SCENARIO TREE GENERATION BY CLUSTERING THE SIMULATED DATA PATHS

Henrikas Pranevicius
Kristina Sutiene
Department of Business Informatics
Kaunas University of Technology
Studentu str. 56-301, Kaunas 51424, Lithuania
E-mail: kristina.sutiene@stud.ktu.lt

KEYWORDS

Stochastic programming, Scenario generation, Scenario tree construction, K-means clustering.

ABSTRACT

Multistage stochastic programs are effective for solving long-term planning problems under uncertainty. Such programs are usually based on a scenario model of future environment developments. A good approximation of the underlying stochastic process may involve a very large number of scenarios and their probabilities. We discuss 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 the 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 in order to model the sequential decisions. The described scenario tree generation method is implemented on sampled data of nominal interest rate.

INTRODUCTION

The concept of scenarios is usually employed for the modeling of randomness in stochastic programming models (Yu et al. 2003; Dupačová et al. 2002), 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 sub-models for each individual stochastic parameter. These sub-models are then 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 the risky events are approximated by a discrete set of scenarios, or sequence of events. This process is known as scenario generation. There are a lot of scenario generation methods, see for example (Dupačová et al. 2003; Høyland and Wallace 2001; Heitsch and Römisch 2005; Pflug 2001; Høyland et al. 2003; Yu et al. 2003). They are based on different principles: conditional sampling, sampling from given marginals and correlations, moment matching, path based methods, optimal discretization.

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 path structure (Hibiki 2003). But the number of scenarios is limited by the available computing power. According to the complexity of stochastic model, the scenario tree structure is used to approximate the random process (Heitsch and Römisch 2005).

In the present paper we concentrate on the scenario generation when the underlying stochastic parameters have been determined and the data paths of their realizations can be generated. Then using the sampled paths, the scenario tree is constructed using the classifying method, such as clustering analysis. An approach similar to our work is introduced in the article (Dupačová et al. 2000), but without a detailed clustering algorithm. Due to this, the K-means clustering method was modified to cluster the data paths, capturing the interstage dependence. 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.

SCENARIO GENERATION FOR MULTISTAGE STOCHASTIC PROGRAMS

In general, the scenario generation consists of following steps (Domenica et al. 2003):

- Choosing the appropriate model to describe the stochastic parameters. For instance, Econometric

Proceedings 21st European Conference on Modelling and Simulation
Ivan Zelinka, Zuzana Oplatková, Alessandra Orsoni ©ECMS 2007
ISBN 978-0-9553018-2-7 / ISBN 978-0-9553018-3-4 (CD)

models and Time Series (Autoregressive models, Moving Average models, Vector Auto Regressive models), Diffusion Processes (Wiener Processes).

- Calibration of model parameters using historical data.

- 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.

- Constructing the scenario tree with the desired properties.

The aim of scenario generation is to create a tree structure of scenarios, which is input in stochastic model. Let introduce some notations used in stochastic programming.

The original multivariate stochastic data process $\xi = \{\xi_t\}_{t=0}^{T}$ is defined on some probability space $(\Omega, \mathcal{F}, P)$ with ξ_t taking values in some $\mathcal{R}^d$. In the stochastic programming model the observations and decisions are given as a sequence $x_0, (\xi_0, x_1), (\xi_1, x_2), \ldots, (\xi_{T-1}, x_T)$, 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 $\mathcal{F}_{t-1} \subseteq \mathcal{F}$. Then, following the (Dupačová et al. 2000), 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.

According to the third step of scenario generation process (described at the beginning of this Section), the process has to be discrete in time, i.e. $\mathbf{T} = \{0, \ldots, T\}$. The points in time $t \in \mathbf{T}$ are called as stage index. Then, the probability distribution of ξ is replaced by a finite support, i.e. a finite number of realizations ξ^s, $s = 1, \ldots, S$ with probability $p_s = P(\xi^s)$, $p_s \geq 0$ and $\sum_{s=1}^{S} p_s = 1$. The set of such realizations $\xi^s = (\xi_0^s, \xi_1^s, \ldots, \xi_T^s)$, $s = 1, \ldots, S$ is called as simulated data paths, or as scenario fan (see Figure 1), if we assume that all scenarios coincide at the first time period $t = 0$, i.e. $\xi_0^1 = \ldots = \xi_0^S$, and form the initial root node. 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 Figure 1.

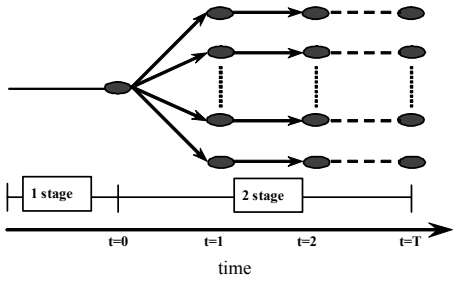

Figure 1: Scenario Fan

If such scenario fan is used as input in multistage stochastic program, the model is of 2-stage problem, as all σ-fields $\mathcal{F}_t$, $t = 1, \ldots, T$ coincide. The properties of 2-stage multiperiod stochastic program are (Dupačová et al. 2000):

- Decisions at all time instances $t = 0, 1, \ldots, T$ are made at once and no further information is expected.

- Hedging against all considered unrelated scenarios of possible developments is assumed.

- 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 (see Figure 2) which can be used for multistage models.

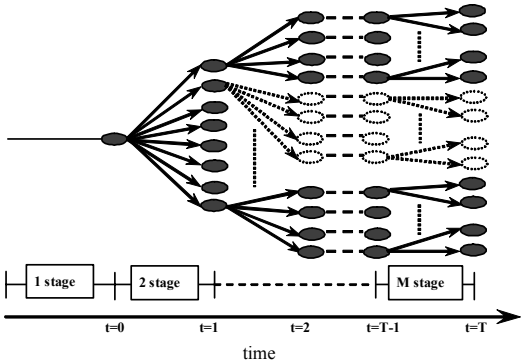

Figure 2: Multistage Scenario Tree

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 the nodes represent various realizations of random variables. This branching continues for $t < T$, resulting the multistage tree.

The distinction between stages, which correspond to the decision moment, and time periods is essential, because it is important in practical application 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.

K-MEANS CLUSTERING: PATH TO TREE

To construct 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 objects in such set are scenarios with dimension equal to time horizon T. The idea of bundling the scenarios to the clusters is depicted in the Figure 3.

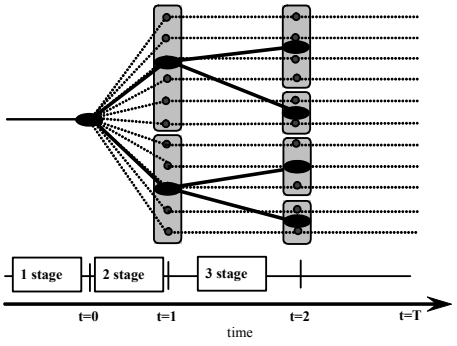

Figure 3: Illustration of 3-stage Tree Construction

In the Figure 3, the 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 sub-clusters 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 visual scenario tree structure is obtained. The discussed technique allows to produce the tree with such characteristics:

- The projection of random variable nearer the time horizon are 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.

- 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.

Before starting to implement this clustering idea, we need to define the initial structure of scenario tree: the number of stages and the branching scheme. Besides, we need to choose some criteria for bundling the scenarios. 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 have to perform 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 (Kaufmann and Rousseeuw 1990) is chosen for constructing the scenario tree from the set of simulated paths. Clustering consists in partitioning of a data set into subsets (clusters), so that the data in each cluster share the common attribute. This similarity is often defined by some distance measure. Next, we will give the formulation of K-means clustering problem.

Given a fan of individual scenarios $\xi^s = \left(\xi_0^s, \xi_1^s, \ldots, \xi_T^s\right), \quad s = 1, \ldots, S$ and the number K of desired clusters $C^1, \ldots, C^K$, 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, i.e.

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

The K-means clustering algorithm, after making it suitable for data paths, is given as follows. At the beginning the decision moments are set, according 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 C^k, such that center

$\bar{\xi}^k$ is nearest to ξ^s in the 2-norm, which is modified to exploit the whole sequence of simulated data path

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

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

$$\bar{\xi}^k = \mathrm{E}\left\{\xi^s\right\}_{\xi^s \in C^k}$$

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 moves group.

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

Step 7: *Repeat*. Go to Step 1 if next stage index exists and employ this algorithm for each cluster individually. In the following iteration, the scenarios, obtained from Step 6, are clustered.

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 $\bar{\xi}_k$ with their probabilities and the branching scheme.

NUMERICAL EXAMPLE

In the analysis, Hibbert, Mowbray and Turnbull (HMT) stochastic asset model for long-term financial planning purposes (Hibbert et al. 2001) is used to simulate a representative set of scenarios of nominal interest rate. The role of scenario generator is to develop a model that posits plausible projections of future interest rate levels rather to explain the past movements in interest rates. This model is composed of a number of component parts that are driven by a set of stochastic drivers. In HMT model presented here, the underlying movements in inflation and real interest rates generate the process for nominal interest rates. It is also important that these variables have to be projected in such a 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 alternative method – copula-based dependency measure (Embrechts et al. 2002). In the paper (Pranevicius and Sutiene 2003) Gaussian copula and Student's t-copula are investigated to model contemporaneous dependencies between real interest rate and inflation rate. In the present paper we employ only the Gaussian copula with correlation coefficient $\rho = 0.25$. To generate the nominal interest rate, such conditions about the environment are assumed: 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 is 5.58%. At the output of this scenario generator the data consist of a finite number of scenarios ($S = 1000$), representing the realizations of a monthly nominal interest rate for a time horizon of 10 years. The dimension of the scenario fan of nominal interest rate is given in Table 1.

Table 1: Dimension of Scenario Fan

	Nodes	Time periods	Scenarios
Nominal interest rate	120000	120	1000

Such scenario fan is aimed to transform to scenario trees with different number of stages, employing the clustering algorithm discussed in Section "K-Means Clustering: Path to Tree". The number of stages depends on the number of decision moments. The branching scheme of scenario tree depends from the number of clusters. For instance, we choose $K = 2$ and $K = 3$ number of scenarios which should be generated per scenario tree node. The five types of scenario trees are generated for the analysis: 2-stage scenario tree with decision moment $t = 10$, 3-stage scenario tree with decision moments $t = 2,10$, 4-stage scenario tree with decision moments $t = 2,4,10$, 5-stage scenario tree with decision moments $t = 2,4,6,10$, 6-stage scenario tree with decision moments $t = 2,4,6,8,10$. Table 2 shows the dimensions of scenario trees for the cases with $K = 2$ scenarios and with $K = 3$ scenarios from each node.

Table 2: Dimensions of Scenario Trees

	K=2		K=3	
	Nodes	Scenarios	Nodes	Scenarios
2-stage tree	3	2	4	3
3-stage tree	7	4	13	9
4-stage tree	15	8	40	27
5-stage tree	31	16	121	81
6-stage tree	63	32	364	243

Table 2 shows that the bigger number of stages and more detailed branching scheme expand the size of constructed scenario tree. While transforming the scenario fan to scenario tree, the dimension of scenarios is notably reduced (see Table 1 and Table 2).

Some of statistical characteristics, the mean value and the dispersion, of nominal interest rate are calculated for the evaluation of generated scenario trees. These characteristics are computed at different time moments for each of scenario tree with different branching scheme. Table 3 and Table 4 provide the obtained results (values in percent): in the intersection of rows and columns the first number denotes the mean value of nominal interest rate, and the second number denotes the dispersion of nominal interest rate. Table 5 shows the mean value and the dispersion calculated from the scenario fan at defined time moments.

Table 3: Characteristics of Scenario Trees with $K = 2$

		Decision moments, in Years				
		t=2	t=4	t=6	t=8	t=10
Scenario trees	2-stage tree	_	_	_	_	7.274 0.041
	3-stage tree	5.331 0.001	_	_	_	7.274 0.055
	4-stage tree	5.331 0.001	5.784 0.016	_	_	7.274 0.062
	5-stage tree	5.331 0.001	5.784 0.016	6.286 0.043	_	7.274 0.069
	6-stage tree	5.331 0.001	5.784 0.016	6.286 0.043	6.744 0.064	7.274 0.072

Table 4: Characteristics of Scenario Trees with $K = 3$

		Decision moments, in Years				
		t=2	t=4	t=6	t=8	t=10
Scenario trees	2-stage tree	_	_	_	_	7.274 0.049
	3-stage tree	5.331 0.003	_	_	_	7.274 0.064
	4-stage tree	5.331 0.003	5.784 0.025	_	_	7.274 0.072
	5-stage tree	5.331 0.003	5.784 0.025	6.286 0.046	_	7.274 0.076
	6-stage tree	5.331 0.003	5.784 0.025	6.286 0.046	6.744 0.066	7.274 0.081

Table 5: Characteristics of Scenario Fan

	Time moments, in Years				
	t=2	t=4	t=6	t=8	t=10
Scenario fan	5.331 0.019	5.784 0.037	6.286 0.055	6.744 0.072	7.274 0.089

The scenario trees are built to approximate the scenario fan. Table 3 – Table 5 show that while building the scenario tree from scenario fan, the statistical properties, the mean value and dispersion, of data process are retained.

The structure of some constructed scenario trees are displayed graphically. Figure 4 and Figure 5 depict the 6-stage scenario tree with $K = 2$ and $K = 3$ branching structure.

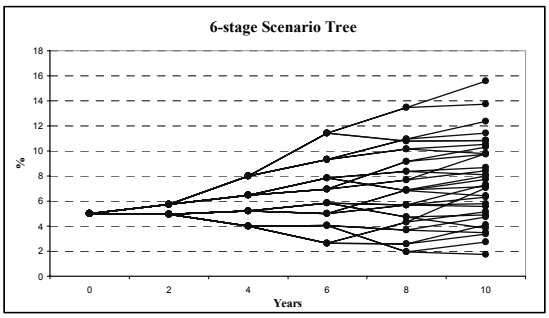

Figure 4: Multistage Scenario Tree with 2 Branches from Each Node

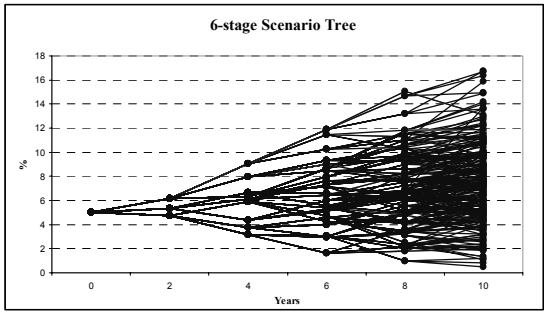

Figure 5: Multistage Scenario Tree with 3 Branches from Each Node

CONCLUDING REMARKS

In the present paper, we described the procedure based on simulation and clustering to generate the scenario tree from simulated paths. It was shown that the constructed trees are much smaller than the given scenario fans, and nevertheless, they are good approximations with respect to the Euclidean distance used to measure the data paths. In the future, the constructed scenario tree will be used as an input to a decision model. Besides, in the scenario tree the effect of using different copulas as dependence structure will be investigated.

REFERENCES

Domenica, N.D.; Birbilis, G.; Mitra, G. and Valente, P. 2003. Stochastic Programming and Scenario Generation within a Simulation Framework: an Information Systems Perspective." Technical Report. CARISMA, Brunel University, UK.

Dupačová, J.; Gröwe-Kuska, N. and Römisch, W. 2003. "Scenario Reduction in Stochastic Programming." *Mathematical Programming 95(3),* 493–511.

Dupačová, J.; Hurt, J. and Štěpán, J. 2002. *Applied Optimization 75: Stochastic Modeling in Economics and Finance.* Kluwer Academic Publishers.

Dupačová, J.; Consigli, G. and Wallace, S.W. 2000. "Scenarios for Multistage Stochastic Programs." *Annals of* Operations *Research 100,* 25–53.

Embrechts, P.; McNeil, A. and Straumann, D. 2002. „Correlation and Dependency in Risk Management: Properties and Pitfalls". In *Risk management: value at risk and beyond,* M.A.H. Dempster (Eds.), 176–224.

Heitsch, H. and Römisch, W. (2005). „Generation of Multivariate Scenario Trees to Model Stochasticity in Power Management". IEEE St. Petersburg Power Tech.

Hibbert, J.; Mowbray, P. and Turnbull, C. 2001. "A stochastic Asset Model & Calibration for Long-Term Financial Planning Purposes". Technical Report. Barrie&Hibbert Limited.

Hibiki, N. 2003. "A Hybrid Simulation / Tree Stochastic Optimization Model for Dynamic Asset Allocation". In *Asset and Liability Management Tools*, B. Scherer, (Eds.), 279–304.

Høyland, K.; Kaut, M. and Wallace, S.W. 2003. "A Heuristic for Moment-matching Scenario Generation." *Computational optimization and Applications 24,* 169–185.

Høyland, K. and Wallace, S.W. 2001. "Generating Scenario Trees for Multistage Decision Problems." *Management Science 47(2),* 295–307.

Kaufmann, L. and Rousseeuw, P.J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley & Sons.

Pflug, G. 2001. "Scenario Tree Generation for Multiperiod Financial Optimization by Optimal Discretization." *Mathematical Programming B89,* 251–271.

Pranevicius, H. and Sutiene, K. 2006. "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.* Tallinn, Estonia, 23–28.

Yu, L.Y.; Ji, X.D. and Wang, S.Y. 2003. "Stochastic Programming Models in Financial Optimization: Survey." *AMO – Advanced Modeling and Optimization 5(1),* 1–26.

AUTHOR BIOGRAPHIES

HENRIKAS PRANEVICIUS, Kaunas, Lithuania. Professor, Dr.habil., the Head of the department of Business Informatics at Kaunas University of Technology. Author received his doctor degree at Kaunas Polytechnic Institute in 1970, habilitation at Ryga Electronic and Computer Technical Institute in 1984. His research interests include the formal specification, validation and simulation of complex systems, knowledge-based simulation, and development of numerical models of systems specified by Markov Processes. The results of investigations have been successfully applied in creating the computerized systems for specification, validation and simulation/modeling of computer network protocols, logistics and industrial systems. The theoretical background of investigation is Piece-Linear Aggregate formalism, which permits to use the formal specification for model development and behavior analysis.

Email: hepran@if.ktu.lt

Web:http://www.vdu.lt/staff/informatics/CVPDF/CV_Pranevicius_en.pdf

KRISTINA SUTIENE, Kaunas, Lithuania. PhD student in Business Informatics Department at Kaunas University of Technology. The title of her thesis is "Statistical evaluation of insolvency duration using simulation model of insurance activity". Student's research interests include statistical forecasting, stochastic modeling and simulation, operational research applied to business processes.

Email: kristina.sutiene@stud.ktu.lt

Web: http://www.stud.ktu.lt/~krisjan/