Depend. Model. 2020; 8:70–92 DE GRUYTER

Research Article Open Access

Viktor Kuzmenko, Romel Salam, and Stan Uryasev*

Checkerboard copula defined by sums of random variables

https://doi.org/10.1515/demo-2020-0004
Received February 19, 2019; accepted March 2, 2020

Abstract: We consider the problem of finding checkerboard copulas for modeling multivariate distributions. A *checkerboard* copula is a distribution with a corresponding density defined almost everywhere by a step function on an m-uniform subdivision of the unit hyper-cube. We develop optimization procedures for finding copulas defined by multiply-stochastic matrices matching available information. Two types of information are used for building copulas: 1) Spearman Rho rank correlation coefficients; 2) Empirical distributions of sums of random variables combined with empirical marginal probability distributions. To construct checkerboard copulas we solved optimization problems. The first problem maximizes entropy with constraints on Spearman Rho coefficients. The second problem minimizes some error function to match available data. We conducted a case study illustrating the application of the developed methodology using property and casualty insurance data. The optimization problems were numerically solved with the AORDA Portfolio Safeguard (PSG) package, which has precoded entropy and error functions. Case study data, codes, and results are posted at the web.

Keywords: multivariate distributions, checkerboard copula, Spearman Rho rank correlation, entropy, case study, optimization procedure, Portfolio Safeguard, PSG

MSC: 62H05, 62H12, 90C25, 91-10

1 Introduction

The objective of the paper is to build a joint distribution of incurred losses (or loss ratios) for a set of correlated classes in insurance business. Some empirical information about the dependence structure of these classes of business is available and we want to build a copula of a joint distribution. This methodology is relevant in any situation where an aggregate loss distribution across correlated classes of business needs to be found. It is especially helpful in representing simultaneous large losses in many classes, a particularly thorny problem in actuarial science.

Suppose that the following information about random variables (losses) is available: a) distributions for m one-dimension random variables; b) distributions for sums of some of these random variables. For instance, we know distributions of 3 random variables, and we know the distribution of the sum of the first two random variables. The main objective of this paper is to develop methodology for finding a copula which matches available information. To our knowledge, this is an original research contribution which is not covered in other papers. We do not know other approaches for calibrating a copula based on such information. For instance, papers [9, 10] consider copulas based on general partitions-of-unity. However, the problem addressed in this paper was not considered in these publications.

Viktor Kuzmenko: V.M.Glushkov Institute of Cybernetics, Kyiv, Ukraine
Romel Salam: Fellow of the Casualty Actuarial Society and Member of the American Academy of Actuaries
***Corresponding Author: Stan Uryasev:** Applied Mathematics and Statistics, Stony Brook University, USA, E-mail: Stanislav.Uryasev@stonybrook.edu

The second objective of the paper is to review earlier results on calibrating checkerboard copulas by maximizing entropy. This review is motivated by the case study for finding a copula with known Spearman's rank correlation coefficients (which are called also grade correlation coefficients).

We conducted a case study illustrating the application of the developed methodology using property and casualty insurance data. The conducted case studies (codes, data) are posted at the web.

This paper relies on results for checkerboard copulas of maximum entropy developed in [2–4, 8, 11, 12]. An m-dimensional *copula* where $m \geq 2$, is a continuous, m-increasing, probability distribution function $C : [0,1]^m \mapsto [0,1]$ on the unit m-dimensional hyper-cube with uniform marginal probability distributions. A *checkerboard* copula is a distribution with a corresponding density defined almost everywhere by a step function on an m-uniform subdivision of the hyper-cube. I.e., the checkerboard copula is a distribution on the unit hyper-cube $[0,1]^m$ defined by subdividing the hyper-cube into n^m identical small hyper-cubes $I_{\boldsymbol{i}}$ with constant density on each one. Suppose that the density on $I_{\boldsymbol{i}}$, where $\boldsymbol{i} = (i_1, i_2, \dots i_m)$ is defined by the expression $n^{m-1} h_{\boldsymbol{i}}$, where $h_{\boldsymbol{i}}$ is an element of hyper-matrix $\boldsymbol{h} = [h_{\boldsymbol{i}}] \in \mathbb{R}^{n^m}$ with $h_{\boldsymbol{i}} \in [0,1]$. The marginal distributions of C are uniform if hypermatrix $\boldsymbol{h}$ is *multiply-stochastic*, i.e., satisfies

$$\sum_{\boldsymbol{i}\,:\,i_r = j} h_{\boldsymbol{i}} = 1 \;\; \forall \;\; r = 1, \dots, m, \;\; j = 1, \dots, n\,.$$

The paper develops optimization procedures for finding copulas defined by a multiply-stochastic matrices matching available information.

2 Checkerboard Copula with Prescribed Spearman Rho Coefficients

The definitions and statements for the multi-dimensional copulas in this section are taken from [12]. This section contains the maximum entropy optimization problem for a checkerboard copula with prescribed Spearman rho coefficients. This problem is solved in the case study in Section 4.2. Also, this introductory section provides definitions for the main methodological Section 3.

2.1 2-Dimensional Checkerboard Copula Defined by Doubly-Stochastic Matrix

Definitions and notations for multidimensional copulas are quite complicated, therefore, we start with the two-dimensional case, similar to paper [11].

Let (X_1, X_2) be a pair of real valued random variables on $\mathbb{R}^2$ and let $g(x_1, x_2)$ be the joint probability density. The corresponding marginal probability densities are

$$g_1(x_1) = \int_{\mathbb{R}} g(x_1, x_2)\, dx_2 \quad \text{and} \quad g_2(x_2) = \int_{\mathbb{R}} g(x_1, x_2)\, dx_1$$

for $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}$. In practice we often wish to construct a joint probability distribution where the corresponding marginal distributions are already known. The method of copulas is one possible solution method. Let $c(u_1, u_2)$ be a density of two-dimension copula, i.e., the joint probability density on the unit square with marginal densities

$$c_1(u_1) = \int_0^1 c(u_1, u_2)\, du_2 = 1 \quad \text{and} \quad c_2(u_2) = \int_0^1 c(u_1, u_2)\, du_1 = 1$$

for each $u_1 \in [0,1]$ and each $u_2 \in [0,1]$. Let $g_1(x_1)$ and $g_2(x_2)$ be the known probability densities with corresponding cumulative distribution functions $F_1(x_1)$ and $F_1(x_1)$ for real valued random variables X_1 and X_2. The joint density, defined by the copula density $c(u_1, u_2)$, equals

$$g(x_1, x_2) = c\left(F_1(x_1), F_2(x_2)\right) g_1(x_1) g_2(x_2) \quad \text{for} \quad (x_1, x_2) \in \mathbb{R}^2\,.$$

Let $\boldsymbol{h} = [h_{ij}] \in \mathbb{R}^{n\times n}$ be a *a doubly-stochastic* matrix, i.e., matrix satisfying conditions

$$\sum_{i=1}^{n} h_{ij} = \sum_{l=1}^{n} h_{jl} = 1 \quad \text{for all} \quad j = 1, \ldots, n\,.$$

Let us define a partition $0 = a(1) < a(2) < \cdots < a(n+1) = 1$ of the interval $[0, 1]$ where $a(k) = (k-1)/n$, $k = 1, \ldots, n+1$. Define the step function $c(u_1, u_2)$ almost everywhere on the region $[0, 1]\times[0, 1]$ by the formula

$$c(u_1, u_2) = n\,h_{ij} \quad \text{for} \quad (u_1, u_2) \in \big(a(i), a(i+1)\big)\times\big(a(j), a(j+1)\big),\ i = 1, \ldots, n,\ j = 1, \ldots, n\,.$$

It can be easily shown that $c(u_1, u_2)$ is a joint density function on the unit square with uniform marginal densities, i.e., it is a density of a copula. Indeed, the integral of $c(u_1, u_2)$ over the unit square equals

$$\begin{aligned}\int_0^1\!\!\int_0^1 c(u_1, u_2)\,du_1\,du_2 &= \sum_{i=1}^{n}\sum_{j=1}^{n}\int_{a(i)}^{a(i+1)}\int_{a(j)}^{a(j+1)} n\,h_{ij}\;du_1\,du_2 = \sum_{i=1}^{n}\sum_{j=1}^{n} nh_{ij}\int_{a(i)}^{a(i+1)}\int_{a(j)}^{a(j+1)} du_1\,du_2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} n\,h_{ij}\frac{1}{n^2} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij} = \frac{1}{n}\sum_{i=1}^{n} 1 = 1\,.\end{aligned}$$

Also, suppose that $u_1 \in \big[a(i), a(i+1)\big]$, then

$$\int_0^1 c(u_1, u_2)\,du_2 = \sum_{j=1}^{n}\int_{a(j)}^{a(j+1)} n\,h_{ij}\;du_2 = \sum_{j=1}^{n} n\,h_{ij}\int_{a(j)}^{a(j+1)} du_2 = \sum_{j=1}^{n} n\,h_{ij}\,\frac{1}{n} = \sum_{j=1}^{n} h_{ij} = 1\,.$$

Therefore, the marginal density for any $u_1 \in [0,\ 1]$, equals

$$c_1(u_1) = \int_0^1 c(u_1, u_2)\,du_2 = 1\,,$$

and similar for any $u_2 \in [0,\ 1]$,

$$c_2(u_2) = \int_0^1 c(u_1, u_2)\,du_1 = 1\,.$$

The corresponding checkerboard copula $C : [0, 1]\times[0, 1] \mapsto [0, 1]$ is defined as follows,

$$C(u_1, u_2) = \int_0^{u_1}\!\!\int_0^{u_2} c(v_1, v_2)\,dv_1\,dv_2 \quad \text{for} \quad u_1 \in [0,\ 1],\ u_2 \in [0,\ 1]\,.$$

Example of two-dimensional copula.
Let us consider the case with m=2, n=4. Table 1 shows an example of hyper-matrix $\boldsymbol{h}$. Table 2 and Figure 1 show the density of the checkerboard copula. Table 3 and Figure 2 show the Checkerboard copula.

2.2 Definitions for *m*-Dimensional Case

Let $m \in \mathbb{N}$ with $m \geq 2$ and let $\boldsymbol{X} = (X_1, \ldots, X_m) \in \mathbb{R}^m$ be a vector–valued random variable with joint probability density $g : \mathbb{R}^m \mapsto \mathbb{R}$. The corresponding marginal probability densities are

$$g_r(x_r) = \int_{\mathbb{R}^{m-1}} g(\boldsymbol{x})\;d\boldsymbol{\pi}_r^c\boldsymbol{x}$$

Table 1: Example. Hyper-matrix $\boldsymbol{h}$.

	1	2	3	4
1	0.167	0.333	0.333	0.167
2	0.167	0.333	0.333	0.167
3	0.167	0.333	0.333	0.167
4	0.5	0	0	0.5

Table 2: Example. Density values of the checkerboard copula.

	1	2	3	4
1	0.667	1.333	1.333	0.667
2	0.667	1.333	1.333	0.667
3	0.667	1.333	1.333	0.667
4	2	0	0	2

Table 3: Example. Checkerboard copula values.

	1	2	3	4
1	0.042	0.125	0.208	0.25
2	0.083	0.250	0.417	0.5
3	0.125	0.375	0.625	0.75
4	0.25	0.5	0.75	1.0

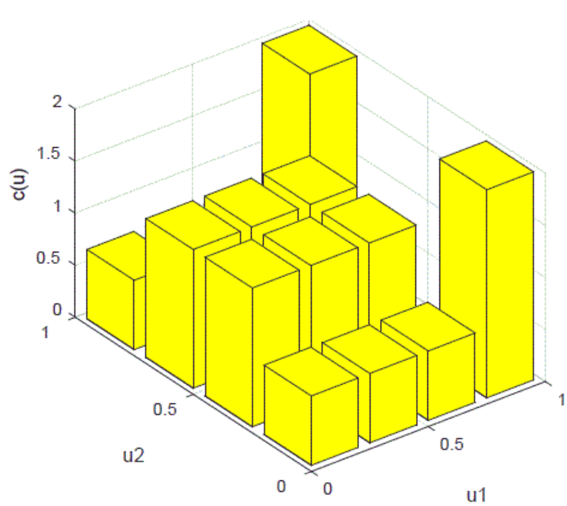

Figure 1: Example. Density of the checkerboard copula.

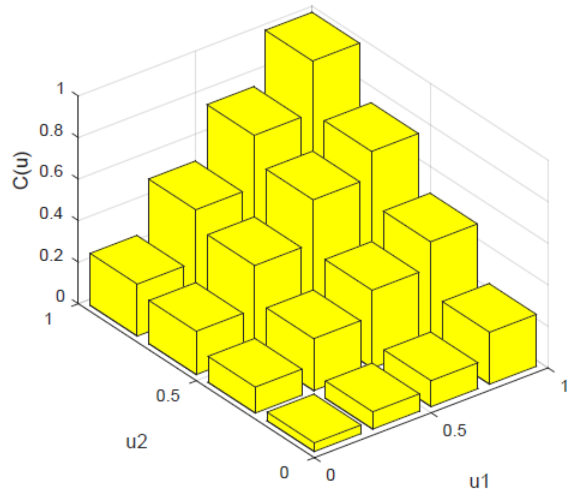

Figure 2: Example. Checkerboard copula.

for all $x_r \in \mathbb{R}$ and each $r = 1, \dots, m$ where we write $\boldsymbol{x} = (x_1, \dots, x_m) \in \mathbb{R}^m$ and where the projection $\boldsymbol{\pi}_r : \mathbb{R}^m \mapsto \mathbb{R}$ onto the x_r-axis and the complementary projection $\boldsymbol{\pi}_r^c : \mathbb{R}^m \mapsto \mathbb{R}^{m-1}$ are defined for each $r = 1, 2, \dots, m$ by

$$\boldsymbol{\pi}_r \boldsymbol{x} = x_r \quad \text{and} \quad \boldsymbol{\pi}_r^c \boldsymbol{x} = \begin{cases} (x_2, \dots, x_m) & \text{if} \quad r = 1 \\ (x_1, \dots, x_{r-1}, x_{r+1}, \dots, x_m) & \text{if} \quad r = 2, \dots, m-1 \\ (x_1, \dots, x_{m-1}) & \text{if} \quad r = m. \end{cases}$$

Frequently, in simulation of random events it is needed to construct a joint probability distribution where the corresponding marginal distributions are already known. The method of copulas provides a theoretical basis for such analysis. If the joint distribution is known and the marginal distributions are continuous then the copula is uniquely defined. We refer to the book by Nelsen [7] for the fundamental theory. It is convenient to assume that the given marginal distributions are continuous. Let $c : [0, 1]^m \mapsto [0, \infty)$ be a joint probability density on the unit m–dimensional hyper-cube with uniform marginal densities. That is, the marginal densities $c_r : [0, 1] \mapsto [0, \infty)$, satisfy the conditions

$$c_r(u_r) = 1 \quad \Leftrightarrow \quad \int\limits_{[0,1]^{m-1}} c(\boldsymbol{u})\, d\boldsymbol{\pi}_r^c \boldsymbol{u} = 1$$

for all $u_r \in [0, 1]$ and each $r = 1, \dots, m$. The distribution $C : [0, 1]^m \mapsto [0, 1]$ defined by

$$C(\boldsymbol{u}) = \int\limits_{\times_{i=1}^{n} [0, u_i]} c(\boldsymbol{v}) d\boldsymbol{v}$$

for all $\boldsymbol{u} \in [0, 1]^m$ is an m-dimensional copula. The copula C defines a joint distribution for a vector valued random variable $\boldsymbol{U} = (U_1, \dots, U_m)$ on the unit hyper-cube $[0, 1]^m$. Let $f_s : \mathbb{R} \mapsto \mathbb{R}$ be a given probability density with corresponding cumulative distribution function $F_s : \mathbb{R} \mapsto [0, 1]$ for each $s = 1, \dots, m$. Write $\boldsymbol{f} = (f_1, \dots, f_m) : \mathbb{R}^m \mapsto [0, \infty)^m$ and $\boldsymbol{F} = (F_1, \dots, F_m) : \mathbb{R}^m \mapsto [0, 1]^m$, also

$$\boldsymbol{u} = \boldsymbol{F}(\boldsymbol{x}) \quad \Leftrightarrow \quad (u_1, \dots, u_m) = (F_1(x_1), \dots, F_m(x_m))$$

for each $\boldsymbol{x} = (x_1 \dots, x_r) \in \mathbb{R}^m$.

The joint density $g : \mathbb{R}^m \mapsto [0, \infty)$ defined for random vector $\boldsymbol{X} = (X_1, \dots, X_m)$ by the formula

$$g(\boldsymbol{x}) = c(\boldsymbol{F}(\boldsymbol{x})) \prod_{s=1}^{m} f_s(x_s) \quad \text{for} \quad \boldsymbol{x} \in \mathbb{R}^m$$

has prescribed marginal densities for $X_r,\ r = 1, \dots, m$ given by

$$g_r(x_r) = f_r(x_r) \int_{\mathbb{R}^{m-1}} c(\boldsymbol{F}(\boldsymbol{x})) \prod_{s \neq r} f_s(x_s)\, d\boldsymbol{\pi}_r^c \boldsymbol{x} = f_r(x_r) \int_{[0,1]^{m-1}} c(\boldsymbol{u})\, d\boldsymbol{\pi}_r^c \boldsymbol{u} = f_r(x_r) \ \text{ for } \ x_r \in \mathbb{R}\,.$$

The corresponding m-dimensional distribution $G : \mathbb{R}^m \mapsto [0, 1]$ is defined in terms of the copula C and the marginal distributions $\boldsymbol{F}$ by the formula

$$G(\boldsymbol{x}) = C(\boldsymbol{F}(\boldsymbol{x})) \ \text{ for } \ \boldsymbol{x} \in \mathbb{R}^m\,.$$

2.3 Checkerboard Copulas and Multiply-Stochastic Matrices

Let $n \in \mathbb{N}$ be a natural number and let $\boldsymbol{h}$ be a non-negative m–dimensional hyper–matrix given by $\boldsymbol{h} = [h_{\boldsymbol{i}}] \in \mathbb{R}^{\ell}$ where $\ell = n^m$ and $\boldsymbol{i} \in \{1, \dots, n\}^m$ with $h_{\boldsymbol{i}} \in [0, 1]$. For instance, suppose that $n = 3,\ m = 2$, then, $\{1, 2, 3\}^2$ is the list of pairs $\{1, 1\}, \{2, 1\}, \{3, 1\}, \{1, 2\}, \{2, 2\}, \{3, 2\}, \{1, 3\}, \{2, 3\}, \{3, 3\}$. If, for instance, $\boldsymbol{i} = \{3, 2\}$, than $h_{\boldsymbol{i}} = h_{\{3,2\}}$.

Define the *marginal sums* $\sigma_r : \{1, \dots, n\} \mapsto \mathbb{R}$ by the formulae

$$\sigma_r(i_r) = \sum_{\boldsymbol{\pi}_r^c \boldsymbol{i}\, \in\, \{1,2,\dots,n\}^{m-1}} h_{\boldsymbol{i}}$$

for each $i_r = 1, 2, \dots, m$. If $\sigma_r(i_r) = 1$ for all $r = 1, 2, \dots, m$, then we say that $\boldsymbol{h}$ is *multiply stochastic*. Define the partition $0 = a(1) < a(2) < \cdots < a(n + 1) = 1$ of the interval $[0, 1]$ by setting $a(k) = (k - 1)/n$ for each $k = 1, \dots, n + 1$ and define a step function $c_{\boldsymbol{h}} : [0, 1]^m \mapsto \mathbb{R}$ almost everywhere by the formula

$$c_{\boldsymbol{h}}(\boldsymbol{u}) = n^{m-1} \cdot h_{\boldsymbol{i}} \quad \text{if} \quad \boldsymbol{u} \in I_{\boldsymbol{i}} = \times_{r=1}^{m} [a(i_r), a(i_r + 1)]$$

for each $\boldsymbol{i} = (i_1, \dots, i_m) \in \{1, 2, \dots, n\}^m$. Now it follows that

$$\int_{[0,1]^m} c_{\boldsymbol{h}}(\boldsymbol{u}) \cdot d\boldsymbol{u} = \sum_{\boldsymbol{i}\, \in\, \{1,\dots,n\}^m} \int_{I_{\boldsymbol{i}}} c_{\boldsymbol{h}}(\boldsymbol{u}) \cdot d\boldsymbol{u} = \sum_{\boldsymbol{i}\, \in\, \{1,\dots,n\}^m} n^{m-1}\, h_{\boldsymbol{i}}\, \frac{1}{n^m} = 1\,. \tag{1}$$

Since

$$(c_{\boldsymbol{h}})_r(u_r) = \int_{[0,1]^{m-1}} c_{\boldsymbol{h}}(\boldsymbol{u}) \cdot d\boldsymbol{\pi}_r^c \boldsymbol{u} = \sum_{\boldsymbol{\pi}_r^c \boldsymbol{i}\, \in\, \{1,\dots,n\}^{m-1}} n^{m-1}\, h_{\boldsymbol{i}}\, \frac{1}{n^{m-1}} = 1$$

for all $r = 1, 2, \dots, m$, then the step function $c_{\boldsymbol{h}} : [0, 1]^m \mapsto [0, \infty)$ is a joint density function for a corresponding checkerboard copula $C_{\boldsymbol{h}} : [0, 1]^m \mapsto [0, 1]$ defined by

$$C_{\boldsymbol{h}}(\boldsymbol{u}) = \int_{\times_{i=1}^{n} [0, u_i]} c_{\boldsymbol{h}}(\boldsymbol{v})\, d\boldsymbol{v} \quad \text{for } \boldsymbol{u} \in [0, 1]^m\,.$$

The joint density $g_{\boldsymbol{h}} : \mathbb{R}^m \mapsto [0, \infty)$ for the random variable $\boldsymbol{X} = (X_1, \dots, X_m)$ is defined by

$$g_{\boldsymbol{h}}(\boldsymbol{x}) = c_{\boldsymbol{h}}(\boldsymbol{F}(\boldsymbol{x})) \prod_{s=1}^{m} f_s(x_s) \ \text{ for } \ \boldsymbol{x} \in \mathbb{R}^m\,,$$

and the corresponding distribution function $G_{\boldsymbol{h}} : \mathbb{R}^m \mapsto [0, 1]$ is defined in terms of the copula $C_{\boldsymbol{h}}$ and the prescribed marginal distributions $\boldsymbol{F}$ by the formula

$$G_{\boldsymbol{h}}(\boldsymbol{x}) = C_{\boldsymbol{h}}(\boldsymbol{F}(\boldsymbol{x})) \ \text{ for } \ \boldsymbol{x} \in \mathbb{R}^m\,.$$

2.4 Spearman Rho Correlation Coefficient

The most widely known are Kendall's tau and Spearman rho, both of which measure a form of dependence known as concordance. Spearman rho is often called the grade correlation coefficient. If x_r are observations from a real valued random variable X_r with cumulative distribution function F_r then the *grade* of x_r is given by $u_r = F_r(x_r)$. Note that the grade u_r can be regarded as an observation of the uniform random variable $U_r = F_r(X_r)$ on $[0, 1]$ and that U_r has mean $1/2$ and variance $1/12$. The *grade correlation coefficient* for the continuous random variables X_r and X_s where $r < s$ is defined as the correlation for the grade random variables $U_r = F_r(X_r)$ and $U_s = F_s(X_s)$ by the formula

$$\rho_{r,s} = \frac{E[(U_r - 1/2)(U_s - 1/2)]}{E[(U_r - 1/2)^2]^{1/2} E[(U_s - 1/2)^2]^{1/2}} = 12\left(E[U_r U_s] - 1/4\right).$$

We refer the reader to Nelsen [7] for further details. The Spearman rho correlation coefficient for the checkerboard copula is given by

$$\rho_{r,s} = 12\left(\frac{1}{n^3}\sum_{\boldsymbol{i}\,\in\,\{1,\dots,n\}^m} h_{\boldsymbol{i}}(i_r - 1/2)(i_s - 1/2) - \frac{1}{4}\right). \tag{2}$$

2.5 Entropy

Let $\boldsymbol{h} \in \mathbb{R}^\ell$ be a multiply stochastic hyper–matrix and let $c_{\boldsymbol{h}} : [0, 1]^m \to \mathbb{R}$ be the associated elementary joint density defined previously. The *entropy* of $\boldsymbol{h}$ is defined by

$$J(\boldsymbol{h}) = (-1)\left[\frac{1}{n}\sum_{\boldsymbol{i}\,\in\,\{1,\dots,n\}^m} h_{\boldsymbol{i}} \log_e h_{\boldsymbol{i}} + (m-1)\log_e n\right]. \tag{3}$$

2.6 Maximum Entropy Problem with Prescribed Spearman Rho Coefficients

We wish to select a multiply stochastic hyper–matrix $\boldsymbol{h} = [h_{\boldsymbol{i}}] \in \mathbb{R}^\ell$ to match known grade correlation coefficients $\rho_{r,s}$ for all $r < s$ in such a way that the entropy is maximized. We now formulate the optimization problem for finding copula with prescribed Spearman rho coefficients.

Optimization problem with prescribed Spearman rho coefficients
Find the hyper-matrix $\boldsymbol{h} \in \mathbb{R}^\ell$ maximizing the entropy

$$J(\boldsymbol{h}) = (-1)\left[\frac{1}{n}\sum_{\boldsymbol{i}\,\in\,\{1,\dots,n\}^m} h_{\boldsymbol{i}} \log_e h_{\boldsymbol{i}} + (m-1)\log_e n\right] \tag{4}$$

subject to the constraints

$$12\left[\frac{1}{n^3}\cdot\sum_{\boldsymbol{i}\,\in\,\{1,\dots,n\}^m} h_{\boldsymbol{i}}(i_r - 1/2)(i_s - 1/2) - 1/4\right] = \rho_{r,s}\,, \quad 1 \le r < s \le m \tag{5}$$

$$\sum_{\boldsymbol{\pi}_r^c \boldsymbol{i}\,\in\,\{1,\dots,n\}^{m-1}} h_{\boldsymbol{i}} = 1\,, \quad i_r \in \{1,\dots,n\}\,, \quad r = 1,\dots,m \tag{6}$$

$$h_{\boldsymbol{i}} \ge 0\,, \quad \boldsymbol{i} \in \{1,\dots,n\}^m \tag{7}$$

In general terms the problem is well posed. There are a finite number of linear constraints on $\boldsymbol{h}$ and so the *feasible* set $\mathcal{F}$ of hyper–matrices satisfying (5,6,7) is a bounded (closed) convex set in $\mathbb{R}^{\ell}$. The function $J : \mathcal{F} \mapsto [0, \infty)$ is strictly concave. If the interior or core of $\mathcal{F}$ is non-empty then there must be a unique solution for $\boldsymbol{h}$ with strictly positive coordinates. The reader is referred to [1, 5] for a general discussion of the requisite convex analysis and nonlinear optimization.

3 Copula Defined by Sums of Random Variables

Let X_r, $r = 1, 2, \dots, m$ are random values with distributions $F_r(x)$ and single-valued quantile functions $F_r^{-1}(u)$ for all $u \in (0, 1)$. We denote by $\boldsymbol{g}$ a subset indexes of these random values. For instance, suppose that $r = 1, 2, \dots, 7$, then, we may have $\boldsymbol{g} = \{2, 5\}$ or $\boldsymbol{g} = \{1, 3, 4, 6\}$. We will denote by $Z_{\boldsymbol{g}}$ the sum of random values with indexes $r \in \boldsymbol{g}$, i.e. $Z_{\boldsymbol{g}} = \sum_{r\in\boldsymbol{g}} X_r$. We denote by $F_{\boldsymbol{g}}(z)$ the distribution for the random value $Z_{\boldsymbol{g}}$.

Let us assume that the distributions $F_r(x)$, $r = 1, 2, \dots, m$, and the distribution $F_{\boldsymbol{g}}(z)$ are available. We want to build a copula for random values X_r, $r = 1, 2, \dots, m$, based on available information about these distributions.

Let us denote

$$Ind(a \leqslant b) = \begin{cases} 1 & \text{if } a \leqslant b \\ 0 & \text{otherwise.} \end{cases}$$

We define the projection $\boldsymbol{\pi}_{rs} : \mathbb{R}^m \mapsto \mathbb{R}^2$ onto the $u_r u_s$-plane and the complementary projection $\boldsymbol{\pi}_{rs}^c : \mathbb{R}^m \mapsto \mathbb{R}^{m-2}$ for $1 \le r < s \le m$ by the formulae

$$\boldsymbol{\pi}_{rs}\boldsymbol{u} = (u_r, u_s)$$

and

$$\boldsymbol{\pi}_{rs}^c\boldsymbol{u} = \begin{cases} (u_3, \dots, u_m) & \text{if } \; r = 1,\ s = 2 \\ (u_2, \dots, u_{s-1}, u_{s+1}, \dots, u_m) & \text{if } \; r = 1,\ 2 < s < m \\ (u_2, \dots, u_{m-1}) & \text{if } \; r = 1,\ s = m \\ (u_1, \dots, u_{r-1}, u_{r+1}, \dots, u_{s-1}, u_{s+1}, \dots, u_m) & \text{if } \; 1 < r < s < m \\ (u_1, \dots, u_{r-1}, u_{r+1}, \dots, u_m) & \text{if } \; 1 < r < m-1, s = m \\ (u_1, \dots, u_{m-2}) & \text{if } \; r = m-1, s = m. \end{cases}$$

We will explain the approach with a simple case when the sum includes only two random values X_r, X_s and $\boldsymbol{g} = \{r, s\}$. By definition,

$$Pr\{X_r + X_s \le z\} = F_{\boldsymbol{g}}(z) \quad \forall\, z \in \mathbb{R}\ . \tag{8}$$

On the other hand, using copula we have

$$\begin{aligned} Pr\{X_r + X_s \le z\} &= \sum_{\boldsymbol{i}\,\in\,\{1,\dots,n\}^m} \int\limits_{I_{\boldsymbol{\pi}_{rs}\boldsymbol{i}}} Ind\left\{F_r^{-1}(u_r) + F_r^{-1}(u_s) \le z\right\} \left[\int\limits_{I_{\boldsymbol{\pi}_{rs}^c\boldsymbol{i}}} c_{\boldsymbol{h}}(\boldsymbol{u}) \cdot d\boldsymbol{\pi}_{rs}^c\boldsymbol{u}\right] d\boldsymbol{\pi}_{rs}\boldsymbol{u} \\ &= \sum_{\boldsymbol{i}\,\in\,\{1,\dots,n\}^m} \int\limits_{I_{\boldsymbol{\pi}_{rs}\boldsymbol{i}}} Ind\left\{F_r^{-1}(u_r) + F_r^{-1}(u_s) \le z\right\} \left[\int\limits_{I_{\boldsymbol{\pi}_{rs}^c\boldsymbol{i}}} n^{m-1} \cdot h_{\boldsymbol{i}} \cdot d\boldsymbol{\pi}_{rs}^c\boldsymbol{u}\right] d\boldsymbol{\pi}_{rs}\boldsymbol{u} \\ &= \sum_{\boldsymbol{i}\,\in\,\{1,\dots,n\}^m} \int\limits_{I_{\boldsymbol{\pi}_{rs}\boldsymbol{i}}} Ind\left\{F_r^{-1}(u_r) + F_r^{-1}(u_s) \le z\right\} \left[n^{m-1} \cdot h_{\boldsymbol{i}} \cdot \frac{1}{n^{m-2}}\right] d\boldsymbol{\pi}_{rs}\boldsymbol{u} \\ &= \sum_{\boldsymbol{i}\,\in\,\{1,\dots,n\}^m} \left[n \int\limits_{a(i_r)}^{a(i_r+1)} \int\limits_{a(i_s)}^{a(i_s+1)} Ind\left\{F_r^{-1}(u_r) + F_r^{-1}(u_s) \le z\right\} du_r\, du_s \right] h_{\boldsymbol{i}}\ . \end{aligned}$$

The last equation and (8) imply

$$\sum_{\boldsymbol{i} \in \{1,\dots,n\}^m} \gamma_{\boldsymbol{i}}^{\boldsymbol{g}}(z)\, h_{\boldsymbol{i}} = F_{\boldsymbol{g}}(z) \quad \forall\, z \in \mathbb{R}\,, \tag{9}$$

where

$$\gamma_{\boldsymbol{i}}^{\boldsymbol{g}}(z) = n \int_{a(i_r)}^{a(i_r+1)} \int_{a(i_s)}^{a(i_s+1)} Ind\left\{F_r^{-1}(u_r) + F_s^{-1}(u_s) \le z\right\}\, du_r\, du_s\ . \tag{10}$$

Similar to (8) we consider the case when cardinality $|\boldsymbol{g}|$ of the set $\boldsymbol{g}$ is higher or equal than 2, i.e. $2 \le |\boldsymbol{g}| \le m$,

$$Pr\{Z_{\boldsymbol{g}} \le z\} = F_{\boldsymbol{g}}(z) \quad \forall\, z \in \mathbb{R}\ . \tag{11}$$

Equation (10) is generalized, in this case, as follows

$$\gamma_{\boldsymbol{i}}^{\boldsymbol{g}}(z) = n^{|\boldsymbol{g}|-1} \int_{I_{\boldsymbol{\pi_g} \boldsymbol{i}}} Ind\left\{\sum_{r \in \boldsymbol{g}} F_r^{-1}(u_r) \le z\right\}\, d\boldsymbol{\pi_g u}\ . \tag{12}$$

So far we have not made any specific assumptions about the distribution of the random value $Z_{\boldsymbol{g}}$. In the considered case we assume that k observations of the random value $Z_{\boldsymbol{g}}$ are available. Therefore, further we suppose that the random value $Z_{\boldsymbol{g}}$ is discretely distributed with equally probable atoms and the distribution function $F_{\boldsymbol{g}}(z)$ takes k values $\frac{1}{k}, \frac{2}{k}, \dots, \frac{k}{k}$. Let us denote by $L^{\boldsymbol{g}}(\boldsymbol{h}, j)$ the loss function, having k equally probable scenarios,

$$L^{\boldsymbol{g}}(\boldsymbol{h}, j) = \frac{j}{k} - \sum_{\boldsymbol{i} \in \{1,\dots,n\}^m} \gamma_{\boldsymbol{i}}^{\boldsymbol{g}}\left(F_{\boldsymbol{g}}^{-1}\left(\frac{j}{k}\right)\right) h_{\boldsymbol{i}}\,, \quad j = 1, \dots k. \tag{13}$$

With this notation, equation (9) can be rewritten as follows,

$$L^{\boldsymbol{g}}(\boldsymbol{h}, j) \;=\; 0\,, \quad j = 1, \dots, k\ . \tag{14}$$

Pay attention that in case if the distribution $F_{\boldsymbol{g}}(z)$ is continuous, we still can use the finite system of equations (14) as an approximation of the infinite system of equations (9). The system of equations (14) may be infeasible. In this case, we can find hyper-matrix $\boldsymbol{h}$ by minimizing an error function. Further, we will consider three error functions:

1) Mean Squared Error,

$$\varepsilon_{MSE}^{\boldsymbol{g}}(\boldsymbol{h}) = \frac{1}{k} \sum_{j=1}^{k} \left[L^{\boldsymbol{g}}(\boldsymbol{h}, j)\right]^2, \tag{15}$$

2) Mean Absolute Error,

$$\varepsilon_{MAE}^{\boldsymbol{g}}(\boldsymbol{h}) = \frac{1}{k} \sum_{j=1}^{k} \left|L^{\boldsymbol{g}}(\boldsymbol{h}, j)\right|, \tag{16}$$

3) CVaR Absolute Error, see [6],

$$\varepsilon_{CVaRAE}^{\boldsymbol{g}}(\boldsymbol{h}, \alpha) = CVaR_\alpha(|L(\boldsymbol{h}, j)|) = min_\xi \left[\xi + \frac{1}{(1-\alpha)k} \sum_{j=1}^{k} \left(\left|L^{\boldsymbol{g}}(\boldsymbol{h}, j)\right| - \xi\right)\right], \tag{17}$$

with confidence parameter $\alpha \in [0, 1)$.

Further we formulate regression problem for finding copula with one set of constraints (14) . Let $\boldsymbol{g}$ be a subset of indices of continuous random variables X_r with distributions $F_r(x)$, $r = 1, 2, \dots, m$. Let denote by $\varepsilon^{\boldsymbol{g}}(\boldsymbol{h})$ one of the three considered error functions. We will solve the following optimization problem to find an optimal vector $\boldsymbol{h}$ defining copula.

Regression problem with one sum function

Find the hyper-matrix $\boldsymbol{h} \in \mathbb{R}^{\ell}$ minimizing the error

$$\min_{\boldsymbol{h}} \ \varepsilon^{\boldsymbol{g}}(\boldsymbol{h}) \tag{18}$$

subject to the constraints

$$\sum_{\boldsymbol{\pi}_r^c \boldsymbol{i} \in \{1,\ldots,n\}^{m-1}} h_{\boldsymbol{i}} = 1 \ , \quad i_r \in \{1, \ldots, n\} \ , \quad r = 1, \ldots, m \tag{19}$$

$$h_{\boldsymbol{i}} \geq 0 \ , \quad \boldsymbol{i} \in \{1, \ldots, n\}^m \tag{20}$$

In general terms the problem (18,19,20) is well posed. There is a finite number of linear constraints on $\boldsymbol{h}$ and so the *feasible* set $\mathcal{F}$ of hyper–matrices satisfying (19,20) is a bounded closed convex set in $\mathbb{R}^{\ell}$. The function $\varepsilon : \mathcal{F} \mapsto [0, \infty)$ is convex for the considered error function. The interior of $\mathcal{F}$ is non-empty, therefore there is a convex set of optimal solutions for $\boldsymbol{h}$. The reader is referred to [1, 5] for a general discussion of convex analysis and nonlinear optimization.

It is important to note that the problem (18,19,20) has sense if the error function $\varepsilon^{\boldsymbol{g}}(\boldsymbol{h})$ on optimal solution point is not equal to zero, which means that the system of linear constraints (14) is not feasible. Suppose that the problem (18,19,20) has zero optimal objective function, then we need to solve the following entropy maximization problem to assure that the solution is based only on available information specified by constraints.

Entropy maximization problem with one sum function

Find hyper-matrix $\boldsymbol{h} \in \mathbb{R}^{\ell}$ maximizing the entropy

$$J(\boldsymbol{h}) = (-1) \left[\frac{1}{n} \sum_{\boldsymbol{i} \in \{1,\ldots,n\}^m} h_{\boldsymbol{i}} \log_e h_{\boldsymbol{i}} + (m-1) \log_e n \right] \tag{21}$$

subject to constraints

$$L^{\boldsymbol{g}}(\boldsymbol{h}, j) \ = \ 0 \ , \quad j = 1, \ldots, k \tag{22}$$

$$\sum_{\boldsymbol{\pi}_r^c \boldsymbol{i} \in \{1,\ldots,n\}^{m-1}} h_{\boldsymbol{i}} = 1 \ , \quad i_r \in \{1, \ldots, n\} \ , \quad r = 1, \ldots, m \tag{23}$$

$$h_{\boldsymbol{i}} \geq 0 \ , \quad \boldsymbol{i} \in \{1, \ldots, n\}^m \tag{24}$$

Optimization problems with several sum functions

Let us denote by $\boldsymbol{g}_{\mu}$, $\mu = 1, \ldots, d$ subsets of indexes of random values X_r, $r = 1, 2, \ldots, m$. For instance, suppose that $r = 1, 2, \ldots, 7$, and $d = 3$, then, we may have $\boldsymbol{g}_1 = \{2, 4\}$, $\boldsymbol{g}_2 = \{3, 5\}$, $\boldsymbol{g}_3 = \{1, 3, 4, 6\}$.

If we have d sums of random values and accordingly d subsets of indexes, than the objective (18) in the regression problem can be replaced by the weighted average of the error functions,

$$\sum_{\mu=1}^{d} \lambda_{\mu} \ \varepsilon^{\boldsymbol{g}_{\mu}}(\boldsymbol{h}) \ , \tag{25}$$

where $\lambda_\mu > 0$, $\mu = 1, \dots, d$ and $\sum_\mu \lambda_\mu = 1$. For instance, we can take equal coefficients $\lambda_\mu = \frac{1}{d}$. The set constraints (22) in the entropy minimization problem should be specified for every sum of random variables, i.e.,

$$L^{\boldsymbol{g}_\mu}(\boldsymbol{h}, j) = 0 , \quad j = 1, \dots, k_\mu , \quad \mu = 1, \dots, d . \tag{26}$$

Calculation of loss function

According to definition (13), the loss function $L^{\boldsymbol{g}}(\boldsymbol{h}, j)$ is a simple linear function in variables $h_{\boldsymbol{i}}$ with coefficients,

$$\gamma_{\boldsymbol{i}}^{\boldsymbol{g}}(z) = n^{|\boldsymbol{g}|-1} \int\limits_{I_{\boldsymbol{\pi_g i}}} Ind \left\{ \sum_{r \in \boldsymbol{g}} F_r^{-1}(u_r) \le z \right\} d\boldsymbol{\pi_g u} , \quad \text{where,} \quad z = F_{\boldsymbol{g}}^{-1}\left(\frac{j}{k}\right) . \tag{27}$$

Further we show how to calculate the integral in (27). We will explain the idea with the two dimension case when $|\boldsymbol{g}| = 2$. The integration is done over the variables u_r, u_s in the box

$$I_{\boldsymbol{\pi}_{rs}\boldsymbol{i}} = [a(i_r), a(i_r + 1)] \times [a(i_s), a(i_s + 1)] .$$

As specified in (10), formula (27) can be written as follows

$$\gamma_{\boldsymbol{i}}^{\boldsymbol{g}}(z) = n \int\limits_{a(i_r)}^{a(i_r+1)} \int\limits_{a(i_s)}^{a(i_s+1)} Ind \left\{ F_r^{-1}(u_r) + F_s^{-1}(u_s) \le z \right\} du_r \, du_s . \tag{28}$$

When in interior of the box $I_{\boldsymbol{\pi}_{rs}\boldsymbol{i}}$ the indicator function equals only 1 or only 0, integral in (28) can be easily evaluated. Therefore, 3 cases are valid,

$$\gamma_{\boldsymbol{i}}(z) = n \cdot \begin{cases} n^{-2} , & \text{if } F_r^{-1}\big(a(i_r + 1)\big) + F_s^{-1}\big(a(i_s + 1)\big) \le z , \\ 0 , & \text{if } F_r^{-1}\big(a(i_r)\big) + F_s^{-1}\big(a(i_s)\big) \ge z , \\ \int_{a(i_r)}^{a(i_r+1)} \int_{a(i_s)}^{a(i_s+1)} Ind \left\{ F_r^{-1}(u_r) + F_s^{-1}(u_s) \le z \right\} du_r \, du_s , & \text{otherwise} . \end{cases} \tag{29}$$

When in interior of the box $I_{\boldsymbol{\pi}_{rs}\boldsymbol{i}}$ the indicator function equals both 1 and 0, we can consider, approximately, that the integral in (28) equals, n^{-2}, which is volume of $I_{\boldsymbol{\pi}_{rs}\boldsymbol{i}}$, multiplied by $\frac{1}{2}$. Therefore,

$$\gamma_{\boldsymbol{i}}(z) \approx \begin{cases} n^{-1} , & \text{if } F_r^{-1}\big(a(i_r + 1)\big) + F_s^{-1}\big(a(i_s + 1)\big) \le z , \\ 0 , & \text{if } F_r^{-1}\big(a(i_r)\big) + F_s^{-1}\big(a(i_s)\big) \ge z , \\ \frac{1}{2} n^{-1} , & \text{otherwise} . \end{cases} \tag{30}$$

Now, let us derive the exact formula for $\gamma_{\boldsymbol{i}}(z)$ when in $I_{\boldsymbol{\pi}_{rs}\boldsymbol{i}}$ the indicator function equals both 1 and 0. In this case, $F_r^{-1}\big(a(i_r + 1)\big) + F_s^{-1}\big(a(i_s + 1)\big) > z$ and $F_r^{-1}\big(a(i_r)\big) + F_s^{-1}\big(a(i_s)\big) < z$. Coefficient $\gamma_{\boldsymbol{i}}(z)$ is calculated by integrating in the box $I_{\boldsymbol{\pi}_{rs}\boldsymbol{i}}$ over the area where $Ind\left\{F_r^{-1}(u_r) + F_s^{-1}(u_s) \le z\right\} = 1$. Let us denote the upper bound for integrating variables u_r, u_s by

$$M_{\boldsymbol{i},r}(z) = max\left\{min\left\{F_r(z),\ a(i_r + 1)\right\},\ a(i_r)\right\} ,$$

$$M_{\boldsymbol{i},s}(z) = max\left\{min\left\{F_s\left(z - F_r^{-1}(u_r)\right),\ a(i_s + 1)\right\},\ a(i_s)\right\} .$$

So, we have

$$\gamma_{\boldsymbol{i}}^{\boldsymbol{g}}(z) = n \int\limits_{a(i_r)}^{M_{\boldsymbol{i},r}(z)} \int\limits_{a(i_s)}^{M_{\boldsymbol{i},s}(z)} du_r \, du_s . \tag{31}$$

Now, let us consider the case when the sum may contain more than two variables, i.e., $2 \le |\boldsymbol{g}| \le m$. Then, formula (29) is generalized as follows,

$$\gamma_{\boldsymbol{i}}^{\boldsymbol{g}}(z) = n^{|\boldsymbol{g}|-1} \cdot \begin{cases} n^{-|\boldsymbol{g}|}\,, & \text{if } \sum_{r\in\boldsymbol{g}} F_r^{-1}\left(a(i_r+1)\right) \le z\,, \\ 0\,, & \text{if } \sum_{r\in\boldsymbol{g}} F_r^{-1}\left(a(i_r)\right) \ge z\,, \\ \int_{I_{\boldsymbol{\pi_g i}}} Ind\left\{\sum_{r\in\boldsymbol{g}} F_r^{-1}(u_r) \le z\right\} d\boldsymbol{\pi_g u}\,, & \text{otherwise}\,, \end{cases} \tag{32}$$

and the approximate formula (30) is generalized as

$$\gamma_{\boldsymbol{i}}^{\boldsymbol{g}}(z) \approx \begin{cases} n^{-1}\,, & \text{if } \sum_{r\in\boldsymbol{g}} F_r^{-1}\left(a(i_r+1)\right) \le z\,, \\ 0\,, & \text{if } \sum_{r\in\boldsymbol{g}} F_r^{-1}\left(a(i_r)\right) \ge z\,, \\ \frac{1}{2}\, n^{-1}\,, & \text{otherwise}\,. \end{cases} \tag{33}$$

The third term in (32) is derived similar to (31). Let us denote $\boldsymbol{g} = \{r_1, r_2 \ldots, r_l\}$, where $2 \le l = |\boldsymbol{g}| \le m$. We consider the following case,

$$\sum_{r\in\boldsymbol{g}} F_r^{-1}\left(a(i_r+1)\right) = \sum_{v=1}^{l} F_{r_v}^{-1}\left(a(i_{r_v}+1)\right) > z\,,$$

and

$$\sum_{r\in\boldsymbol{g}} F_r^{-1}\left(a(i_r)\right) = \sum_{v=1}^{l} F_{r_v}^{-1}\left(a(i_{r_v})\right) < z\,.$$

Let us denote

$$M_{\boldsymbol{i},r_1}(z) = max\left\{min\left\{F_{r_1}(z),\ a(i_{r_1}+1)\right\},\ a(i_{r_1})\right\},$$

$$M_{\boldsymbol{i},r_2}(z) = max\left\{min\left\{F_{r_2}\left(z - F_{r_1}^{-1}(u_{r_1})\right),\ a(i_{r_2}+1)\right\},\ a(i_{r_2})\right\},$$

$$M_{\boldsymbol{i},r_3}(z) = max\left\{min\left\{F_{r_3}\left(z - \sum_{v=1}^{2} F_{r_v}^{-1}(u_{r_v})\right),\ a(i_{r_3}+1)\right\},\ a(i_{r_3})\right\},$$

$$\ldots$$

$$M_{\boldsymbol{i},r_l}(z) = max\left\{min\left\{F_{r_v}\left(z - \sum_{v=1}^{l-1} F_{r_v}^{-1}(u_{r_v})\right),\ a(i_{r_l}+1)\right\},\ a(i_{r_l})\right\}.$$

So, finally we have

$$\gamma_{\boldsymbol{i}}^{\boldsymbol{g}}(z) = n^{|\boldsymbol{g}|-1} \int_{a(i_{r_1})}^{M_{\boldsymbol{i},r_1}(z)} \int_{a(i_{r_2})}^{M_{\boldsymbol{i},r_2}(z)} \cdots \int_{a(i_{r_v})}^{M_{\boldsymbol{i},r_v}(z)} du_{r_1}\, du_{r_2} \ldots du_{r_v}\,. \tag{34}$$

4 Case Study

This section presents a case study illustrating application of methodology considered in Sections 2 and 3.

The optimization problems were solved with Portfolio Safeguard (PSG); see http://www.aorda.com. PSG is an optimization package for solving nonlinear and mixed-integer optimization problems; it is free for academic purposes. PSG contains precoded classes of nonlinear functions, which allows for formulation and solving of optimization problems in analytic format. MATLAB code was developed to process data and prepare inputs for PSG.

4.1 Copula Defined by Spearman Rho Coefficients

This section provides a case study illustrates the optimization approach presented in Section 2 for finding checkerboard copula with known Spearman Rho coefficients .

The case study codes, data and results are posted at http://uryasev.ams.stonybrook.edu/index.php/research/testproblems/financial_engineering/case-study-checkerboard-copula-defined-by-sperman-rho-coefficients-entropyr/. We posted several instances of solved problems in TEXT, MATLAB, and R formats. The entropy maximization problem is solved with PSG, which has a precoded entropy function. PSG maximizes entropy with dual formulation. However, the user is not involved in this reduction (just option in the optimization problem statement should be specified). PSG automatically generates the dual problem, solves it, and present the results for the primal problem.

The dataset contains five random variables X_j , representing the incurred losses for five classes of business for an insurance company. Accordingly, ten unique Spearman rho coefficients, denoted by $\rho_{r,s}$, were calculated, where $1 \leqslant r < s \leqslant m = 5$, as shown in the following Table 4.

Table 4: Spearman rho coefficients $\rho_{r,s}$.

1	0.535294	0.664706	0.629412	-0.414706
0.535294	1	0.247059	0.423529	-0.4
0.664706	0.247059	1	0.844118	-0.317647
0.629412	0.423529	0.844118	1	-0.247059
-0.414706	-0.4	-0.317647	-0.247059	1

The optimization problem (4–7) is reduced to the following optimization problem (35–38).

Optimization problem
Find hyper-matrix $\boldsymbol{h} \in \mathbb{R}^5$ by maximizing

$$\max_{\boldsymbol{h}} \quad -\sum_{\boldsymbol{i} \in \{1,\dots,n\}^5} h_{\boldsymbol{i}} \log_e h_{\boldsymbol{i}} \tag{35}$$

subject to constraints

$$\frac{12}{n^3} \sum_{\boldsymbol{i} \in \{1,\dots,n\}^5} h_{\boldsymbol{i}}(i_r - 1/2)(i_s - 1/2) - 3 = \rho_{r,s} \ , \quad 1 \le r < s \le 5 \tag{36}$$

$$\sum_{\boldsymbol{\pi}_r^c \boldsymbol{i} \in \{1,\dots,n\}^4} h_{\boldsymbol{i}} = 1 \ , \quad i_r \in \{1,\dots,n\} \ , \quad r = 1,\dots,5 \tag{37}$$

$$h_{\boldsymbol{i}} \ge 0 \ , \quad \boldsymbol{i} \in \{1,\dots,n\}^5 \tag{38}$$

We solved the optimization problem (35–38) with grid parameter $n = 4, 8, 10$. Table 5 shows the optimal objective value and calculation times. We observe that the solution time is quickly increasing with dimension n. The dimension $n = 10$, on one hand, is sufficiently large to get a good approximation precision of the copula, on the other hand, the optimization time = 7.66 sec, is not significant for a nonlinear optimization problem having $n^5 = 100,000$ prime variables $h_{\boldsymbol{i}}$. We want to emphasize that this is a nonlinear optimization problem with quite large number of variables. PSG package has a precoded entropy function which is very efficiently implemented. Data are posted at the web and a reader can benchmark this problem with some other nonlinear programming software.

Figures 3-5 show two-dimensional projections of density of the optimal checkerboard copula with $n = 10$. Two-dimensional projection of the density to coordinates u_{i_1}, u_{i_2} is done by fixing complementary components (not involved in the projection) at value 0.5 .

Table 5: Calculation results for optimization problem (35–38) with grid parameters $n = 4, 8, 10$.

n	Optimal Value	Solution Time (sec)
4	15.58	0.03
8	55.82	1.91
10	78.98	7.66

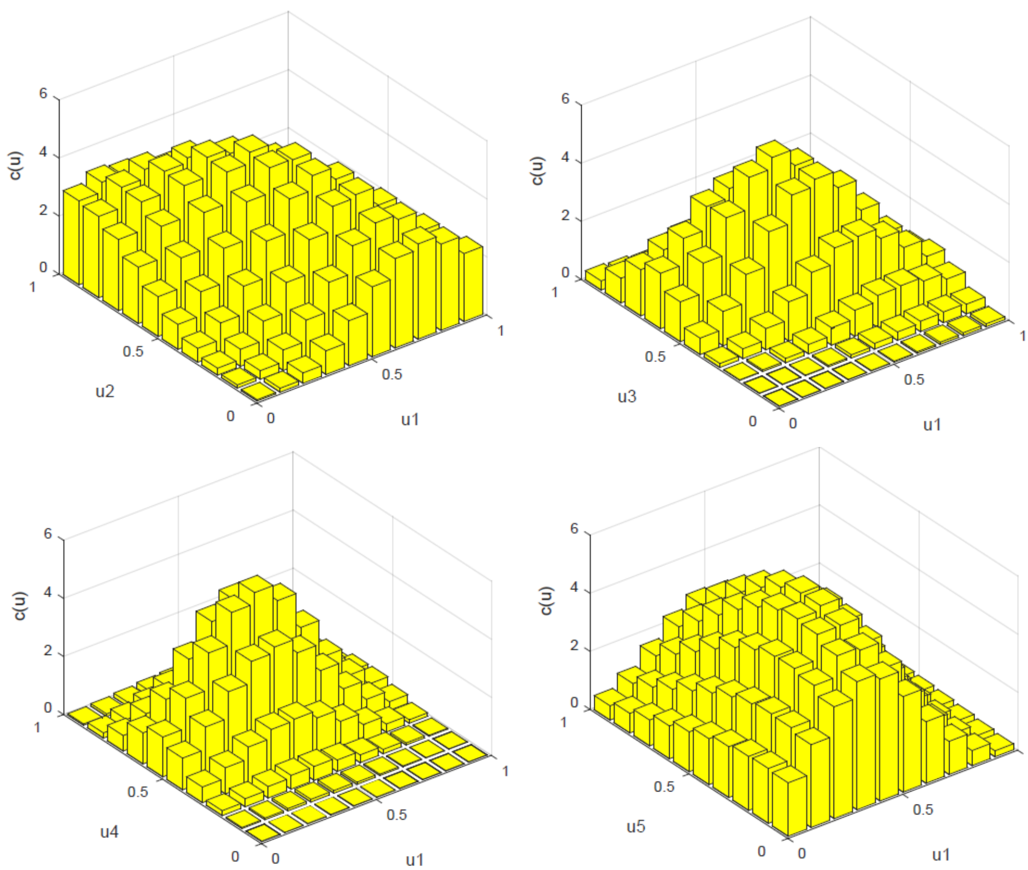

Figure 3: Two-dimensional projections (u1-u2; u1-u3; u1-u4; u1-u5) of density of the checkerboard copula, m=5, n=10, obtained by maximizing entropy.

4.2 Checkerboard Copula Defined by Sums of Random Variables

This section calibrates checkerboard copulas with known marginal distributions and distributions of sums of random variables, as described in Section 3. The case study codes, data and results are posted at http://uryasev.ams.stonybrook.edu/index.php/research/testproblems/financial_engineering/case-study-checkerboard-copula-defined-by-sums-of-random-variables/. We have found m=3-dimensional checkerboard copulas with grid parameter $n = 10$. The error minimization problems were solved with the PSG package (see http://www.aorda.com) which has precoded error functions: Mean Squared, Mean Absolute, and CVaR Absolute. Standard statistical packages have Mean Squared and Mean Absolute minimization capabilities, however, they do not accept constraints. Optimization packages, such as Gurobi can solve very efficiently linear and quadratic optimization problems. Problems considered in this section can be reduced to quadratic or linear programming. However, a significant effort need to be made to make this reduction, write a code, and debug. With PSG it is possible to avoid these time consuming steps.

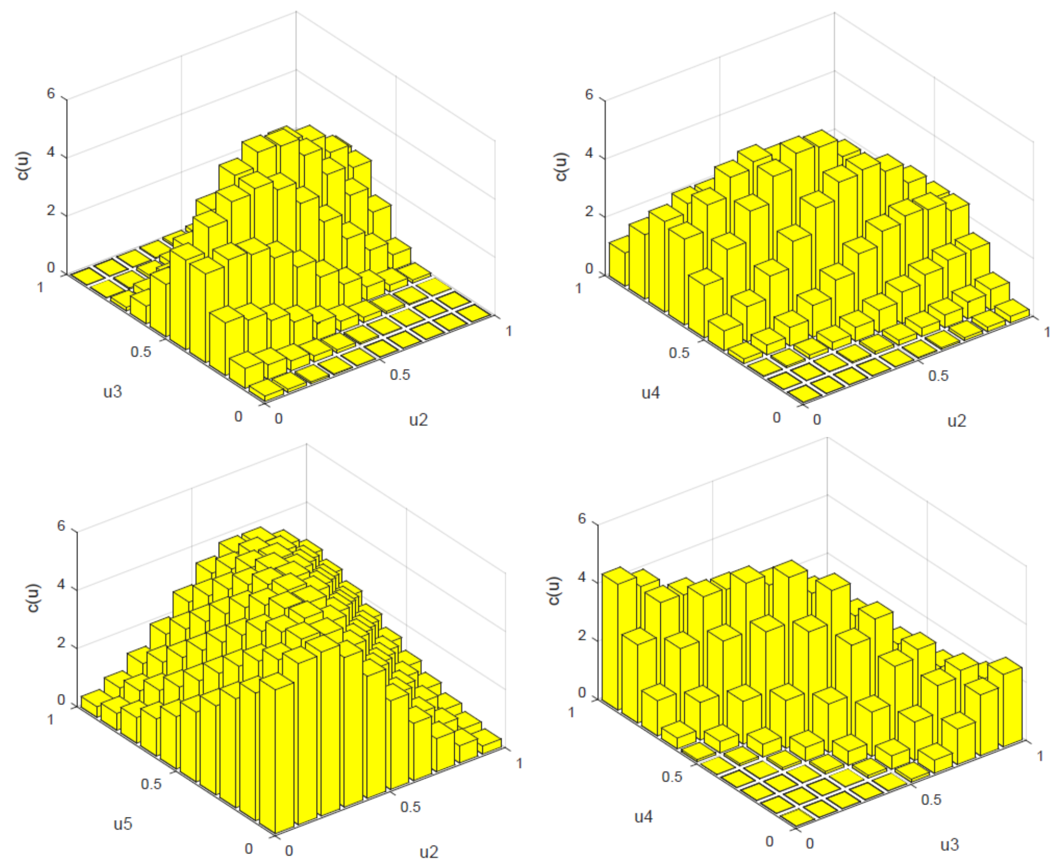

Figure 4: Two-dimensional projections (u2-u3; u2-u4; u2-u5; u3-u4) of density of the checkerboard copula, m=5, n=10, obtained by maximizing entropy.

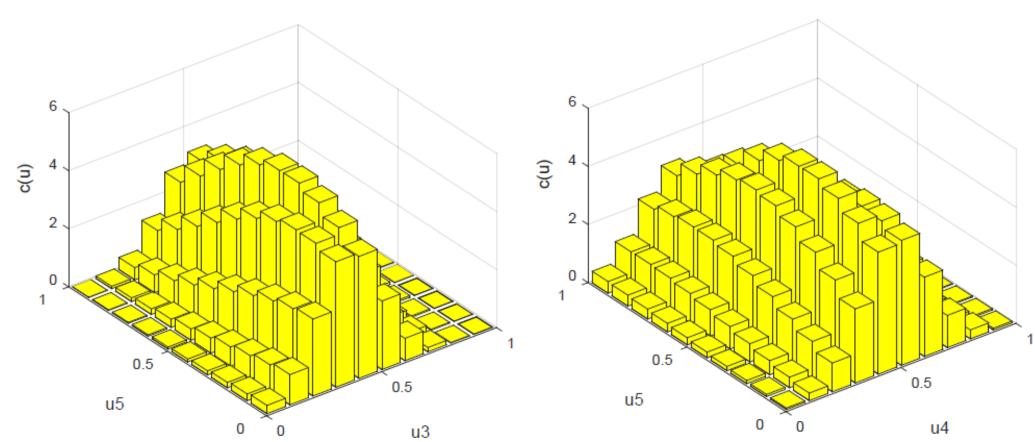

Figure 5: Two-dimensional projections (u3-u5; u4-u5) of density of the checkerboard copula, m=5, n=10, obtained by maximizing entropy.

We assumed that for 3 random variables W, X, and Y the empirical probability distribution functions $F_W(w)$, $F_X(x)$, $F_Y(y)$ are defined with 1000 observations. Assumptions for the sums of the random variables are defined in the following two cases.

Case 1.

For the random value $Z = W + X + Y$, the empirical probability distribution function $F_Z(z)$ is defined with K=16 observations $z_1, \ldots, z_{16}$. We solved an optimization problem and found a checkerboard copula on $n \times n \times n$ grid, where n=10. The 16 scenarios of the loss function $L(\boldsymbol{h}, \boldsymbol{j})$, defined in (13), were calculated as follows,

$$L(\boldsymbol{h}, j) = \frac{j}{16} - \sum_{i_1=1}^{10} \sum_{i_2=1}^{10} \sum_{i_3=1}^{10} \gamma_{i_1 i_2 i_3}(z_j)\ h_{i_1 i_2 i_3}\ , \quad j = 1, \ldots 16\,. \tag{39}$$

We use formula (33) for the approximate calculations of the coefficients $\gamma_{i_1 i_2 i_3}(z_j)$,

$$\gamma_{i_1 i_2 i_3}(z) \approx \begin{cases} n^{-1}\ , & \text{if } F_W^{-1}(a(i_1+1)) + F_X^{-1}(a(i_2+1)) + F_Y^{-1}(a(i_3+1)) \leq z\ , \\ 0\ , & \text{if } F_W^{-1}(a(i_1)) + F_X^{-1}(a(i_2)) + F_Y^{-1}(a(i_3)) \geq z\ , \\ \frac{1}{2}\, n^{-1}\ , & \text{otherwise}\ . \end{cases} \tag{40}$$

Further we formulate the error minimization problem with one sum function as defined in (21–24).

Optimization Problem (Case 1)

Find hyper-matrix $\boldsymbol{h} \in \mathbb{R}^3$ *minimizing an error function*

$$\min_{\boldsymbol{h}}\ err(L(\boldsymbol{h}, \boldsymbol{j})) \tag{41}$$

subject to constraints

$$\sum_{j_2=1}^{10} \sum_{j_3=1}^{10} h_{j_1 j_2 j_3} = 1, \quad j_1 = 1, \ldots, 10 \tag{42}$$

$$\sum_{j_1=1}^{10} \sum_{j_3=1}^{10} h_{j_1 j_2 j_3} = 1, \quad j_2 = 1, \ldots, 10 \tag{43}$$

$$\sum_{j_1=1}^{10} \sum_{j_2=1}^{10} h_{j_1 j_2 j_3} = 1, \quad j_3 = 1, \ldots, 10 \tag{44}$$

$$h_{j_1 j_2 j_3} \geqslant 0, \quad j_1, j_2, j_3 = 1, \ldots, 10 \tag{45}$$

We considered in objective (41) three functions: Mean Squared Error (15), Mean Absolute Error (16), and CVaR Absolute Error (17). Table 6 shows solutions for Optimization Problem (Case 1) with these error functions.

Table 6: Calculation results for Optimization Problem (Case 1) with different error functions.

Error Function	Optimal Value	R^2	Solution Time (sec)
Mean Squared	3.13E-03	0.987	1.45
Mean Absolute	1.56E-04	0.998	5.57
CVaR Absolute, $\alpha = 0.9$	3.13E-02	-	1.42
CVaR Absolute, $\alpha = 0.99$	3.66E-02	-	2.5

Figures 6, 7, 8, 9 shows the two-dimensional projections of density of checkerboard copula, obtained by minimizing Mean Squared, Mean Absolute, CVaR Absolute $\alpha = 0.9$, CVaR Absolute $\alpha = 0.99$ in Optimization Problem, Case 1. Two-dimensional projection of the density to coordinates u_{i_1}, u_{i_2} is done by fixing complementary components (not involved in the projection) at value 0.5 .

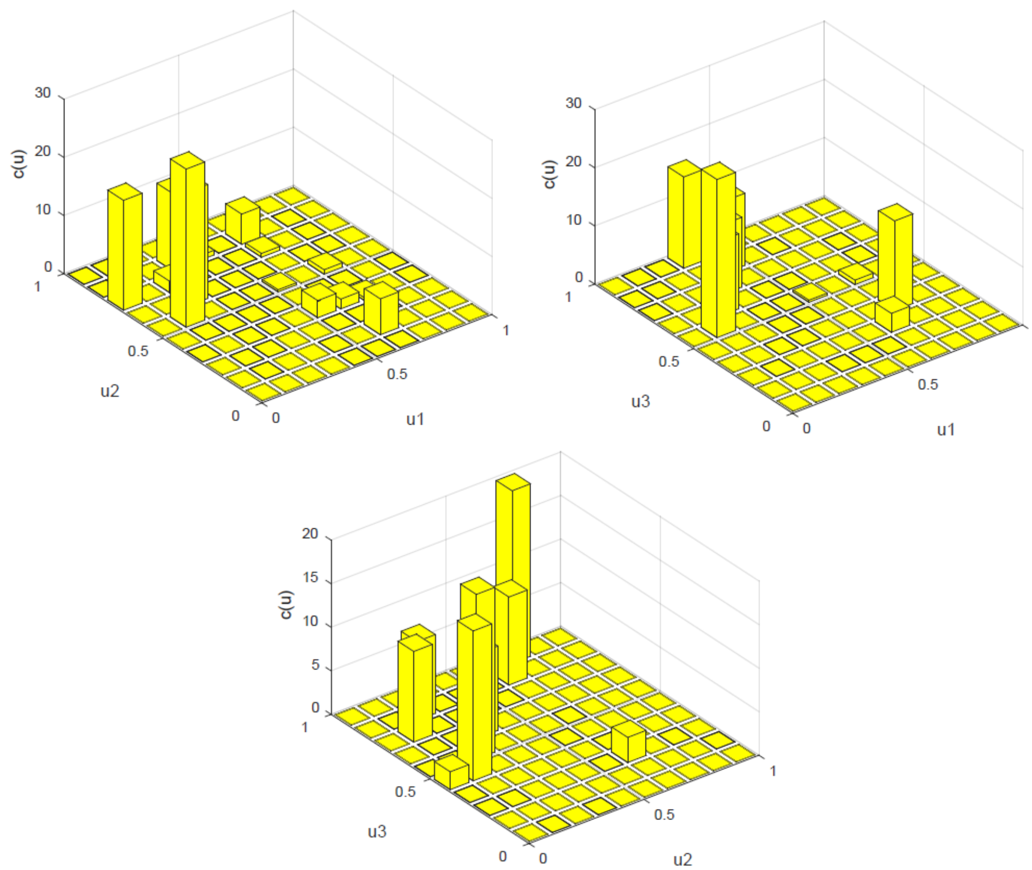

Figure 6: Two-dimensional projections of density of the checkerboard copula, obtained by minimizing Mean Squared Error in Case 1.

Case 2.

For the three sums of random values $Z_1 = W + X$, $Z_2 = W + Y$, $Z_3 = X + Y$, the empirical probability distributions $F_{Z_1}(z)$, $F_{Z_2}(z)$, $F_{Z_3}(z)$ are defined by K=16 observations for every sum. So, we have observations $z_1^1, \dots, s_{16}^1$ for Z_1, observations $z_1^2, \dots, s_{16}^2$, for Z_2, and observations $z_1^3, \dots, s_{16}^3$ for Z_3. Let us denote the following loss functions,

$$L_{i_1 i_2}(\boldsymbol{h}, j) = \frac{j}{16} - \sum_{i_1=1}^{10} \sum_{i_2=1}^{10} \gamma_{i_1 i_2}(z_j^1) \sum_{i_3=1}^{10} h_{i_1 i_2 i_3}\,, \quad j = 1, \dots 16, \tag{46}$$

$$L_{i_1 i_3}(\boldsymbol{h}, j) = \frac{j}{16} - \sum_{i_1=1}^{10} \sum_{i_3=1}^{10} \gamma_{i_1 i_3}(z_j^2) \sum_{i_2=1}^{10} h_{i_1 i_2 i_3}\,, \quad j = 1, \dots 16, \tag{47}$$

$$L_{i_2 i_3}(\boldsymbol{h}, j) = \frac{j}{16} - \sum_{i_2=1}^{10} \sum_{i_3=1}^{10} \gamma_{i_2 i_3}(z_j^3) \sum_{i_1=1}^{10} h_{i_1 i_2 i_3}\,, \quad j = 1, \dots 16\,. \tag{48}$$

We use formula (33) for the approximate calculations of the coefficients $\gamma_{i_1 i_2}(z_j^1)$, $\gamma_{i_1 i_3}(z_j^2)$, $\gamma_{i_2 i_3}(z_j^3)$,

$$\gamma_{i_1 i_2}(z) \approx \begin{cases} n^{-1}\,, & \text{if } F_W^{-1}(a(i_1+1)) + F_X^{-1}(a(i_2+1)) \le z\,, \\ 0\,, & \text{if } F_W^{-1}(a(i_1)) + F_X^{-1}(a(i_2)) \ge z\,, \\ \frac{1}{2} n^{-1}\,, & \text{otherwise}\,. \end{cases} \tag{49}$$

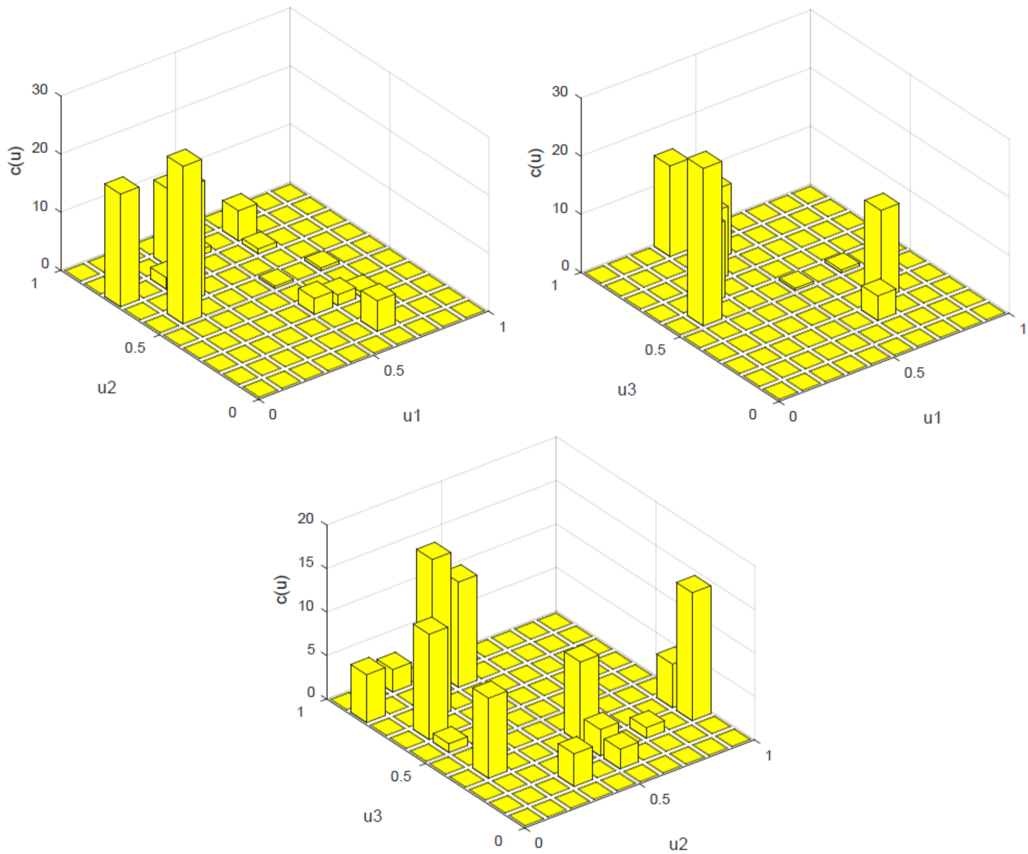

Figure 7: Two-dimensional projections of density of the checkerboard copula, obtained by minimizing Mean Absolute Error in Optimization Problem, Case 1.

$$\gamma_{i_1 i_3}(z) \approx \begin{cases} n^{-1}, & \text{if } F_W^{-1}(a(i_1+1)) + F_Y^{-1}(a(i_3+1)) \le z, \\ 0, & \text{if } F_W^{-1}(a(i_1)) + F_Y^{-1}(a(i_3)) \ge z, \\ \frac{1}{2} n^{-1}, & \text{otherwise}. \end{cases} \tag{50}$$

$$\gamma_{i_2 i_3}(z) \approx \begin{cases} n^{-1}, & \text{if } F_X^{-1}(a(i_2+1)) + F_Y^{-1}(a(i_3+1)) \le z, \\ 0, & \text{if } F_X^{-1}(a(i_2)) + F_Y^{-1}(a(i_3)) \ge z, \\ \frac{1}{2} n^{-1}, & \text{otherwise}. \end{cases} \tag{51}$$

Further we formulate the minimization problem with the weighted average of the error functions defined in (25).

Optimization Problem (Case 2)

Find hyper-matrix $\boldsymbol{h} \in \mathbb{R}^3$ minimizing weighted average of the error functions

$$\min_{\boldsymbol{h}} \frac{1}{3}(err(L_{i_1 i_2}(\boldsymbol{h}, j)) + err(L_{i_1 i_3}(\boldsymbol{h}, j)) + err(L_{i_2 i_3}(\boldsymbol{h}, j))) \tag{52}$$

subject to constraints

$$\sum_{j_2=1}^{10} \sum_{j_3=1}^{10} h_{j_1 j_2 j_3} = 1, \quad j_1 = 1, \dots, 10 \tag{53}$$

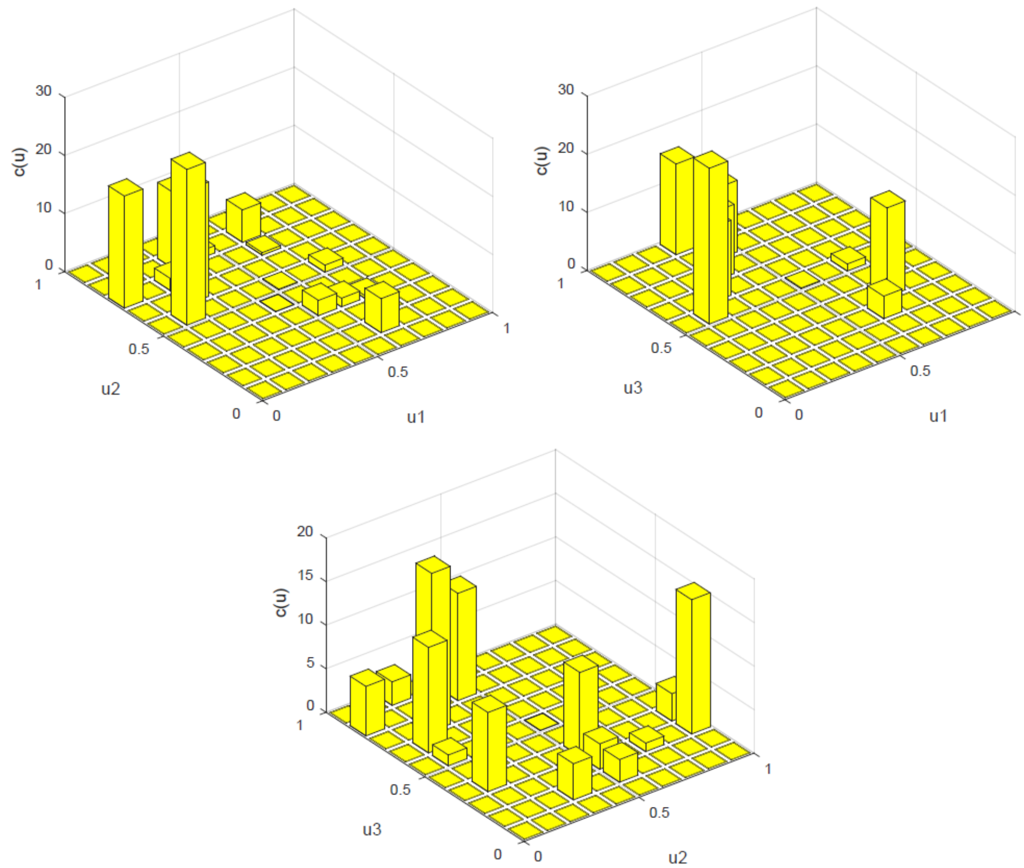

Figure 8: Two-dimensional projections of density of the checkerboard copula, obtained by minimizing CVaR Absolute $\alpha = 0.9$ Error in Optimization Problem, Case 1.

$$\sum_{j_1=1}^{10}\sum_{j_3=1}^{10} h_{j_1j_2j_3} = 1, \quad j_2 = 1,\ldots,10 \tag{54}$$

$$\sum_{j_1=1}^{10}\sum_{j_2=1}^{10} h_{j_1j_2j_3} = 1, \quad j_3 = 1,\ldots,10 \tag{55}$$

$$h_{j_1j_2j_3} \geqslant 0, \quad j_1, j_2, j_3 = 1,\ldots,10. \tag{56}$$

We considered in objective (52) three error functions defined in Section 3: Mean Squared, Mean Absolute, and CVaR Absolute Error. Optimization problems were solved with PSG. Table 7 shows results for the Optimization Problem (Case 2).

Table 7: Calculation results for Optimization Problem (Case 2) with different error functions.

Error Functin	Optimal Value	R^2	Solution Time (sec)
Mean Squared	1.01E-03	0.988	2.10
Mean Absolute	2.06E-02	0.918	1.70
CVaR Absolute, $\alpha = 0.9$	6.56E-02	-	2.02
CVaR Absolute, $\alpha = 0.99$	0.1	-	0.41

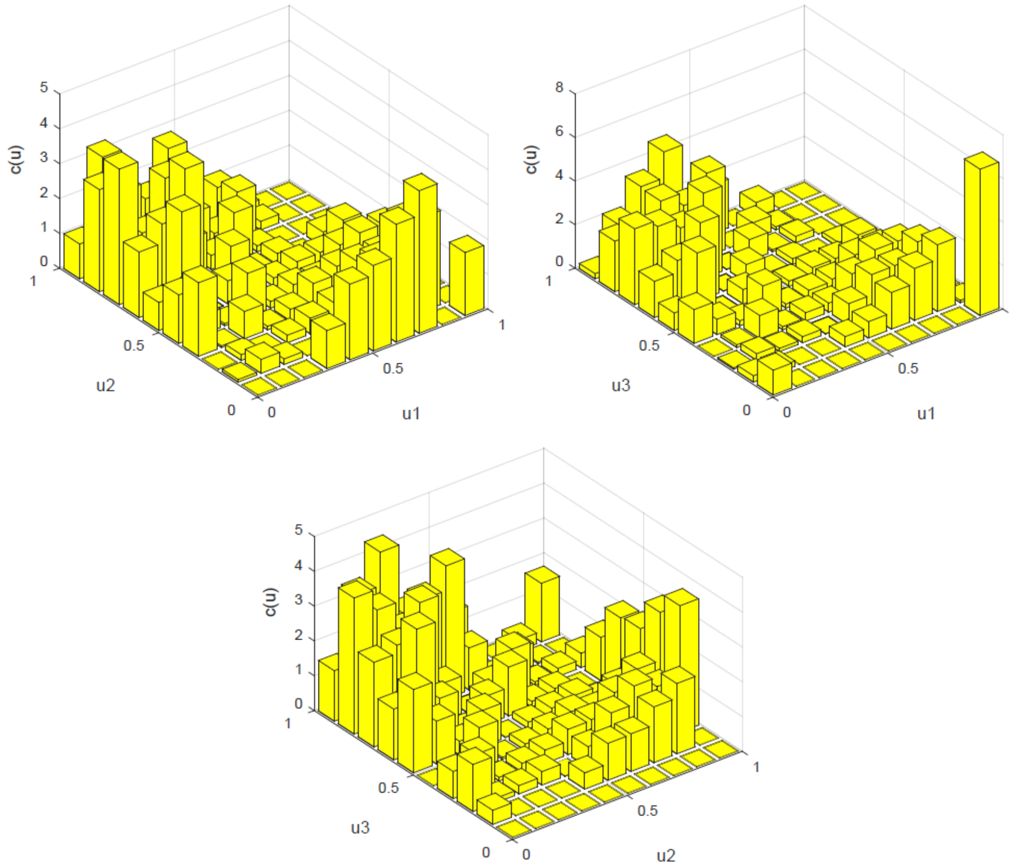

Figure 9: Two-dimensional projections of density of the checkerboard copula, obtained by minimizing CVaR Absolute $\alpha = 0.99$ Error in Optimization Problem, Case 1.

Figures 10, 11, 12, 13 show the two-dimensional projections of density of the checkerboard copula, obtained by minimizing Mean Absolute, Mean Squared, and CVaR Absolute $\alpha = 0.9$, $\alpha = 0.99$ in Optimization Problem, Case 2. Two-dimensional projection of the density to coordinates u_{i_1}, u_{i_2} is obtained by fixing complementary components (not involved in the projection) at value 0.5 .

5 Summary

We consider two setups for finding checkerboard copula, which link a multivariate distribution on a unit hyper-cube to their corresponding one-dimensional marginal distributions. A checkerboard copula is uniquely defined by a multiply-stochastic hyper-matrix. In the first setup Spearman Rho rank correlation coefficients are available. To find optimal values of elements of the hyper-matrix we maximized entropy subject to constraints, which match known Spearman Rho coefficients. With the second setup, distributions of sums of random variables and distributions of marginals are available. We developed a system of equations linking elements of a hyper-matrix with known observations of random variables and their sums. This system of equations is overspecified, therefore, we have used regression to find a hyper-matrix.

The case study was done using property and casualty insurance data. More importantly, the case study represents circumstances often faced by actuaries trying to build aggregate loss distributions across correlated classes of business where the objective is to make the correct representation of the dependencies observed in the data. The optimization problems were numerically solved with the AORDA Portfolio Safeguard (PSG) package, which has precoded entropy and error functions. Case study data, codes and results

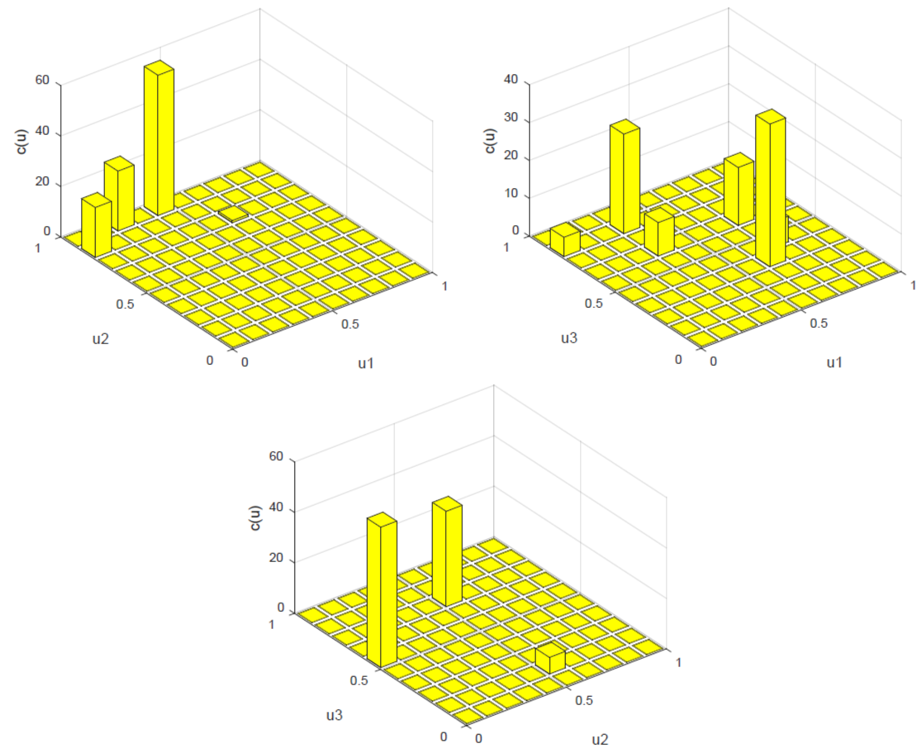

Figure 10: Two-dimensional projections of density of the checkerboard copula, obtained by minimizing the average of Mean Squared Errors in Case 2.

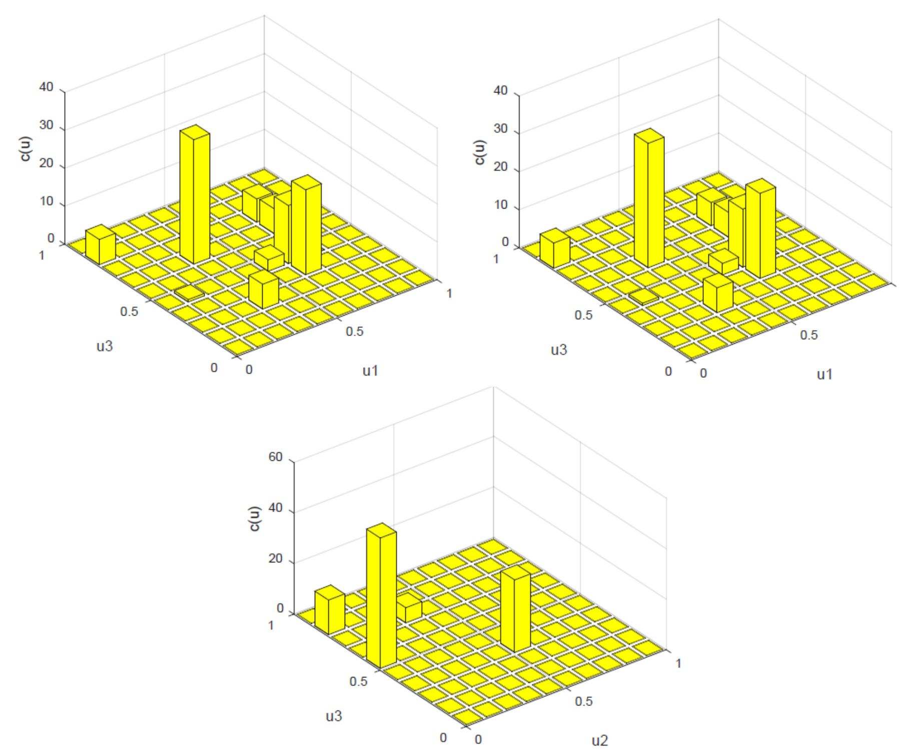

Figure 11: Two-dimensional projections of density of the checkerboard copula, obtained by minimizing the average of Mean Absolute Errors in Case 2.

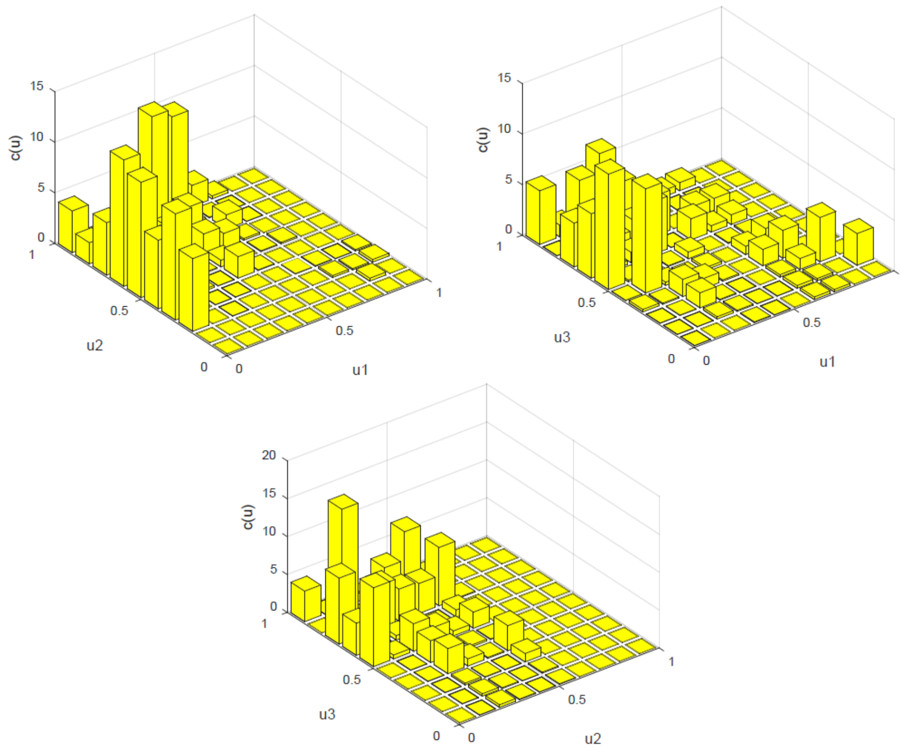

Figure 12: Two-dimensional projections of density of the checkerboard copula, obtained by minimizing the average of CVaR Absolute $\alpha = 0.9$ Errors in Case 2.

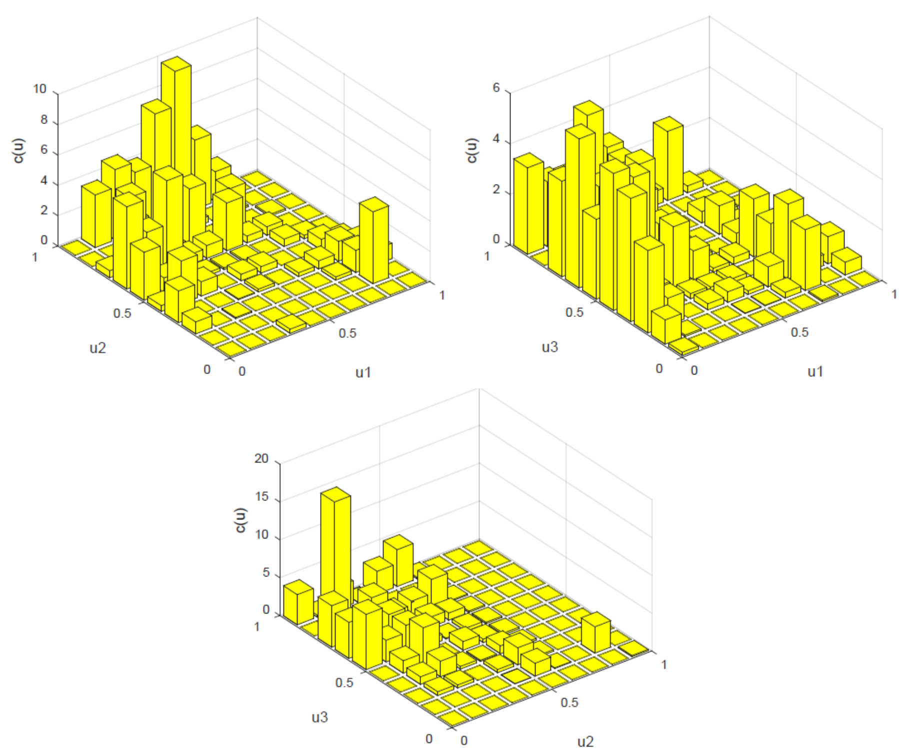

Figure 13: Two-dimensional projections of density of the checkerboard copula, obtained by minimizing the average of CVaR Absolute $\alpha = 0.99$ Errors in Case 2.

are posted at web and are available for verification (PSG is free for academic purposes).

Acknowledgments: We acknowledge that the first variant of the case study for the copula calibration with known sums of random variables was conducted by Dr. Konstantin Pavlikov during his PhD study at the University of Florida. Also, we acknowledge the help of Dr. Alexander Golodnikov with conducting the case study and preparing the paper.

References

[1] Bazaraa, M. S. and C. M. Shetty (1979). *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, New York.
[2] Boland, J., P. Howlett, J. Piantadosi, and R. Zakaria (2016). Modelling and simulation of volumetric rainfall for a catchment in the Murray-Darling basin. *ANZIAM J. 58*(2), 119–142.
[3] Borwein, J. M., P. Howlett, and J. Piantadosi (2014). Modelling and simulation of seasonal rainfall using the principle of maximum entropy. *Entropy 16*(2), 747–769.
[4] Borwein, J. M. and P. Howlett (2019). Checkerboard copulas of maximum entropy with prescribed mixed moments. *J. Aust. Math. Soc. 107*(3), 302–318.
[5] Borwein, J. M. and A. S. Lewis (2006). *Convex Analysis and Nonlinear Optimization.* Second edition. Springer, New York.
[6] Mafusalov, A. and S. Uryasev (2016). CVaR (superquantile) norm: Stochastic case. *European J. Oper. Res. 249*(1), 200–208.
[7] Nelsen, R. B. (2006). *An Introduction to Copulas.* Second Edition. Springer, New York.
[8] Piantadosi, J., J. Boland, and P. Howlett (2009). Generating synthetic rainfall on various timescales - daily, monthly and yearly. *Environ. Model. Assess. 14*(4), 431–438.
[9] Pfeifer, D., H. Awoumlac Tsatedem, A. Mändle, and C. Girschig (2016). New copulas based on general partitions-of-unity and their applications to risk management. *Depend. Model. 4*, 123–140.
[10] Pfeifer, D., A. Mändle, and O. Ragulina (2017). New copulas based on general partitions-of-unity and their applications to risk management (part II). *Depend. Model. 5*, 246–255.
[11] Piantadosi, J., P. Howlett, and J. Boland (2007). Matching the grade correlation coefficient using a copula with maximum disorder. *J. Ind. Manag. Optim. 3*(2), 305–312.
[12] Piantadosi, J., P. Howlett, and J. Borwein (2012). Copulas with maximum entropy. *Optim. Lett. 6*(1), 95–122.