Kybernetika

Vladislav Bína; Radim Jiroušek

A short note on multivariate dependence modeling

Kybernetika, Vol. 49 (2013), No. 3, 420--432

Persistent URL: http://dml.cz/dmlcz/143356

KYBERNETIKA — VOLUME *49* (2013), NUMBER 3, PAGES 420–432

A SHORT NOTE ON MULTIVARIATE DEPENDENCE MODELING

Vladislav Bína and Radim Jiroušek

As said by Mareš and Mesiar, necessity of aggregation of complex real inputs appears almost in any field dealing with observed (measured) real quantities (see the citation below). For aggregation of probability distributions Sklar designed his copulas as early as in 1959. But surprisingly, since that time only a very few literature have appeared dealing with possibility to aggregate several different pairwise dependencies into one multivariate copula.

In the present paper this problem is tackled using the well known Iterative Proportional Fitting Procedure. The proposed solution is not an exact mathematical solution of a marginal problem but just its approximation applicable in many practical situations like Monte Carlo sampling. This is why the authors deal not only with the consistent case, when the iterative procedure converges, but also with the inconsistent non-converging case. In the latter situation, the IPF procedure tends to cycle (when combining three pairwise dependencies the procedure creates three convergent subsequences), and thus the authors propose some heuristics yielding a "solution" of the problem even for inconsistent pairwise dependence relations.

Keywords: Frank copula, IPFP, entropy

Classification: 97K50, 94A17

1. INTRODUCTION

In his research papers, Milan Mareš focused on several important topics. His long-life interests were connected with the game theory and applications in economics. And it was this field of application what led him and his colleague and friend Radko Mesiar to study copulas and write a paper [9] where they described several tools for aggregation of complex quantities.

Abe Sklar introduced copulas in 1959 [11]. It took more than forty years to financial mathematicians to employ them in economic models. It was David X. Li, who first employed Gaussian copulas in financial models [8]. Since that time not many papers have been published in which more than two quantities interconnected with different types of dependence relations are considered. Let us illustrate the type of problems we have in mind by a simple discrete example.

Consider three binary random variables X, Y, Z with the distribution from Table 1. Computing its two-dimensional marginal distributions (see Table 2) one can immediately see that variables Y and Z are independent, and that there is a strong positive correlation

	$X=0$		$X=1$	
	$Z=0$	$Z=1$	$Z=0$	$Z=1$
$Y=0$	0.2	0.2	0.05	0.05
$Y=1$	0.0	0.1	0.25	0.15

Tab. 1. Three-dimensional binary distribution.

	$Z=0$	$Z=1$
$Y=0$	0.25	0.25
$Y=1$	0.25	0.25

	$X=0$	$X=1$
$Y=0$	0.4	0.1
$Y=1$	0.1	0.4

	$X=0$	$X=1$
$Z=0$	0.2	0.3
$Z=1$	0.3	0.2

Tab. 2. Two-dimensional marginal distributions.

between variables X and Y, while the correlation between variables X and Z is negative and much weaker. It means that in this example each pair of the considered variables is interconnected with another type of dependence (here we consider the independence to be a special type of dependence). We fully agree with the claim of Mareš and Mesiar [9] that the necessity of aggregation of complex real inputs appears almost in any field dealing with observed (measured) real quantities. To this claim we only want to add that when modeling a dependence of a group of variables we cannot rely upon the fact that each pair of the variables (or, generally a subgroup of the variables) is linked up by the same type of dependence as the other pairs of variables.

The same type of the problem was studied by Kjersti Aas et al. in [1] and Doris and Ernesto Schirmacher in [12]. They propose to construct complex copulas by decomposing them into a product of (two-dimensional) pair-copulas. From the point of view of Milan Mareš' papers it may be interesting that they illustrate their approach on examples from financial analysis and risk management. However, their approach, though general and mathematically correct, requires to specify copulas connecting conditional density functions. This step is, in our opinion, intuitively incomprehensible. We can hardly imagine a manager or financial analyst following the process they proposed and estimating copulas connecting conditional density functions. So the goal of this paper is to present an alternative approach for composing more-dimensional probability distribution from a system of two-dimensional copulas employing the famous Iterative proportional fitting procedure [4].

2. DESCRIBING UNCERTAINTY

As said above, to describe a dependence of random variables, A. Sklar introduced a general class of functions called *copulas* [11]. These functions can describe any type of dependence among random variables. Restricting ourself to two random variables, any two-dimensional distribution function can be expressed as a copula-combination of its one-dimensional marginal distributions. More precisely, let $F(X,Y)$ be a two-

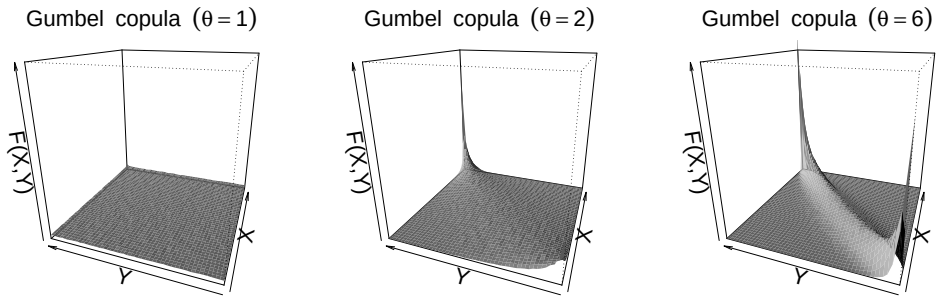

Fig. 1. Gumbel copulas.

dimensional cumulative distribution function and $F(X)$, $F(Y)$ be its respective marginal one-dimensional cumulative distribution functions. Then, due to the famous Sklar's theorem there exists a copula $C : [0,1] \times [0,1] \to [0,1]$ such that

$$F(X,Y) = C(F(X), F(Y)).$$

Similarly to most of other authors (see e. g. Hennessy and Lapan [5]), in this paper we will restrict our considerations to *Archimedean copulas*, i. e. the copulas that can be expressed with the help of a so called *generating function* $\psi : [0, +\infty) \to [0,1]$ in the following simple way

$$C(u,v) = \psi\left(\psi^{-1}(u) + \psi^{-1}(v)\right). \tag{1}$$

This restriction yields two advantages. First, as it is obvious from the formula (1), these functions are associative, and so one can, in case of need, describe the dependence of more than two random variables. The second advantage is connected with the fact that quite often, when solving a problem of practice, one does not have a detailed information about the dependence to be modeled. In most of situations one has only a subjective estimate of the type and strength of the dependence, and it is its strength that can be easily expressed by a parameter θ of the selected family of copulas. In this paper we will use Frank copulas, though it seems that in the problems of practice one can also consider e. g., Gumbel (Figure 1) or Clayton (Figure 2) copulas. As we can see from Figure 2, the latter ones are not symmetric with respect to the minor diagonal which may be desirable in some special situations. The Figures 1, 2 and 3 present three-dimensional surfaces of chosen Archimedean copulas using randomly generated points from the corresponding distribution.

The generating function for the Frank copulas is

$$\psi_\theta(u) = -\frac{1}{\theta}\ln\left(1 - (1 - e^{-\theta})e^{-u}\right),$$

and its inverse is

$$\psi_\theta^{-1}(u) = -\ln\left(\frac{e^{-\theta u} - 1}{e^{-\theta} - 1}\right).$$

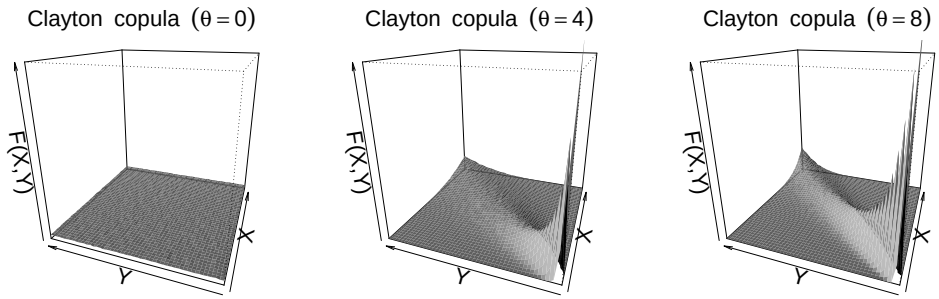

Fig. 2. Clayton copulas.

The parameter θ, which takes its value from[1] $(-\infty, +\infty)$, expresses here both the strength and also the type of the dependence. In the experiments we use the following five possibilities (see Figure 3):

- Strong direct proportion: $\theta = 12$;
- Weak direct proportion: $\theta = 4$;
- Independence: $\theta = 0$;
- Weak indirect proportion: $\theta = -4$;
- Strong indirect proportion: $\theta = -12$.

3. COMPLEX DEPENDENCE

When speaking about the Archimedean copulas in the preceding section we mentioned their advantage connected with the fact that they can easily be used to introduce the dependence of more than two variables:

$$C(u, v, w) = \psi\left(\psi^{-1}(u) + \psi^{-1}(v) + \psi^{-1}(w)\right).$$

However, from the practical point of view this form of dependence is rather unrealistic since such a copula introduces the same type of dependence between all the three pairs of the considered variables. As said in Introduction, in practical situations it happens quite often that considering a group of variables one has to introduce different types of dependence between different pairs of variables. So, the task we are going to tackle is how to model situations when one considers several (usually three, four, or five at maximum) variables with given pairwise dependencies described by different copulas. Here it does not matter whether the given dependencies are defined for all pairs of variables or just for some of them.

[1]The reader perhaps noticed that for $\theta = 0$ the functions ψ_0 and ψ_0^{-1} are not defined. However, we define C_{ψ_0} to be the *independence* copula, i. e., $C_{\psi_0} = \lim_{\theta \to 0} C_{\psi_\theta}$.

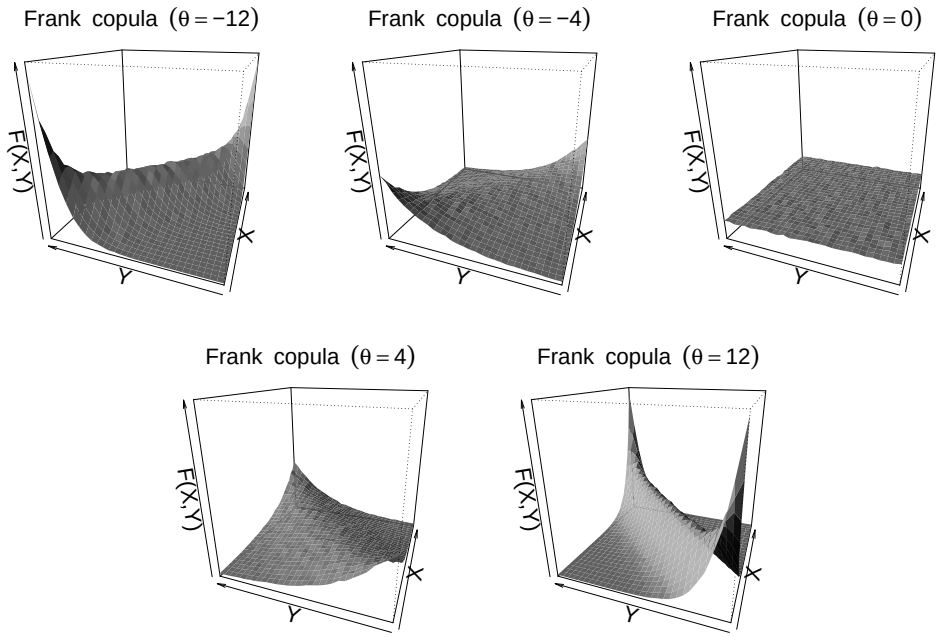

Fig. 3. Frank copulas.

The simplest situation occurs when the pairs of variables with the predefined dependence relations can be ordered in the way, that each pair contains at least one variable that does not occur in the preceding pairs (in this case one can set up from the copulas either a decomposable model [6] or a compositional model [7]). In a general case, as said in Introduction, we propose to solve this problem with the help of the famous *Iterative Proportional Fitting Procedure* [4], which can be described in the following simple way.

For the sake of simplicity we will consider here only three variables X, Y, and Z, and assume their pairwise dependencies be expressed by two-dimensional density functions $f_1(X,Y)$, $f_2(X,Z)$ and $f_3(Y,Z)$. The generalization for more variables and greater number of predefined pairwise dependence relations is straightforward.

As the name suggests, the IPF procedure is iterative. So we will construct an infinite sequence of three-dimensional density functions $g_0(X,Y,Z)$, $g_1(X,Y,Z)$, $g_2(X,Y,Z)$, $g_3(X,Y,Z),\ldots$ and the desired result $g^\star(X,Y,Z)$ will be determined by its limit

$$g^\star(X,Y,Z) = \lim_{n\to\infty} g_n(X,Y,Z),$$

which will be a density function having all the three given functions f_1, f_2, f_3 for its marginals.

Let the starting density function g_0 be defined as an independent product of the uniform one-dimensional density functions. Then the computational process is defined

by the following simple formulae:[2].

$$\begin{aligned}
g_1(X,Y,Z) &= f_1(X,Y)\cdot g_0(Z|X,Y),\\
g_2(X,Y,Z) &= f_2(X,Z)\cdot g_1(Y|X,Z),\\
g_3(X,Y,Z) &= f_3(Y,Z)\cdot g_2(X|Y,Z),\\
g_4(X,Y,Z) &= f_1(X,Y)\cdot g_3(Z|X,Y),\\
g_5(X,Y,Z) &= f_2(X,Z)\cdot g_4(Y|X,Z),\\
g_6(X,Y,Z) &= f_3(Y,Z)\cdot g_5(X|Y,Z),\\
g_7(X,Y,Z) &= f_1(X,Y)\cdot g_6(Z|X,Y),\\
g_8(X,Y,Z) &= f_2(X,Z)\cdot g_7(Y|X,Z),\\
&\vdots
\end{aligned} \tag{2}$$

Unfortunately, the application of this iterative process is connected with a couple of theoretical problems.

First, Csiszár proved in [3] a convergence of this process, however his proof is valid for discrete random variables, only. The convergence in a special case of continuous variables was analyzed by Rüschendorf [10], but general necessary and sufficient conditions for the convergence of this process in case of continuous variables are not known. Moreover, a trivial necessary condition for this convergence is an existence of a three-dimensional density function $f(X,Y,Z)$, for which f_1, f_2 and f_3 are its marginal densities. Unfortunately, in most of situations there is no simple way how to recognize whether such a joint density function exists. This is why we look for the desired joint density function in the following indirect way.

In this paper we do not study an exact mathematical solution of the underlying marginal problem. We look for a model expressing a subjective knowledge of a manager or a financial engineer. We said above that when selecting a model describing the dependence of two variables we can choose Frank or Gumbel copulas. Because of use of expert estimates[3], quite naturally, we can hardly distinguish whether we should consider Frank copula with parameter $\theta = 4.4$ or $\theta = 3.5$. Therefore we do not make a great harm when discretizing the problem in the way that we consider finite valued variables instead of continuous ones. In Figures 4 and 6 one can see that when considering variables with 20 values the quality of discrete model approaches the continuous one (in fact experiments show that in many situations one can do with considering 10-valued variables). We studied the influence of the number of categories on the convergence of IPF procedure. Surprisingly, we did not detect any impact and therefore we did not focus on the choice of grid coarseness parameter. But in a general case, it can be chosen according to the accuracy of this representation in comparison to the original model.

Application of the Iterative proportional procedure to discrete random variables is well explored and one can employ results of Csiszár [3] and Asci and Piccioni [2] saying that

[2] $g_i(Z|X,Y)$ denotes the conditional density function for which $g_i(X,Y,Z) = g_i(Z|X,Y)\cdot g_i(X,Y)$

[3] In case of estimating the parameter values from data we can employ e. g. maximum likelihood or minimum distance estimators (see Weiß [14]).

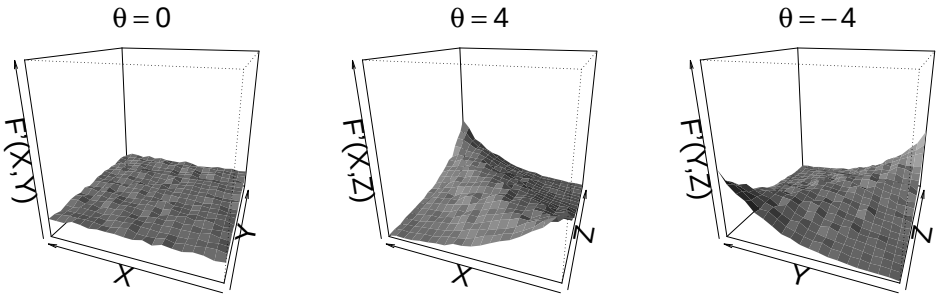

Fig. 4. Discretized Frank copulas for variable pairs X and Y, X and Z, Y and Z: A case of weaker dependencies ($\theta = 0$, $\theta = 4$, $\theta = -4$).

- since we use g_0 uniform, the procedure converges if and only if there exists a distribution having the given marginals;
- if the procedure converges than it yields a maximum entropy distribution with the given marginals;
- if there does not exist a distribution with the given marginals then the procedure tends to cycle (in this case we suggest to consider as a result of the process a *center* of the cycle – see Section 4).

So the computational process constructing a probability distribution representing the required dependence relations among the parameters consists of the following four steps.

1. Choose the pairs of variables for which you want to specify a type of dependence (recall that here we understand *independence* as a precisely defined type of a dependence). For each such pair of variables specify a value of parameter θ expressing the desired type of dependence by the respective Frank copula.
2. Discretize the two-dimensional distributions represented by the copulas.
3. Starting with the uniform distribution apply IPFP to all the discrete two-dimensional probability distributions.
4. If the process does not converge specify a center of the cycle and check whether its marginals do not differ from the copulas given in Step 1 in an undesirable way. If this is the case then apply some of the heuristics proposed in Section 4.

Example. Let us consider a situation similar to the example from Introduction. Consider three random variables X, Y and Z and assume that because of some reason or another we believe that variables X and Y are independent, variables X and Z are weakly positively correlated, and finally Y and Z are weakly negatively correlated. So,

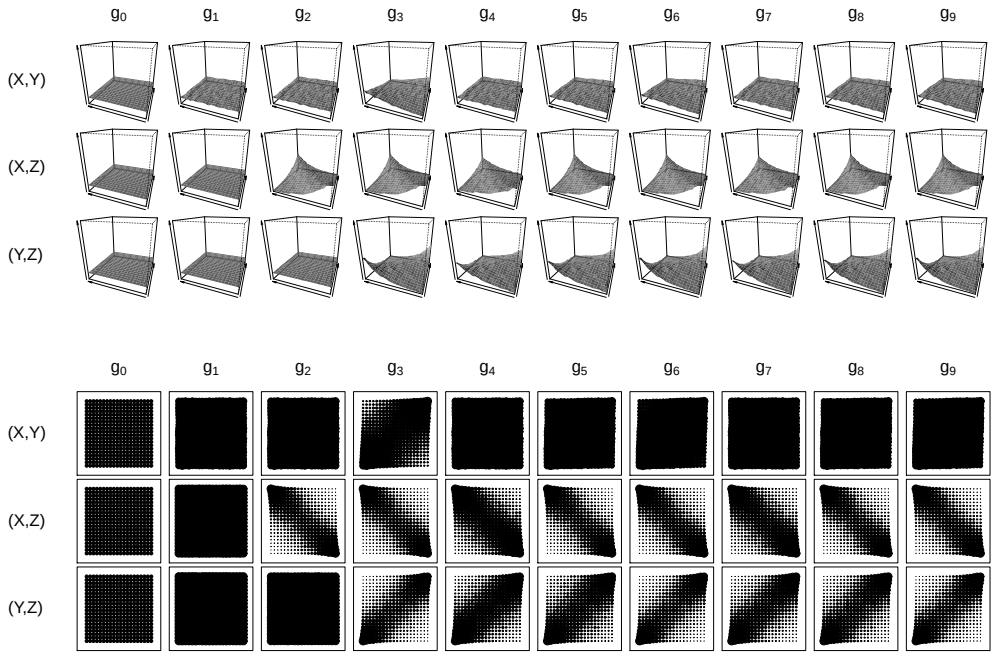

Fig. 5. Two ways of presentation of IPFP: A convergent case.

we have enough reasons to model the situation with the help of Frank copulas as indicated in Figure 4: we choose the parameter $\theta = 0$ for the variables X and Y, $\theta = 4$ for X and Z, and for Y and Z $\theta = -4$.

After application of three cycles (nine iterative steps) of IPFP we can see that the process converges rapidly to the stable solution (for the projections of the distributions computed during this process see Figure 5).

4. HEURISTICS FOR THE CASE WHEN THE PROCESS FAILS

Let us now explain more exactly, what we understand when saying that the sequence of density functions (2) converges to a cycle. By this we express the fact that though there does not exist $\lim_{n\to\infty} g_n$, all the three subsequences

$$g_1, g_4, g_7, \ldots, g_{3n+1}, \cdots$$
$$g_2, g_5, g_8, \ldots, g_{3n+2}, \cdots$$
$$g_3, g_6, g_9, \ldots, g_{3n+3}, \cdots$$

have their limits [13]. Denote them (for $i = 1, 2, 3$)

$$g_i^\star = \lim_{n\to\infty} g_{3n+i}.$$

Since the two-dimensional density function f_i is marginal to all the three-dimensional density functions from the subsequence $g_i, g_{3+i}, g_{6+i}, \ldots, g_{3n+i}, \ldots$, it is obvious that f_i is marginal also to $g_i^\star$. Moreover, it can be deduced from the Csiszár's results ([3]) that $g_2^\star$ is an I-projection of $g_1^\star$ into the set of all density functions having f_2 for its marginal in the sense that

$$g_2^\star = \arg \min_{h \in \Delta(f_2)} \{Div(h; g_1^\star)\},$$

where $\Delta(f_2)$ denotes the set of all the three-dimensional density functions $h(X, Y, Z)$ for which $h(X, Z) = f_2(X, Z)$, and Div denotes the famous Kullback-Leibler divergence (crossentropy) of functions h and $g_1^\star$:

$$Div(h; g_1^\star) = \sum_{(x,y,z)} h(x, y, z) \cdot \log \frac{h(x, y, z)}{g_1^\star(x, y, z)}.$$

Similarly, $g_3^\star$ is an I-projection of $g_2^\star$ into $\Delta(f_3)$ and $g_1^\star$ is an I-projection of $g_3^\star$ into $\Delta(f_1)$. All this theoretical knowledge helps us to select a proper way how to proceed when IPFP converges to a cycle.

As we said in the previous section, this situation occurs when the chosen density functions (or more precisely their discretized versions) are not consistent in the sense that there does not exist a three-dimensional density function $h(X, Y, Z)$ having all the three functions f_1, f_2, f_3 for its marginals. In other words it means that we have to release the system of dependence relations we have encoded into the functions f_1, f_2, f_3. There is a great variety of ways how to do it. We can revoke some of the dependencies, either to decrease the number of predefined density functions, or to decrease their strength, and start the computations with newly predefined marginal functions. Another possibility is to choose an appropriate representative from the set of all convex combinations of the density functions $g_1^\star$, $g_2^\star$, $g_3^\star$.

There is an abundant variety of approaches enabling this choice. For example, one can use the method based on minimization of an I-aggregate as defined by Vomlel (see [13]). But at this moment we just want to advise against application of another possibility, namely the use of the maximum entropy principle. As it will be evident from the following example, its application in this situation virtually corresponds to preferring weak dependencies and to the exclusion of the strong ones. As the contribution is not focused on this task, we do not want to discuss the details, and in the case of irresolution we propose yet another very simple possibility. Namely a choice of an arithmetic mean of the density functions $g_1^\star$, $g_2^\star$, $g_3^\star$.

Naturally, after selecting a resulting representative density function (for nonnegative k, l such that $k + l \leq 1$)

$$h_{k,l}^\star(X, Y, Z) = k \cdot g_1^\star(X, Y, Z) + l \cdot g_2^\star(X, Y, Z) + (1 - k - l) \cdot g_3^\star(X, Y, Z) \qquad (3)$$

we recommend to check whether its marginals $h_{k,l}^\star(X, Y), h_{k,l}^\star(X, Z)$ and $h_{k,l}^\star(Y, Z)$ sufficiently well represent the required pairwise dependencies.

Example. As in the previous example we consider three variables X, Y, Z; this time with strong direct proportion for the pair X, Z and weak indirect proportion in the case of Y, Z – see Figure 6.

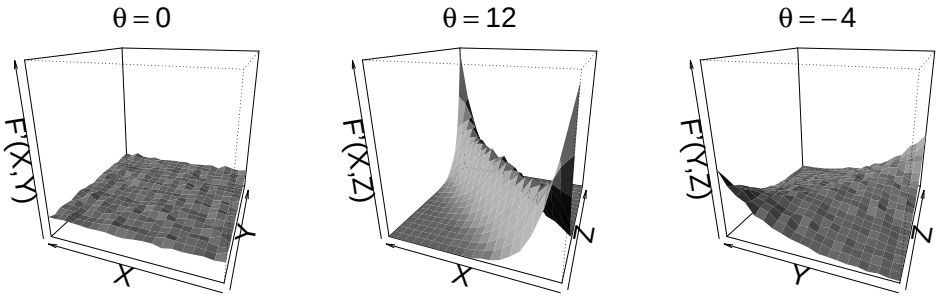

Fig. 6. Frank copulas for pairs of variables X and Y, X and Z, Y and Z: A case of stronger dependencies ($\theta = 0$, $\theta = 12$ and $\theta = -4$).

Again the three cycles of IPFP were performed and we can see that the process converges to a cycle (periodically repeating projections are depicted in Figure 7).

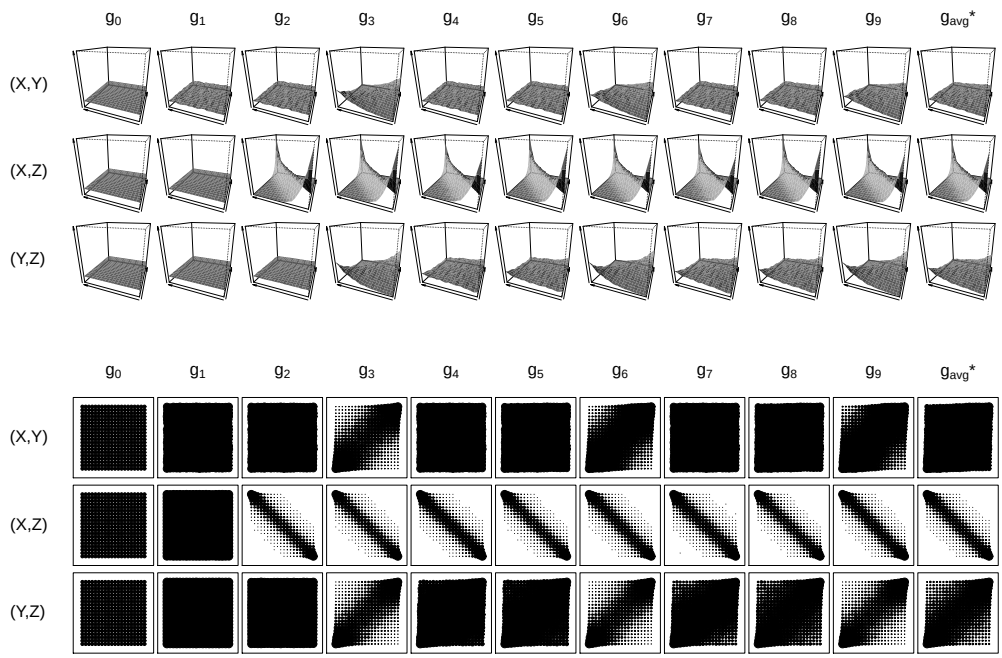

Fig. 7. Two ways of IPFP presentation: Convergence to a cycle augmented by the center of the cycle (an average).

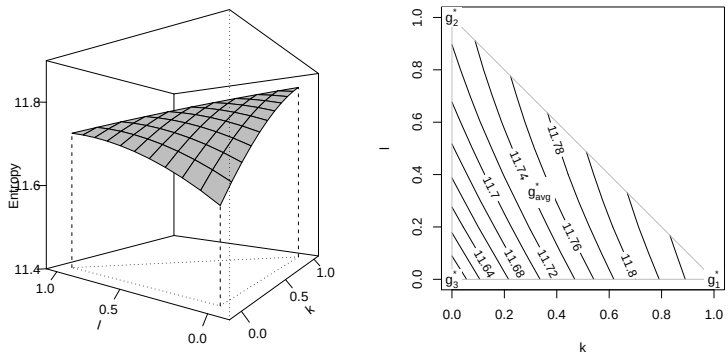

Fig. 8. Entropy of subsequences' limits and their convex combinations in dependence on two parameters k and l.

As said above, in this case the result of IPFP can be represented by the three limits (for $i = 1, 2, 3$): $g_i^\star = \lim_{n\to\infty} g_{i+3n}$, and one can consider all their convex combinations as a result of the computational process.

Since we assume no additional information, a maximum entropy principle could seem to be a natural choice. But let us have a look at the entropy of the limits $g_1^\star$, $g_2^\star$, $g_3^\star$, and their convex combinations (Figure 8). Note that any convex combination is here expressed using only two parameters k and l (as in formula (3)).

We can see that the convex combination with maximal entropy is $h^\star_{1,0}(X, Y, Z) = g_1^\star$. This is because the first copula has the highest entropy and this independence copula is marginal to all the densities from the subsequence g_{1+3n}, and therefore also to $g_1^\star$. Hence, we see that the convex combination with maximum entropy tends to prefer weaker dependencies. Therefore we suggest to choose simply the average of the distributions $g_1^\star$, $g_2^\star$, $g_3^\star$, i. e., $h^\star_{\frac{1}{3},\frac{1}{3}}(X, Y, Z)$. The values of entropy for the distributions from this example are the following:

$$\begin{aligned} H(g_1^\star) &= 11.860, & H(g_2^\star) &= 11.727, \\ H(g_3^\star) &= 11.578, & H(h^\star_{\frac{1}{3},\frac{1}{3}}) &= 11.748. \end{aligned}$$

5. CONCLUSIONS

We have presented an alternative approach to model a more-dimensional copula in case when several two-dimensional copulas (i. e., marginal copulas to the looked for more-dimensional one) are specified. For this, we have proposed to employ the famous Iterative Proportional Fitting Procedure. At this moment we want to conclude the paper with the two important remarks:

- When explaining the approach we assumed that only two-dimensional copulas are given. This was just for the simplicity sake. It is obvious that the approach can be generalized for the case when also some copulas of higher dimension are

given. Naturally, because of necessity to store representation of functions g_i in a computer memory one cannot expect to employ the approach for, let us say, ten random variables.

- The presented approach is not an exact mathematical solution of a marginal problem (for this, the reader is referred e. g. to [6]. The presented approach is designed to get a simple and lucid tool applicable in case that one has subjective rough estimates of the dependence relations between couples (or, triples, ...) of random variables and looks for the way how to employ this knowledge in a method like Monte Carlo.

ACKNOWLEDGEMENT

The research was partially supported by GA ČR under grant no. 403/12/2175 and University of Economics in Prague under projects no. F6/8/2012 and F6/12/2013.

(Received July 6, 2012)

REFERENCES

[1] K. Aas, C. Czado, A. Frigessi, and H. Bakken: Pair-copula construction of multiple dependence. Insurance Math. Econom. *44*, 2 (2009), 182–198.

[2] C. Asci and M. Piccioni: A note on the IPF algorithm when the marginal problem is unsolvable. Kybernetika *39* (2003), 6, 731–737.

[3] I. Csiszár: I-divergence geometry of probability distributions and minimization problems. Ann. Probab. *3* (1975), 146–158.

[4] W. E. Deming and F. F. Stephan: On a least square adjustment of a sampled frequency table when the expected marginal totals are known. Ann. Math. Statist. *11* (1940), 427–444.

[5] D. A. Hennessy and H. E. Lapan: The use of Archimedean copulas to model portfolio allocations. Math. Finance *12* (2002), 2, 143–154.

[6] R. Jiroušek: Solution of the marginal problem and decomposable distributions. Kybernetika *27*, 5 (1991), 403–412.

[7] V. Kratochvíl: Characteristic properties of equivalent structures in compositional models. Internat. J. Approx. Reasoning *52* (2011), 5, 599-612.

[8] D. X. Li: On default correlation: A copula function approach. J. Fixed Income *9* (2000), 4, 43–54.

[9] M. Mareš and R. Mesiar: Aggregation of complex quantities. In: Proceedings of AGOP'2005. International Summer School on Aggregation Operators and Their Applications (R. Mesiar, G. Pasi, and M. Faré, eds.), Universitá della Svizzeria Italiana, Lugano 2005, pp. 85–88.

[10] L. Rüschendorf: Convergence of the iterative proportional fitting procedure. Ann. Statist. *23* (1995), 4, 1160–1174.

[11] A. Sklar: Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris *8* (1959), 229–231 .

[12] D. Schirmacher and E. Schirmacher: Multivariate Dependence Modeling Using Pair-copulas. Technical Report, Society of Acturaries, Enterprise Risk Management Symposium, Chicago 2008.

[13] J. Vomlel: Integrating inconsistent data in a probabilistic model. J. Appl. Non-Classical Logics *14* (2004), 3, 367–386.

[14] G. N. F. Weiß: Copula parameter estimation: numerical considerations and implications for risk management. J. Risk *13* (2010), 1, 17–53.

Vladislav Bína, Faculty of Management in Jindřichův Hradec, University of Economics in Prague, Jarošovská 1117/II, 377 01 Jindřichův Hradec. Czech Republic.
e-mail: bina@fm.vse.cz

Radim Jiroušek, Faculty of Management in Jindřichův Hradec, University of Economics in Prague and Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: radim@utia.cas.cz