any future projection date can be quickly generated. A cascading structure is a characteristic of scenario generator: real interest and inflation rates are simulated, which then, depending on the relationship structure assumed, influence the realization of discount bond yields (Fig. 5). More details about the HMT model can be found in work [13]. In scenario generator the financial variables have to be projected in such way as to reflect the appropriate interdependencies between them. It is reasonable to consider the case when interest rates and inflation rates move together. Interdependencies between these variables are identified through the copula-based dependency measure, discussed in Section II.

In the present paper, two different dependence structures – Gaussian copula and Student's t-copula – are employed to model dependencies between real interest rate and inflation rate. The Gaussian copula is given by

$$\widetilde{C}_{Cor}^{Ga}(u) = \Phi_{Cor}^{d}\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right),$$

where Φ_{Cor}^{d} denotes the joint distribution function of the d-variate standard normal distribution function with matrix Cor of linear correlation coefficient, $\Phi^{-1}(u)$ denotes the inverse of univariate standard normal distribution function. The main property of such dependence structure is that Gaussian copula does not have neither upper nor lower tail dependence. The Student's t-copula allows for joint fat tails and increases the probability of joint extreme events, comparing it with the above-described Gaussian copula. Student's t-copula can be written as

$$\widetilde{C}_{Cor,v}^{t}(u) = t_{Cor,v}^{d}\left(t_v^{-1}(u_1), \ldots, t_v^{-1}(u_d)\right),$$

where Cor, v are the parameters of t-copula, $t_{Cor,v}^{d}$ denotes the joint distribution function of the d-variate Student's t-distribution with v degrees of freedom, t_v^{-1} is the inverse of univariate Student's t-distribution with v degrees of freedom. Student's t-copula has the additional parameter v comparing with Gaussian copula. Increasing the value of v decreases the tendency to discover extreme co-movements.

At this moment, let assume that a matrix $Cor^{\tau} = \left[cor_{ij}^{\tau}\right]$, $-1 \le cor_{ij}^{\tau} \le 1$, $i, j = \overline{1, d}$ of Kendall's tau correlations has already been assessed, which denotes rank-order correlations between two random variables. In HMT model, Kendall's tau correlation coefficient is set equal to 0.25 between short-term real interest rate and short-term inflation rate, equal to 0.25 between long-term real interest rate and long-term inflation rate. While simulating the dependent variables in HMT model, the copula function is employed when $d = 4$: there are two stochastic variables that are described by two-factor model. Simulation procedure for Gaussian copula and Student's t-copula is performed as follows: (a) convert Kendall's tau cor_{ij}^{τ} to linear correlation coefficient cor_{ij} using formula $cor_{ij} = \sin\left(\pi\, cor_{ij}^{\tau} / 2\right)$ and construct the lower triangular matrix $\mathrm{A} = \left\lfloor a_{ij} \right\rfloor$ that holds $Cor = AA'$, (b) generate independent standard normal variables ε_i, $i = \overline{1, d}$ and form a column vector ε, (c) generate a random variate $\xi \sim \chi_v^2$, (d) construct joint probability density function, taking matrix product $\widetilde{\varepsilon} = A\varepsilon$, (e) calculate $\widetilde{\widetilde{\varepsilon}} = \sqrt{v}\widetilde{\varepsilon} / \sqrt{\xi}$, (f) set $\widetilde{u}_i = \Phi\left(\widetilde{\varepsilon}_i\right)$ and $\widetilde{\widetilde{u}}_i = t_v\left(\widetilde{\widetilde{\varepsilon}}_i\right)$, (g) set $\widetilde{x}_i = \Phi^{-1}\left(\widetilde{u}_i\right)$ and $\widetilde{\widetilde{x}}_i = \Phi^{-1}\left(\widetilde{\widetilde{u}}_i\right)$. At the result, $\widetilde{x}_i$ are correlated variables based on Gaussian copula, and $\widetilde{\widetilde{x}}_i$ are correlated variables based on Student's t-copula.

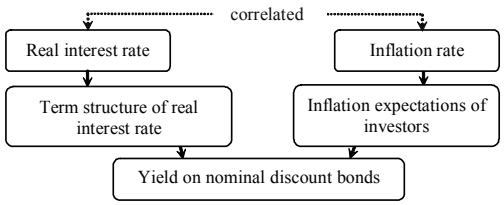

Fig. 5. A cascade structure of HMT model

HMT model is used to simulate 1, 3, 5, 7, 10 year coupon bond yields over a horizon of 20 years with time increments of one month. The initial parameters are set with the reference to the Hibbert's et al. work. The conditions about the environment are assumed as follows: inflation level is 2.5%, long-term inflation level is 2.83%, current 3-month T-bill norm is 5% and current 10-year T-bond yield 5.58%. The lower bounds on the levels of inflation and real interest rates are placed to ensure that negative rates don't appear. At the output of this scenario

Table I: Dimension of scenario fan

	Nodes	Time periods	Scenarios
Scenario fan of discount bond yields	240000	240	1000

Table II: Dimension of scenario trees

	K=2		K=3	
	Nodes	Scenarios	Nodes	Scenarios
3-stage tree	7	4	13	9
5-stage tree	31	16	121	81

generator the data consisted of a finite number of scenarios ($S = 1000$), representing the realizations of discount bond yields. The dimension of the scenario fan is given in Table I. Such scenario fan is aimed to transform to the scenario trees with different number of stages and with different branching factor, employing the clustering algorithm discussed in Section III. The number of stages depends on the number of decision moments. The branching scheme of scenario tree depends on the number of clusters. For instance, we choose the number of scenarios equal to $K = 2$ and $K = 3$ which is generated per scenario tree node. Two types of scenario trees are generated for the analysis: 3-stage scenario tree with decisions at $t = 10, 20$ and 5-stage scenario tree with decisions at $t = 5, 10, 15, 20$. Table II shows the dimension of scenario trees for the cases $K = 2$ and $K = 3$. Table I and Table II show that the dimension of scenario fan is notably reduced while transforming the scenario fan to scenario tree.

In the analysis, we aim to investigate how dependence structure affects the values of target variables and the structure of scenario tree. For instance, we consider 1-year and 10-year coupon bond yields. Some of statistical characteristics, the mean value and the dispersion, of discount bond returns are calculated for the evaluation of generated scenario trees. Table III – Table VI provide the obtained numerical results.

Table VII – Table VIII display the mean value and the dispersion calculated from the scenario fan at defined time moments. It turns out that for a larger branching factor K, the data of discount bond returns become more diverse, but the mean value remains the same. It holds for both Gaussian copula and Student's t-copula based dependence structures. The more stage number is set, the bigger dispersion is obtained.

Table III: Characteristics of scenario trees when $K = 2$

Gaussian Dependence			Decision moments, in Years			
			t=5	t=10	t=15	t=20
1Y coupon bond return, %	2-stage tree	Mean	–	7.380	–	8.880
		Dispersion	–	0.041	–	0.078
	5-stage tree	Mean	6.116	7.380	8.313	8.880
		Dispersion	0.007	0.050	0.096	0.104
10Y coupon bond return, %	2-stage tree	Mean	–	7.892	–	8.999
		Dispersion	–	0.036	–	0.057
	5-stage tree	Mean	6.877	7.892	8.601	8.999
		Dispersion	0.008	0.045	0.074	0.078

Table IV: Characteristics of scenario trees when $K = 3$

Gaussian dependence			Decision moments, in Years			
			t=5	t=10	t=15	t=20
1Y coupon bond return, %	2-stage tree	Mean	–	7.380	–	8.880
		Dispersion	–	0.047	–	0.094
	5-stage tree	Mean	6.116	7.380	8.313	8.880
		Dispersion	0.008	0.065	0.103	0.117
10Y coupon bond return, %	2-stage tree	Mean	–	7.892	–	8.999
		Dispersion	–	0.041	–	0.073
	5-stage tree	Mean	6.877	7.892	8.601	8.999
		Dispersion	0.009	0.054	0.080	0.091

Table V: Characteristics of scenario trees when $K = 2$

Student's t-dependence			Decision moments, in Years			
			t=5	t=10	t=15	t=20
1Y coupon bond return, %	2-stage tree	Mean	–	7.178	–	8.695
		Dispersion	–	0.052	–	0.084
	5-stage tree	Mean	6.184	7.178	8.063	8.655
		Dispersion	0.011	0.069	0.108	0.113
10Y coupon bond return, %	2-stage tree	Mean	–	7.677	–	8.858
		Dispersion	–	0.040	–	0.059
	5-stage tree	Mean	6.879	7.677	8.334	8.858
		Dispersion	0.011	0.048	0.081	0.080

Table VI: Characteristics of scenario trees when $K = 3$

Student's t-dependence			Decision moments, in Years			
			t=5	t=10	t=15	t=20
1Y coupon bond return, %	2-stage tree	Mean	–	7.178	–	8.695
		Dispersion	–	0.064	–	0.105
	5-stage tree	Mean	6.184	7.178	8.063	8.655
		Dispersion	0.015	0.081	0.119	0.134
10Y coupon bond return, %	2-stage tree	Mean	–	7.677	–	8.858
		Dispersion	–	0.049	–	0.074
	5-stage tree	Mean	6.879	7.677	8.334	8.858
		Dispersion	0.015	0.061	0.087	0.095

Table VII: Characteristics of scenario fan

Gaussian dependence		Decision moments, in Years			
		t=5	t=10	t=15	t=20
Scenario fan of 1Y coupon bond return, %	Mean	6.116	7.380	8.313	8.880
	Dispersion	0.044	0.087	0.118	0.139
Scenario fan of 10Y coupon bond return, %	Mean	6.878	7.892	8.601	8.999
	Dispersion	0.042	0.072	0.093	0.109

Table VIII: Characteristics of scenario fan

Student's t dependence		Decision moments, in Years			
		t=5	t=10	t=15	t=20
Scenario fan of 1Y coupon bond return, %	Mean	6.184	7.178	8.063	8.695
	Dispersion	0.060	0.106	0.135	0.159
Scenario fan of 10Y coupon bond return, %	Mean	6.879	7.678	8.334	8.858
	Dispersion	0.051	0.081	0.101	0.115

The initial fan of individual scenarios and the constructed scenario trees show that the data obtained under Student's t-dependence are more diverse, showing the higher risk, and have smaller mean value than data obtained under Gaussian dependence.

The scenario fan is illustrated using the "funnel of doubt" plot, resulting from uncertainty in the future values. In Fig. 6, the "funnel of doubts" graph displays the 1st, 5th, 25th, 50th, 75th, 95th, 99th percentile values and the mean sample value (light dashed line). The spread around its median expands as the time increases, carrying a certain risk of uncertainty that increases with time, but tends to stabilize at the end of time horizon, which is the effect of mean reversion value. The assumption of avoiding negative values of nominal interest rate determines that the expected value of discount bond yields drifts up over time. VaR (Value-at-Risk) type conclusion is that in $100(1-p)\%$ of the cases the yield is higher or equal to VaR_p value (vertical axis), where $0 < p < 1$ is a percentile value. The spread of 10-year discount bond yields is less than the spread of 1-year discount bond yields, because of the effect of mean reversion. More visual representations on sampled data paths can be found in the work [14]. Let's analyze the scenario trees generated from scenario fan of discount bond yields (Fig. 7–Fig. 10).

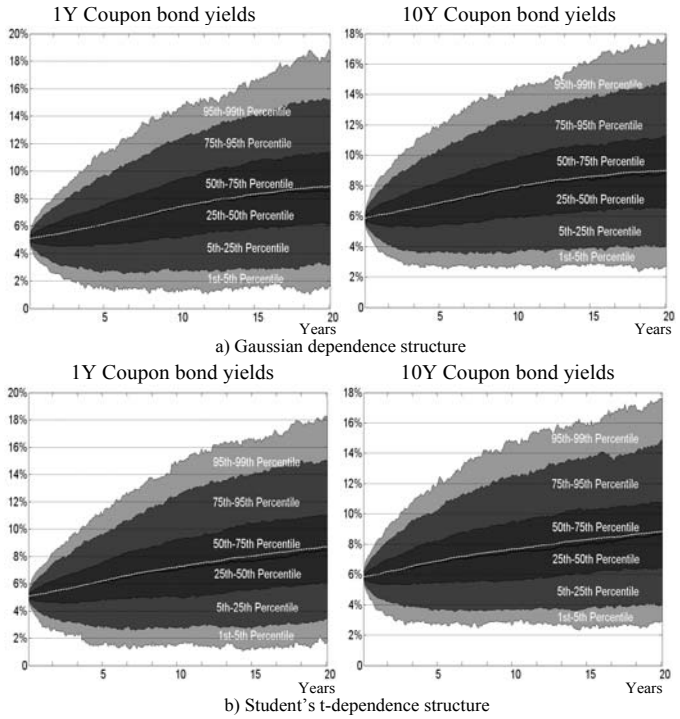

Fig. 6. Scenario fan of 1Y and 10Y Coupon bond yields

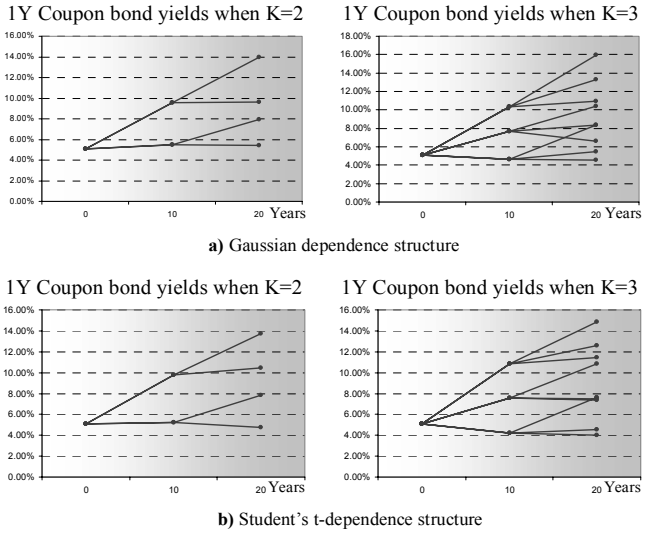

Fig. 7. 3-stage scenario trees of 1Y Coupon bond yields with decisions at t={10,20} years

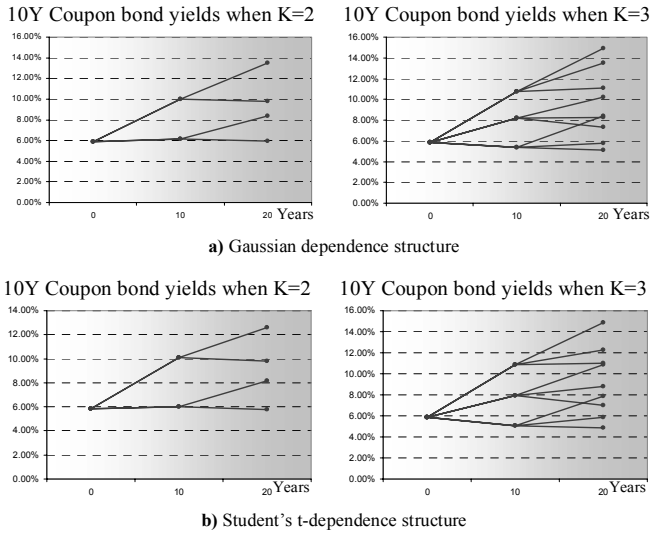

Fig. 8. 3-stage scenario trees of 10Y Coupon bond yields with decisions at t={10,20} years

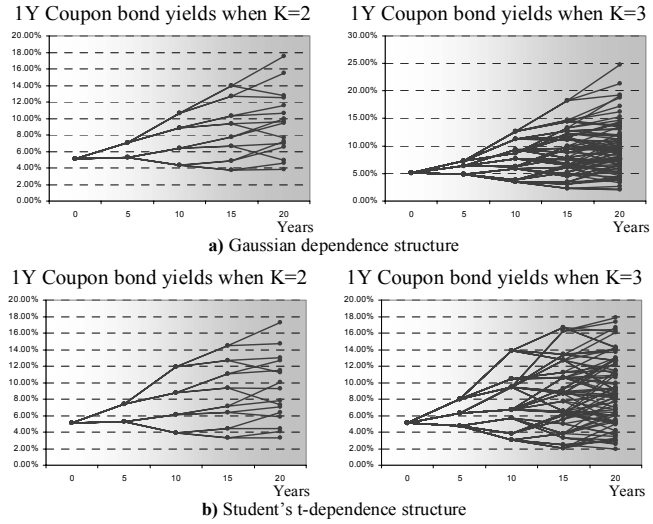

Fig. 9. 5-stage scenario trees of 1Y Coupon bond yields with decisions at t={5,10,15,20} years

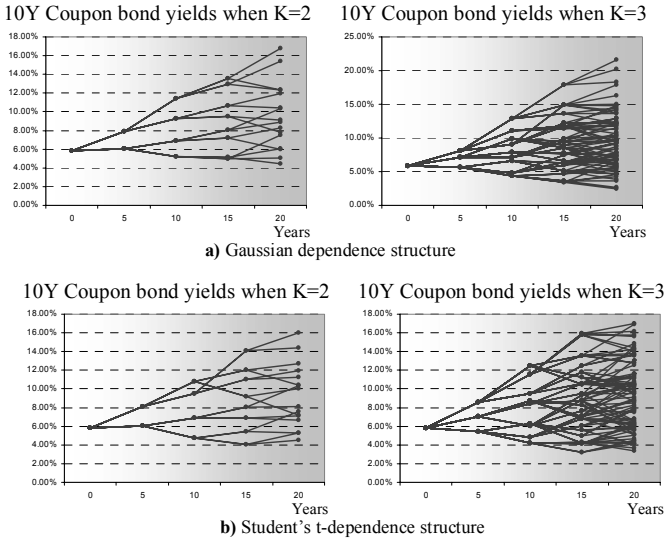

Fig. 10. 5-stage scenario trees of 10Y Coupon bond yields with decisions at t={5,10,15,20} years

Scenario tree with a higher branching factor lets to model more extreme scenarios. Using of Student's t-copula as dependence measure between real interest rate and inflation rate has effect to obtain more diverse values of discount bond yields comparing with the case when the Gaussian copula is used.

V. Concluding Remarks

In the present paper, we described the procedure based on both simulation and clustering to generate the scenario trees out of data paths. The computational experiment showed that the size of generated scenario trees is much smaller than the dimension of scenario fan, and nevertheless, they are good approximations with respect to the Euclidean distance used to measure the time-dependent data paths. Answering to our question, does the copula features are captured in the approximate representation of uncertainty in the form of scenario tree, we conclude that different dependence structures with the same correlation coefficient between stochastic variables affect the structure of multistage scenario tree. The graphic representation of scenario trees shows that scenario trees generated from dependent data based on Student's t-copula are more chaotic than generated from dependent data employing Gaussian copula. The effect of using Student's t-copula as dependence measure between real interest rate and inflation rate is to decrease value of discount bond yields. It results from the feature that using Gaussian copula the extreme events are independent, so we don't get really extreme scenarios. In the future, the constructed scenario trees will be used as an input to the multistage stochastic program.

References

[1] L. Y. Yu, X. D. Ji, and S. Y. Wang, "Stochastic programming models in financial optimization: survey," *AMO – Advanced Modeling and Optimization*, vol. 5(1), 2003, pp. 1–26.

[2] J. Dupačová, J. Hurt, and J. Štěpán, *Applied Optimization 75: Stochastic Modeling in Economics and Finance*. Dordrecht, Holland: Kluwer Academic Publishers, 2002, ch. 3.

[3] S. Mitra, "Scenario generation for stochastic programming," White Paper, Optirisk Systems, UK, 2006.

[4] J. Dupačová, N. Gröwe-Kuska, and W. Römisch, "Scenario reduction in stochastic programming," *Mathematical Programming*, vol. 95(3), 2003, pp. 493–511.

[5] K. Høyland and S. W. Wallace, "Generating scenario trees for multistage decision problems," *Management Science*, vol. 47(2), 2001, pp. 295–307.

[6] H. Heitsch and W. Römisch, „Generation of multivariate scenario trees to model stochasticity in power management," in *IEEE St. Petersburg PowerTech Proceedings*, Russia, 2005.

[7] G. Pflug, "Scenario tree generation for multiperiod financial optimization by optimal discretization," *Mathematical Programming*, vol. 89, 2001, pp. 251–271.

[8] J. Dupačová, G. Consigli, and S. W. Wallace, "Scenarios for multistage stochastic programs," *Annals of Operations Research*, vol. 100, 2000, pp. 25–53.

[9] P. Embrechts, A. McNeil, and D. Straumann, „Correlation and Dependency in Risk Management: Properties and Pitfalls," in *Risk management: value at risk and beyond*, M. A. H. Dempster, Ed. UK: Cambridge University Press, 2002, pp. 176–224.

[10] K. Aas, "Modelling the dependence structure of financial assets: a survey of four copulas," Research Report, SAMBA/22/04, Norwegian Computing Center, Norway, 2004.

[11] N. D. Domenica, G. Birbilis, G. Mitra, and P. Valente, "Stochastic programming and scenario generation within a simulation framework: an information systems perspective," Technical Report, CARISMA, UK, 2003.

[12] L. Kaufmann and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Canada : Wiley-Interscience, 1990, ch. 3.

[13] J. Hibbert, P. Mowbray, and C. Turnbull, "A stochastic asset model & calibration for long-term financial planning purposes," Technical Report, Barrie&Hibbert Limited, UK, 2001.

[14] H. Pranevicius and K. Sutiene, "Simulation of dependence between asset returns in insurance," in *Proceedings of the 5th International Conference on Operational Research: Simulation and Optimization in Business and Industry*, Estonia, 2006, pp. 23–28.