Scenario generation methods for discrete data

Mgr. Ondřej Komora

Supervisor: Ing. Vít Procházka, Ph.D.

Charles University
Faculty of Mathematics and Physics
Department of Probability and Mathematical Statistics

17.10.2024

Introduction to scenario generation

Scenario = potential realization of randomness

- Allows formulation of stochastic optimization problems

Scenario generation = process of creating scenarios out of data.

- Has impact on
    1. Computational complexity.
    2. Quality of solutions.

However...

- Scenario generation is difficult for discrete data.
- Also, there is a relative lack of research.

  $\Rightarrow$ The only easy-to-use method is sampling for discrete data.
- We propose a new easy-to-use copula-based alternative to sampling.
- We show this method outperforms sampling significantly.

Copula and Sklar's theorem

Copula

Copula is the distribution function of a random vector with uniform margins on interval $[0,1]$.

Sklar's theorem

Let F be a joint distribution function of random vector $X = (X_1, \ldots, X_n)$. Then there exists copula C such that for $t_1, \ldots, t_n \in \overline{\mathbb{R}}$ it holds

$$F(t_1, \ldots, t_n) = C\left(F_{X_1}(t_1), \ldots, F_{X_n}(t_n)\right).$$

Copula C is uniquely determined on $\times_{i=1}^{n} \operatorname{Ran} F_{X_i}$.

A copula-based method from [Kaut, 2014]

- Sklar's theorem allows us to model dependence structure (copula) and marginal distributions independently.
- Assume:
  1. We have input random vector $X = (X_1, \ldots, X_n)$.
  2. Copula C of X (or its estimate).
  3. We aim to generate S scenarios.
- Method works in two steps:
  1. Model copula C using so-called *copula sample*.
  2. Transform copula sample to reflect marginal distributions.

Step 1: Generate copula sample

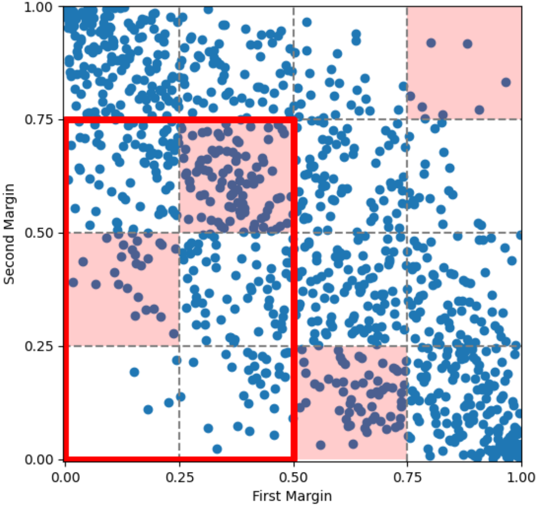
- Copula sample is defined as

$$\mathcal{C} := \{(r_1, \ldots, r_n) : 1 \leq r_i \leq S, \forall i \leq n\}$$

  where each value appears exactly once in each dimension.
- For a target copula C, we try to find copula sample $\mathcal{C}$ minimizing

$$\mathrm{dev}_{\mathrm{avg}}(\mathcal{C}, C) = \frac{1}{S^n} \sum_{r_1=1}^{n} \cdots \sum_{r_n=1}^{n} |\mathcal{C}_r(r_1, \ldots, r_n) - C_r(r_1, \ldots, r_n)|$$

Step 1: Generate copula sample



Step 2: Transform copula sample

- Assume we generated copula sample $\{(r_s^1, \dots, r_s^n) : s \in \{1, \dots, S\}\}$.
- We need to transform ranks r_s^i into reasonable values.

  $\Rightarrow$ Choose value x_s^i from region

$$\left[F_{X_i}^{-1}\left(\frac{r_s^i - 1}{S}\right), F_{X_i}^{-1}\left(\frac{r_s^i}{S}\right)\right].$$

- The options are
  1. Conditional median
  2. Conditional expectation
  3. And so on ...

Extension for discrete data

- Method is designed for continuous data.

  $\Rightarrow$ Fails to generate reasonable scenarios for discrete data.
- We demonstrate this on uniform distribution on $\{0, 1\}^2$.

  $\Rightarrow$ The algorithm produces only scenarios $(0, 0)$ and $(1, 1)$.
- Main idea: transform discrete variables into continuous ones.

Discrete extension

Assume that X with $\operatorname{supp} X \subseteq \mathbb{N}_0$ is a discrete random variable and U is a continuous random variable on $[0, 1]$ with strictly increasing distribution function on $[0, 1]$ which is independent of X. Then we define the extension of X as a random variable $X^* = X + U - 1$.

Illustration of a uniform discrete extension

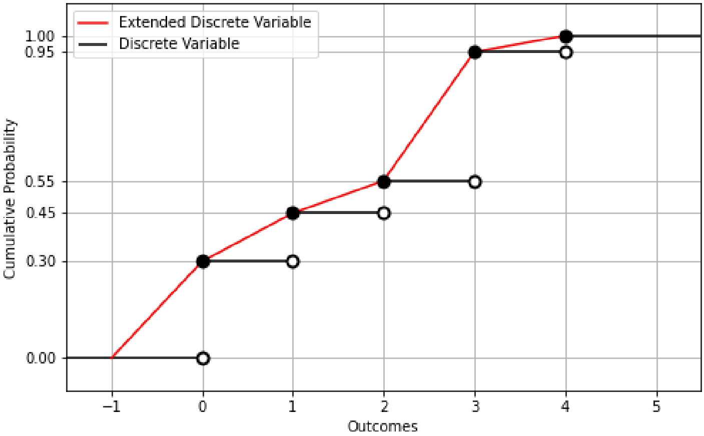


Figure: Comparison of a discrete distribution function and its uniform extension.

Step 1 revisited: Generate copula sample

We use the following procedure

1. Replace all discrete margins of X with their extensions.
2. Compute the copula C^* of the resulting vector. Call it *extension copula*.
3. Use this copula to generate a copula sample.

Properties of extension copula

Expression for extension copula:

$$C^*(u_1, \dots, u_k, v_1, \dots, v_p) = \sum_{S \subseteq \{1,\dots,p\}} C\left(u_1, \dots, u_k, v_1^S, \dots, v_p^S\right) \prod_{i \in S} \lambda_i(v_i) \prod_{j \notin S} (1 - \lambda_j(v_j)),$$

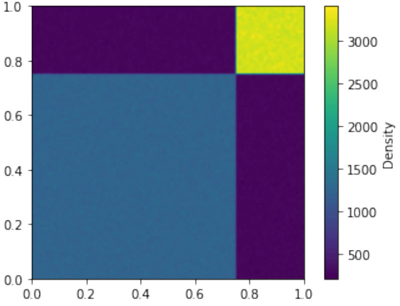
where

$$\lambda_i(v_i) = \begin{cases} \frac{v_i - v_i^-}{v_i^+ - v_i^-} & v_i \notin \text{Ran } F_{Y_i}, \\ 0 & v_i \in \text{Ran } F_{Y_i}, \end{cases}$$

Properties of C^*:

- According to Sklar's theorem, copulas are uniquely defined on $\times_{i=1}^n \text{Ran } F_{X_i}$.
- The extension copula linearly interpolates these points.
- It does not depend on the extension type!

Extension is a natural one



(a) Discrete margins. (b) Mixed margins.

Step 2 revisited: Transform copula sample

- Assume we generated copula sample $\{(r_s^1, \ldots, r_s^n) : s \in \{1, \ldots, S\}\}$.
- The algorithm replaces discrete variables X_i by their extensions X_i^*.

  $\Rightarrow$ We obtain regions

$$\left[F_{X_i^*}^{-1}\left(\frac{r_s^i - 1}{S}\right), F_{X_i^*}^{-1}\left(\frac{r_s^i}{S}\right)\right].$$

- Problems:
  1. Conditional expectation/median might be non-integral.
  2. Region might not contain any possible realization of X_i.
- Question: Into which realization of X_i transform ranks r_s^i?

On discrete transformation of copula samples

Identification of reasonable realizations

Let L_X be a function defined as

$$L_X(u) = \begin{cases} 0 & u = 0, \\ F_X^{-1}(u) + \mathbf{1}[u \in \operatorname{Ran} F_X] & u \in (0,1), \\ \sup(\operatorname{supp} X) & u = 1. \end{cases}$$

Then only for the realizations $n \in \operatorname{supp} X$ fulfilling

$$L_X\left(\frac{r_s^i - 1}{S}\right) \leq n \leq F_X^{-1}\left(\frac{r_s^i}{S}\right)$$

it holds

$$P\left(F_{X^*}^{-1}\left(\frac{r_s^i - 1}{S}\right) \leq X^* \leq F_{X^*}^{-1}\left(\frac{r_s^i}{S}\right) \middle| X = n\right) > 0.$$

Approach No.1 for discrete transformation

Select realization of X with the greatest contribution to

$$P\left(L_X\left(\frac{r_s^i - 1}{S}\right) \leq X \leq F_X^{-1}\left(\frac{r_s^i}{S}\right)\right).$$

This translates to problem

$$\begin{aligned} \max_{n \in \mathbb{N}_0} \quad & P(X = n) \\ \text{s.t.} \quad & L_X\left(\frac{r_s^i - 1}{S}\right) \leq n \leq F_X^{-1}\left(\frac{r_s^i}{S}\right). \end{aligned}$$

Approach No.2 for discrete transformation

Select realization of X with the greatest contribution to

$$P\left(F_{X^*}^{-1}\left(\frac{r_s^i - 1}{S}\right) \leq X^* \leq F_{X^*}^{-1}\left(\frac{r_s^i}{S}\right)\right).$$

This translates to problem

$$\begin{aligned} \max_{n \in \mathbb{N}_0} \quad & P\left(1 - n + F_{X^*}^{-1}\left(\frac{r_s^i - 1}{S}\right) \leq U \leq 1 - n + F_{X^*}^{-1}\left(\frac{r_s^i}{S}\right)\right) \cdot P(X = n) \\ \text{s.t.} \quad & L_X\left(\frac{r_s^i - 1}{S}\right) \leq n \leq F_X^{-1}\left(\frac{r_s^i}{S}\right). \end{aligned}$$

Approach No.3 for discrete transformation

If $\operatorname{supp} X$ is large, we have following options:

1. $\operatorname{med}\left(X \,\middle|\, L_X\left(\frac{r_s^i - 1}{S}\right) \le X \le F_X^{-1}\left(\frac{r_s^i}{S}\right)\right)$,
2. $\mathrm{E}\left[X \,\middle|\, L_X\left(\frac{r_s^i - 1}{S}\right) \le X \le F_X^{-1}\left(\frac{r_s^i}{S}\right)\right]$.

Case study: Stochastic knapsack

- The Knapsack problem is a traditional optimization problem.
- We make appearance of items and prices uncertain.
- Two versions of the problem:
  1. Uncertain appearances of items.
  2. Uncertain appearances of items and prices.
- Versions represent problems with discrete and mixed data.
- Two-stage stochastic problem:
  1. First stage: Decide if we try to put item into knapsack.
  2. Second stage: Item appears or not and prices are determined. Value of knapsack is calculated.

Problem formulation

- Model the appearance of items using scenario variables

$$q_j^s = \begin{cases} 1 & \text{if item } j \text{ appears in scenario } s, \\ 0 & \text{otherwise.} \end{cases}$$

- Problem formulation is

$$\begin{aligned} \max_{x_i,\, e_s} \quad & \sum_{s\in\mathcal{S}} p^s \left(\sum_{j=1}^{K} c_j x_j q_j^s - Q e_s \right) \\ \text{s.t.} \quad & \sum_{j=1}^{K} w_j x_j q_j^s \leq W + e_s \quad s \in \mathcal{S}, \\ & x_i \in \{0,1\} \qquad i = 1,\dots,K, \\ & e_s \geq 0 \qquad s \in \mathcal{S}. \end{aligned}$$

- If prices are uncertain, we replace c_j by their scenario values c_j^s.

Problem-oriented method

- Denote
  1. f objective function.
  2. $f(x, \eta)$ so-called *out-of-sample* evaluation. Represents "true" objective value.
  3. $f(x, \tau)$ so-called *in-sample* evaluation. Approximates $f(x, \eta)$.
- Is based on minimizing the discrepancy between in-sample and out-of-sample evaluations on a pool of heuristic solutions.
- Obtain scenario set τ by solving

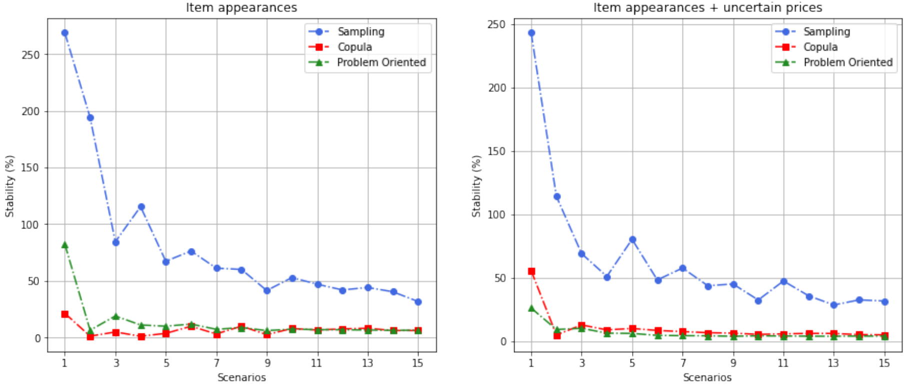
$$\min_{\tau} L(\tau; \mathcal{X}) := \sum_{x \in \mathcal{X}} (f(x, \tau) - f(x, \eta))^2 \cdot (\alpha \cdot \mathbf{1}[f(x, \tau) > f(x, \eta)] + \beta \cdot \mathbf{1}[f(x, \tau) < f(x, \eta)])$$

- See [Prochazka and Wallace, 2020] for more details.

In-sample stability

- Defined as

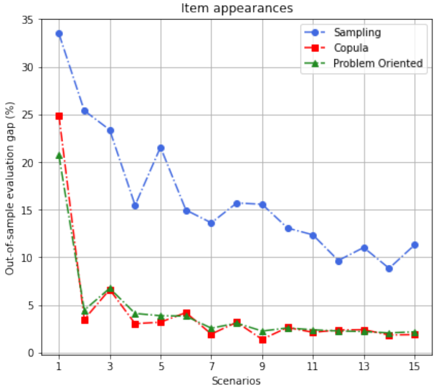
$$ST_n = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \frac{\max_{\tau \in \mathcal{T}_n} f(x,\tau) - \min_{\tau \in \mathcal{T}_n} f(x,\tau)}{\min_{\tau \in \mathcal{T}_n} f(x,\tau)}.$$

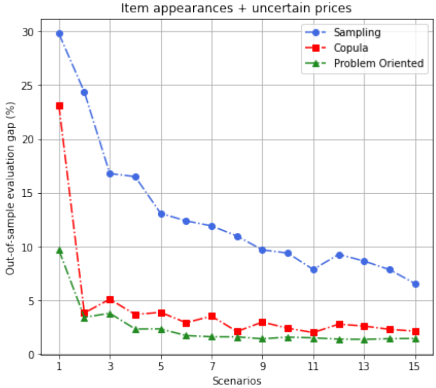


Out-of-sample evaluation gap

- Defined as

$$EG_n = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \sqrt{\frac{1}{K} \sum_{\tau \in \mathcal{T}_n} \left(\frac{f(x,\tau) - f(x,\eta)}{f(x,\eta)} \right)^2}.$$

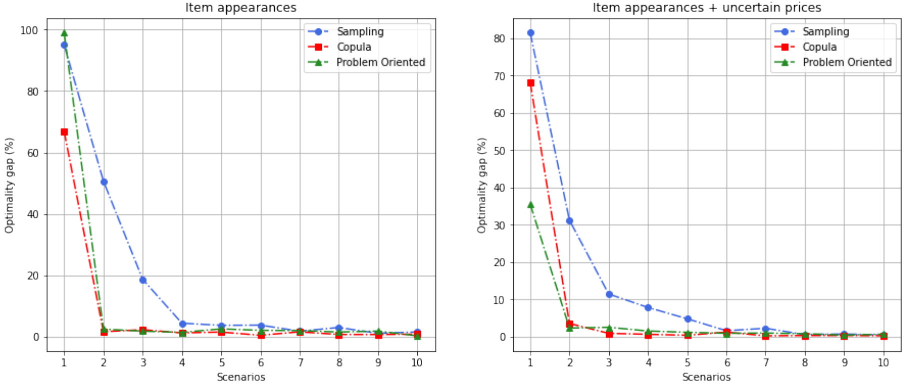




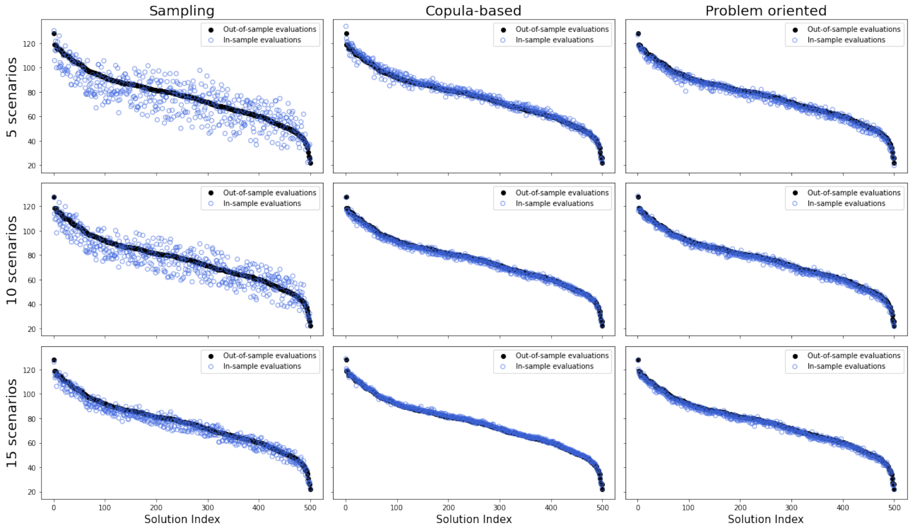
Optimality gap

- Defined as

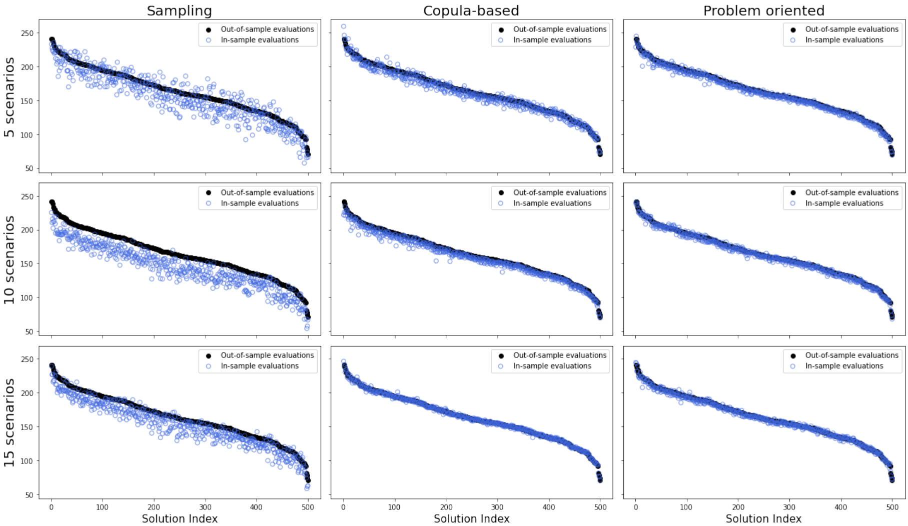
$$OG_n = \frac{1}{K} \sum_{\tau \in \mathcal{T}_n} \frac{f(x^*, \eta) - f(x^*_\tau, \eta)}{f(x^*_\tau, \eta)}.$$



Ranking visual assessment I.



Ranking visual assessment II.



Ranking assessment using Kendall's τ

Number of Scenarios	Sampling	Copula-based	Problem-oriented
5 scenarios	0.729	0.935	0.931
10 scenarios	0.750	0.952	0.953
15 scenarios	0.835	0.969	0.959
20 scenarios	0.898	0.968	0.961
25 scenarios	0.901	0.972	0.962

Table: Stochastic knapsack problem with uncertain item appearances.

Number of Scenarios	Sampling	Copula-based	Problem-oriented
5 scenarios	0.758	0.905	0.939
10 scenarios	0.817	0.945	0.954
15 scenarios	0.844	0.954	0.961
20 scenarios	0.896	0.960	0.958
25 scenarios	0.900	0.961	0.962

Table: Stochastic knapsack problem with uncertain item appearances and prices.

Conclusion

We conclude the analysis as follows

- Method outperforms sampling significantly.
- Method is comparable with some problem-oriented methods.
- However, problem-oriented methods are difficult to develop.
- Meanwhile the proposed method is easy to use.

Contributions

Contributions of our thesis:

1. A new method for generating scenarios for discrete data. Namely
   - Use of extension copula in method from [Kaut, 2014].
   - New approaches to the transformation of discrete margins.
2. Illustrational examples.
   - Demonstration of why the unextended method fails.
   - Motivating the use of extension copula.
3. Extension copula for mixed random vectors.
   - Generalization of extension copula for mixed random vectors.
   - Derivation of the generalized form (based on [Denuit and Lambert, 2005]).

References

Denuit, M. and Lambert, P. (2005).
Constraints on concordance measures in bivariate discrete data.

Genest, C. and Nešlehová, J. (2007).
A primer on copulas for count data.

Genest, C., Nešlehová, J. G., and Rémillard, B. (2014).
On the empirical multilinear copula process for count data.

Kaut, M. (2014).
A copula-based heuristic for scenario generation.

Prochazka, V. and Wallace, S. W. (2020).
Scenario tree construction driven by heuristic solutions of the optimization problem.

Research ideas

- Scenario generation for discrete data for two-stage and multi-stage problems.
- Ideas:
  - Relax discrete distributions to continuous ones (discrete extensions or use continuous scenarios to describe discret ones)
  - Adjust methods using Wasserstein distance for discrete data