GENERATING SIMULATION INPUT WITH APPROXIMATE COPULAS

Feras Nassaj
Johann Christoph Strelen
Rheinische Friedrich-Wilhelms-Universitaet Bonn
Institut fuer Informatik IV
Roemerstr. 164, 53117 Bonn, Germany

KEYWORDS

Simulation, Input Modeling, Dependency, Multivariate Random Numbers, Generation

ABSTRACT

Copulas are used in finance and insurance for modeling stochastic dependency. They comprehend the entire dependence structure, not only the linear correlations. Here they serve the purpose to analyze measured samples of random vectors, to estimate a multivariate distribution for them, and to generate random vectors with this distribution. This can be applied as well to time series.

INTRODUCTION

Stochastic models and discrete simulation are indispensable means for the quantitative analysis of systems. It is well known that missing to carefully model the influences from outside, especially the load, may lead to wrong results and ultimately to wrong decisions based on the simulation results. One reason for bad load models may be to ignore dependencies, i.e. to use independent random variables instead of proper commonly distributed random vectors or stochastic processes.

Influence from outside of the model like load or failure of system components can be incorporated into the model using observed traces or input models, namely random variables, random vectors, or stochastic processes. Data from traces can be used directly. If input is modeled, data are realisations of the model.

The use of random variates is well understood and common since long time, the use of generated random vectors and stochastic processes is much more difficult, not so popular, a topic of actual research.

In this paper, we propose to use copulas for the analysis of observed data and for the generation of dependent random variates and time series. The use of copulas is common in finance and insurance.

The copula of a multivariate distribution describes its dependence structure completely, not only the correlations of the random variables. It is uncoupled from the marginal distributions which can be modeled as empirical distributions or fitted standard distributions.

The use of copulas might make a difficult task, finding a multivariate distribution, more facile by performing two easier tasks. The first step is modeling the marginal distributions, the second consists in estimating the copula. Moreover, it is quite simple to generate random vectors with copulas.

In our approach, the marginal distributions might be modeled as empirical distributions or as theoretical distributions as usual. However, we estimate the copula as a frequency distribution, which is not common. Usually one of the known families of copulas is fitted. There are many such families, see e.g. (Nelsen 1998), but most of these families are for only two dimensions. For simulation, more dimensions might be needed. Moreover, as remarked in (Blum and Dias and Embrechts 2002), fitting a copula is essentially as difficult as estimating the joint distribution in the first place. Thirdly, different families of copulas account for different kinds of dependence. Hence, the input modeler must choose the family according to the actual dependence nature. In contrast, an empirical copula incorporates the dependence form automatically. For these reasons, we use some kind of empirical copulas (the frequency distribution) instead of fitting families of copulas.

A chi-square test is proposed for the evaluation of the goodness of the fitted approximate distribution.

The new technique contrasts with other proposed input models. For example, autoregressive processes (AR) model allow to model linear dependencies in time series with Gaussian random variables. They are conveniently fitted to measured data with the linear Yule-Walker equations.

ARTA-like models (ARTA (Cario and Nelson 1996) for univariate time-series, NORTA (Cario and Nelson 1997) for random vectors, VARTA (Biller and Nelson 2003) for processes of random vectors) allow also to model linear dependencies, more over, they allow for general distributions by means of a Gaussian AR or a multivariate Gaussian random variable as basis whose random variables are transformed into the desired distributions. The correlations of the basis process are different from the desired correlations. Therefore, a transformation is required. Sometimes this transformation results in unfeasible correlation matrices of the basis process (Ghosh

and Henderson 2002b), the *defective matrix problem.*
TES processes (Melamed 1997) rely on empirical distributions of the random variables. They comprise lag 1 correlations. The interactive software system TEStool serves the purpose of fitting measured data to a TES process.
AR, ARTA-like, and TES processes as input modeling approaches for random vectors and time-series consider only the linear correlations, not the whole dependence structure. In contrast, copulas take into account the entire dependency, hence this new technique as well.
In (Nassaj and Strelen 2005) we propose some kind of nonlinear non-Gaussian autoregressive processes. The dependence structure is more general, nonlinear dependencies are accounted for. The distributions of the random variables are general. The procedure of fitting to measured date is done in two successive steps. The first one for the dependence structure applies optimization with respect to some parameters. The second one concerns the distribution of univariate random variables. This separation is similar to the copula approach. However, this procedure requires knowledge or assumption about the type of dependencies. This problem does not appear in the approach we describe in this paper.
In the next section, some material about copulas is provided. Section 3 describes the procedure of building an approximate distribution which fits measured data, and how random vectors are generated from this distribution. The chi-square test for the evaluation of the distribution is given in section 4. Section 5 contains some examples.

COPULAS

A compact definition of copulas is given in (Pfeifer and Neslehova 2003):
Definition A *copula* is a function C of D variables on the unit D-cube $[0,1]^D$ with the following properties:
1. The range of C is the unit interval $[0,1]$
2. $C(\mathbf{u})$ is zero for all $\mathbf{u} \in [0,1]^D$ for which at least one coordinate equals zero
3. $C(\mathbf{u}) = u_d$ if all coordinates of $\mathbf{u}$ are 1 except the d-th one
4. C is D-increasing in the sense that for every $\mathbf{a} \leq \mathbf{b}$ in $[0,1]^D$ the measure ΔC_a^b assigned by C to the D-box $[\mathbf{a},\mathbf{b}] = [a_1,b_1] \times ... \times [a_D,b_D]$ is nonnegative, i.e.

$$\Delta C_a^b := \sum_{(\epsilon_1,...,\epsilon_D)\in\{0,1\}} (-1)^{\epsilon_1+...+\epsilon_D} C\Big(\epsilon_1 a_1 + (1-\epsilon_1)b_1, ..., \epsilon_D a_D + (1-\epsilon_D)b_D\Big) \geq 0.$$

In fact, a copula is a multivariate distribution function for the random vector $\mathbf{U} = (U_1, ..., U_D)$ with univariate uniform margins restricted to the unit D-cube. All partial derivatives exist almost everywhere, hence the conditional distribution functions and the density as well.
The key theorem due to Sklar clarifies the relations of dependence and the copula of a distribution:
Theorem (Sklar): Let F denote a D-dimensional distribution function with margins $F_1, ..., F_D$. Then there exists a D-copula C such that for all real $\mathbf{z} = (z_1, ..., z_D)$,

$$F(\mathbf{z}) = C\Big(F_1(z_1), ..., F_D(z_D)\Big).$$

If all the margins are continuous, then the copula is unique; in general, it is determined uniquely on the ranges of the marginal distribution functions. Moreover, if we denote by $F_1^{-1}, ..., F_D^{-1}$ the generalized inverses of the marginal distribution functions, then for every $\mathbf{u} = (u_1, ..., u_D)$ in the unit D-cube,

$$C(\mathbf{u}) = F\Big(F_1^{-1}(u_1), ..., F_D^{-1}(u_D)\Big).$$

For a proof, see (Nelsen 1998). In the next section, we define pseudo-inverses of distribution functions.
Multivariate random numbers $(z_1, ..., z_D)$ can be generated using copulas. First, we consider the special case of two dimensions:
1. Generate independent random numbers u_1 and u_2, uniform on $(0,1)$.
2. Use the pseudo-inverse of the conditional distribution function $C_2(u_2|U_1 = u_1) = P\{U_2 \leq u_2 | U_1 = u_1\}$ for the generation of the random number u_2:

$$u_2 = C_2^{-1}(v|U_1 = u_1).$$

The conditional distribution function is equal to the partial derivative $\frac{\partial}{\partial u_1} C(u_1, u_2)$.
3. Univariate random variates can be generated with the inverse distribution function method $z_1 = F_1^{-1}(u_1)$ and $z_2 = F_2^{-1}(u_2)$. z_1 and z_2 are the elements of the desired random vector.
The generalization to higher dimensions is straightforward, the generation of the dependent u_d can be done in the usual way. See (Law and Kelton 2000),page 479, for example.

THE APPROXIMATE MULTIVARIATE DISTRIBUTION

We begin with a sketch of the method to fit an approximate multivariate distribution to data samples, the precise algorithm is presented subsequently.
A. Building the approximate distribution $\mathcal{A}$
1. Approximations $F_d(x)$, $d = 1, ..., D$, of the unknown marginal distribution functions are built from the given sample. This can be empirical distribution functions of some kind, or fitted standard distributions like exponential, Weibull etc.
2. The observed sample points $\mathbf{z}_i$, $i = 1, ..., n$, are transformed into points $\mathbf{u}_i$ of the unit D-cube $[0,1]^D$ by means of the marginal distribution functions.

3. The density of the approximate copula, that is, the density of the $\mathbf{u}_i$, is estimated. To this end, the D-cube is partitioned into sub-cubes. In each sub-cube, the density of the approximate copula is estimated from the number of points $\mathbf{u}_i$ in the subcube, divided by n and the volume of the sub-cube.

B. Generating random vectors.

In principle, using the approximate copula, random points $\hat{\mathbf{u}} \in [0,1]^D$ can be generated. From this, random vectors $\hat{\mathbf{z}}$ are obtained by means of the pseudo-inverses $F_d^{-1}(u) = \min\{z,\, F_d(z) = u\}$ of the estimated marginal distribution functions. Later on we indicate problems with this which occur if the marginal distribution functions are not continuous, and how to solve this problem.

Now we describe precisely the algorithm for the approximate distribution $\mathcal{A}$. Input data are a sample of random vectors $\mathbf{z}_i = (z_1, ..., z_D) \in \mathcal{Z}$, $i = 1, ..., n$, which are drawn from the unknown multivariate distribution $\mathcal{D}$ of the random vector $\mathbf{Z} = (Z_1, ..., Z_D)$.

Example 1. In Simulation, stochastic processes are more interesting. They can be analyzed and realized using a sliding window over a stochastic process. A special example is: $Z_{1,i} = A_i,\ Z_{2,i} = A_{i+1},\ i = 1, ..., n$, where $A_i,\ i = ..., -1, 0, 1, 2, ...,$ is the stationary stochastic process defined by $A_{i+1} = 0.5\Big(1 - 4(A_i - 0.5)^2\Big) + 0.5X_i$, where the X_i are independent and uniformly distributed over $[0, 1]$.

Example 2, observed data at an Internet server. The data consists of inter-arrival times (A_i) and packet lengths (B_i). In this case, the stochastic process is $Z_{1,i} = A_i,\ Z_{2,i} = B_i,\ Z_{3,i} = A_{i+1},\ Z_{4,i} = B_{i+1},\ i = 1, ..., n$.

1. The empirical marginal distribution functions $F_d(z)$ are estimated. To this end, the sequences $z_{d,1}, z_{d,2}, ..., z_{d,n},\ d = 1, ..., D$, are ordered: $z_{d,(1)}, z_{d,(2)}, ...,$ where $i < j$ implies $z_{d,(i)} \le z_{d,(j)}$. For $z_{d,(i)}$, where $z_{d,(i-1)} < z_{d,(i)} = z_{d,(i+1)} = ... = z_{d,(i+m-1)} < z_{d,(i+m)}$, $F_d(z_{d,(i)}) = m/n$ holds, $F_{d,i}$ for short (here we define $z_{d,(0)} = 0$ and $z_{d,(i,n+1)} = \infty$; these values are not used for estimation). For $z \in (z_{d,(i)}, z_{d,(i+1)})$, we define $F_d(z) = F_{d,i}$, but we will not use this. Alternatively, $F_d(z), d = 1, ..., D$, are fitted standard distributions.

2. Using the empirical or the standard marginal distributions, we get the transformed points $\mathbf{u}_i = (u_{1,i}, ..., u_{D,i}) \in [0,1]^D$, where $u_{d,i} = F_{d,i},\ d = 1, ..., D,\ i = 1, ..., n$.

3. For the sub-cubes, in each dimension d, the set $[0, 1]$ is partitioned into K_d subsets $S_{d,j}$ as follows: $S_{d,j} = [(j-1)\delta_d,\ j\delta_d),\ j = 1, ..., K_d - 1,\ S_{d,K_d} = [(K_d - 1)\delta_d,\ K_d\delta_d]$, where $\delta_d = 1/K_d,\ d = 1, ..., D$. With these subsets, the sub-cubes of $[0,1]^D$ are $\mathcal{S}_{\mathbf{j}} = S_{1,j_1} \times ... \times S_{D,j_D},\ \mathbf{j} \in \mathcal{K} = \{1, ..., K_1\} \times ... \times \{D, ..., K_D\}$.

Example. $D = 2$ dimensions, $K_1 = 3,\ K_2 = 4,\ \delta_1 = 1/3,\ \delta_2 = 1/4$

In a two dimensional cube, a partition $\mathcal{S}_{\mathbf{j}},\ \mathbf{j} \in \mathcal{K}$, induces a partition $\mathcal{T}_{\mathbf{j}} = T_{1,j_1} \times ... \times T_{D,j_D},\ \mathbf{j} \in \mathcal{K}$, in the original space $\mathcal{Z}$ of the observed random vectors $\mathbf{z}_i$ by means of $\mathbf{u} \in \mathcal{S}_{\mathbf{j}} \Leftrightarrow \mathbf{z} \in \mathcal{T}_{\mathbf{j}} \Leftrightarrow \forall d = 1, ..., D : z_d \in T_{d,j_d}$ where $\mathbf{u} = (u_1, ..., u_D)$, $\mathbf{z} = (z_1, ..., z_D)$, and $z_d = F_d^{-1}(u_d)$. See figure 1. This induced partition is unique only if the marginal distribution functions $F_d(z)$ are strictly increasing.

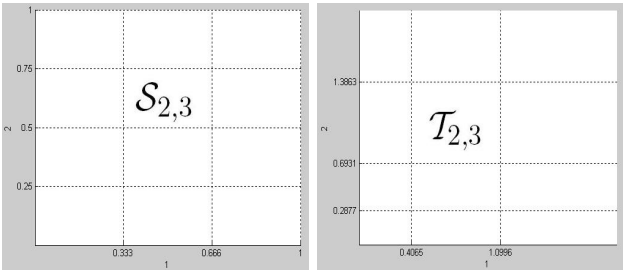

Figure 1: LEFT: SUB-CUBES OF THE UNIT D-CUBE. RIGHT: OF SPACE $\mathcal{Z}$, where $F_d(z) = 1 - \exp(-z)$

The approximate density of the copula is constant within each sub-cube $S_{\mathbf{j}},\ \mathbf{j} \in \mathcal{K}$. With the number $N_{\mathbf{j}}$ of points $u_{\mathbf{j}}$ in the sub-cubes $\mathcal{S}_{\mathbf{j}}$ and $H_{\mathbf{j}} = N_{\mathbf{j}}/n$, the density has the value $H_{\mathbf{j}}/(\delta_1 \cdot ... \cdot \delta_D)$. $H_{\mathbf{j}},\ \mathbf{j} \in \mathcal{K}$, is a frequency distribution for tuples $\mathbf{j}$. The reader may note that these approximations are not really a copula, in general: The marginal distributions are only approximately uniform. However, the empirical copulas, and thus their derivatives, the frequency copulas, converge to true copulas. See (Goorbergh and Genest and Werker 2005).

This (approximate) frequency copula, together with the pseudo-inverses of the marginal distributions, defines the approximate distribution $\mathcal{A}$. Its goodness of fit can be tested statistically with a chi-square test if a further sample of the same population is available. We present this in section 4.

Now we indicate how random vectors $\hat{\mathbf{z}}$ are generated form the fitted approximate distribution $\mathcal{A}$:

1. First a sub-cube $S_{\hat{\mathbf{j}}}$ is selected randomly with equal probability $1/n$ according to the distribution $H_{\mathbf{j}},\ \mathbf{j} \in \mathcal{K}$, in the usual way, see for example (Law and Kelton 2000), page 479.

2. If the marginal distribution functions $F_d(z),\ d = 1, ..., D$, are continuous, a random point $\hat{\mathbf{u}} = (\hat{u}_1, ..., \hat{u}_D)$ is generated with a uniform distribution over the selected sub-cube. With the pseudo-inverses of the marginal distribution functions, the elements of the random vector are $\hat{z}_d = F_d^{-1}(\hat{u}_d),\ d = 1, ..., D$.

If the marginal distributions are discrete, we proceed differently. This is also the case for our empirical distribution functions. From all points $\mathbf{u}_i$ in the sub-cube, one point is selected randomly , say $\hat{\mathbf{u}}$. If one point occurs $m > 1$ times in the sample, the same point will be present several times in the sub-cube, say m-fold. In

this case point probability is m/n.
For the transformation into the original space $\mathcal{Z}$ we use $F_d(z|Z_d \in T_{d,\hat{j}_d})$, the marginal distribution functions conditioned on Z_d lying in the interval $T_{d,\hat{j}_d}$ according to the selected sub-cube, for all dimensions d: $\hat{z}_d = F_d^{-1}(\hat{u}_d|Z_d \in T_{d,\hat{j}_d}),\ d = 1, ..., D$.
Why do we proceed differently for discrete distributions? Consider the following situation. Let $u' = F_d(z_{d,(i-1)}) < j\delta_d < F_d(z_{d,(i)}) = u'', \hat{u} = u'(1+\epsilon),\ \epsilon > 0$ so small that $\hat{u} < j\delta_d$. Hence, $u' \in S_{d,j},\ u'' \in S_{d,j+1}$ and $z_{d,(i-1)} \in T_{d,j},\ z_{d,(i)} \in T_{d,j+1}$. If in the generating process a sub-cube $... \times S_{d,j} \times ... \in [0,1]^D$ was selected and $\hat{u}$ was generated for $\hat{z}_d$, the $\hat{\mathbf{z}}$ which is generated via $\hat{z}_d = F_d^{-1}(\hat{u}) = z_{d,(i)}$ would lie in the sub-cube $\mathcal{T} \in \mathcal{Z}$ which corresponds to the sub-cube $... \times S_{d,j+1} \times ... \in [0,1]^D$, not $... \times S_{d,j} \times ... \in [0,1]^D$ which was determined by the algorithm. This problem could alternatively be omitted with a different definition of the sub-cubes.
The computational cost for the method is $O(n \log n + nK_1 \cdot ... \cdot K_D)$. In our examples, calculated with MATLAB on a 1GHz PC, the computing times were seconds or few minutes.

A CHI-SQUARE TEST FOR THE QUALITY OF THE APPROXIMATE DISTRIBUTION

In this section, we compare the approximate distribution $\mathcal{A}$ with a second sample $\mathbf{z}'_i,\ i = 1, ..., n'$, from the same population as the sample $\mathbf{z}_i,\ i = 1, ..., n$, which we used to build $\mathcal{A}$, by means of a chi-square goodness-of-fit test. If the hypothesis is not rejected, we take this as an indication of the quality of $\mathcal{A}$.
For the chi-square test the sample must consist in independent points, but this is not fulfilled in general. Therefore we start with a larger sample and discard points between $\mathbf{z}'_i$ and $\mathbf{z}'_{i+1}$ and hope that spaced points are nearly independent.
For the test, the space $\mathcal{Z}$ must be partitioned. We use the partition $\mathcal{T}_\mathbf{j},\ \mathbf{j} \in \mathcal{K}$, but in each subset must be enough points of the first sample. This is in general not the case. Therefore we combine sub-cubes to a subset $\mathcal{R}_l$ until $n' \cdot p_l \geq 5$ where p_l is the sum of the probabilities $H_\mathbf{j}$ of all combined sub-cubes $\mathcal{T}_\mathbf{j}$; this is a usual heuristic. Thus we obtain r subsets $\mathcal{R}_l$ and probabilities $p_l,\ l = 1, ..., r$. The precise definition of this combination is provided by an algorithm in the appendix.
As we said before, the partition $\mathcal{T}_\mathbf{j},\ \mathbf{j} \in \mathcal{K}$, is not unique, in general. Here we define the intervals in each dimension as follows: $T_{d,j} = [l_{d,j},\ h_{d,j})$ for all d and all $j = 1, ..., K_d$, where $l_{d,1} = 0,\ l_{d,j} = F_d^{-1}(u)$ with $u = \min_{1 \leq i \leq n}\{u_{d,i} \in \mathcal{S}_{d,j}\},\ h_{d,j} = l_{d,j+1}, j = 1, ..., K_d - 1$, and $h_{d,K_d} = \infty$.
Let $N'_\mathbf{j}$ denote the number of points $\mathbf{z}'_\mathbf{j}$ in the sub-cube $\mathcal{T}_\mathbf{j}$ and y_l the number in the subset $\mathcal{R}_l,\ l = 1, ..., r$.
The test statistic is the χ^2-distance function $Q = \sum_{l=1}^{r} y_l^2/(n'p_l) - n'$ which is compared with the $(1-\alpha)$-quantil of the χ^2-distribution with $r-1$ degrees of freedom.

EXAMPLES

In the numerical examples, correctness and accuracy are verified with the chi-square test, with some statistics and, visually, with scatter diagrams.
Statistics and diagrams are calculated for the measured sample and time series which are generated with the approximate distribution $\mathcal{A}$. The statistics are means, coefficients of variation, and correlations of the $z_{d,i}$, the latter between $z_{d,i}$ and $z_{d',i},\ d \neq d'$. First we calculated the differences of corresponding coefficients of variation and correlations, and relative differences of corresponding means. In order to not bother the reader with many figures, we only give the maximum of the absolute values of these differences, the *maximum statistics difference*.
The method primarily serves the purpose to analyze the multivariate distribution of random vectors and to generate random vectors for simulation input. In our examples, it is indicated how it can be used for time series.
Under these circumstances, when the next $\hat{\mathbf{z}}_i$ is generated, some elements from the previous $\hat{\mathbf{z}}_{i-1}$ can be taken, e.g. in example 2, $\hat{z}_{1,i} = \hat{z}_{2,i-1}$. If for the generation of $\hat{\mathbf{z}}_{i-1}$ the sub-cube was $\mathcal{T}_\mathbf{j}$, then for the next generation, the interval T_{1,k_1} of the next sub-cube $\mathcal{T}_\mathbf{k}$ equals T_{2,j_2} from the previous sub-cube.
Here, two kinds of errors may occur. First, if the following holds: In the sample $\mathbf{z}_i,\ i = 1, ...n$, from which the distribution $\mathcal{A}$ was built, there is no point $\mathbf{z}_i$ which lies in any sub-cube $\mathcal{T}_\mathbf{k}$ with $T_{1,k_1} = T_{2,j_2}$. That means the row $H(j_2, .)$ has only zero entries. Hence, in the selected sub-cube, no point can be generated, the generation leads into a *dead end*.
Secondly, the probabilities $H_\mathbf{j}$ can be so that the generated points end in a cycle of points which recur again and again.
These errors occur with some probability if the sample is small, in bigger samples this probability becomes smaller and smaller. In fact, we observed these problems only when very small samples were used for the distribution $\mathcal{A}$, ten points or so.
If a bigger sample is impossible or unwanted, there is a remedy: When a dead end or a cycle occurs. $\hat{\mathbf{z}}_i$ is generated completely new under violation of $\hat{z}_{1,i} = \hat{z}_{2,i-1}$.

Example 1

We consider a sliding window over a stochastic process. $Z_{1,i} = A_i,\ Z_{2,i} = A_{i+1},\ i = 1, ..., n$, where $A_i,\ i = ..., -1, 0, 1, 2, ...$, is the stationary stochastic process defined by $A_{i+1} = 0.5\Big(1 - 4(A_i - 0.5)^2\Big) + 0.5X_i$, where the X_i are independent and uniformly distributed over $[0, 1]$.
Sample size $n = 4000$. The number of subintervals

in each of the two dimensions $K_1 = K_2 = 40$. The maximum statistics difference is 0.018, no dead end occurred. The scatter diagrams of the sample and the generated points indicate that there are probably regions where no points can exist, and that these regions are observed by the generated process with good accuracy, as seen in figure 2. For comparison, we generated

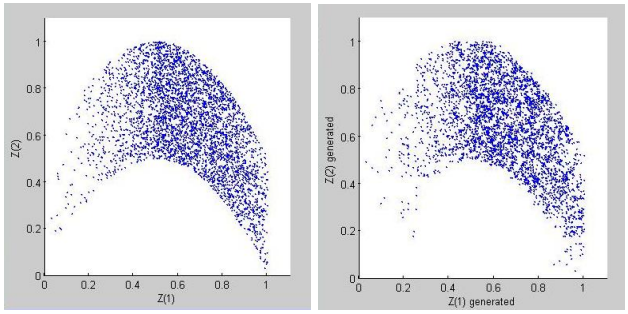

Figure 2: LEFT: THE ORIGINAL SAMPLE. RIGHT: THE GENERATED PROCESS

a process with a fitted linear nGAR model. Obviously, here many generated points lie in impossible regions as seen in figure 3. For four different streams of random

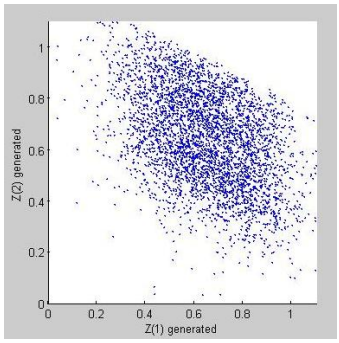

Figure 3: PROCESS GENERATED WITH AN AR MODEL

numbers, we built the approximate distribution $\mathcal{A}$ with small samples, $n = 100$, and larger samples, $n = 400$, and $K_1 = K_2 = 20$. For the small sample sizes, the chi-square test indicated three of four times "reject", for the larger sample sizes not once.

Example 2

We consider observed data by (Klemm, Lindemann and Lohmann 2002) at an Internet server. The data consists of inter-arrival times (A_i) and packet lengths (B_i).
In this case, the stochastic process is $Z_{1,i} = A_i,\ Z_{2,i} = B_i,\ Z_{3,i} = A_{i+1},\ Z_{4,i} = B_{i+1},\ i = 1, ..., n$.
Sample size $n = 4000$, $K_1 = K_2 = K_3 = K_4 = 40$. The maximum statistics difference is 0.03, no dead end occurred. The scatter diagrams of the sample and the generated points indicate good accuracy. See figures 4 and 5. Figure 6 is a scatter diagram of the points

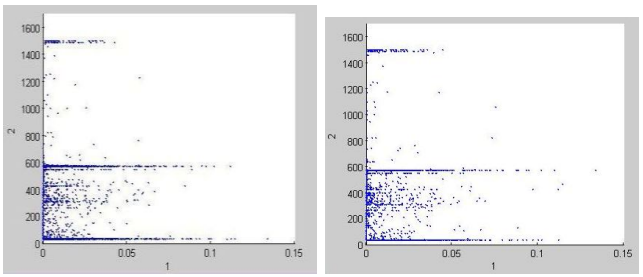

Figure 4: LEFT: THE ORIGINAL SAMPLE. RIGHT: THE GENERATED PROCESS; DIMENSIONS 1 & 2

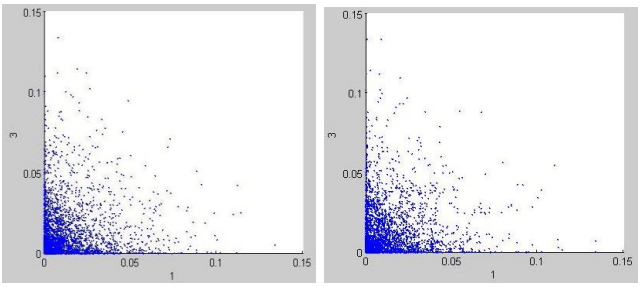

Figure 5: LEFT: THE ORIGINAL SAMPLE. RIGHT: THE GENERATED PROCESS; DIMENSIONS 1 & 3

$(u_{1,i}, u_{2,i})$, hence some visualization of the marginal density of the copula, dimensions 1 and 2. Figure 7 visual-

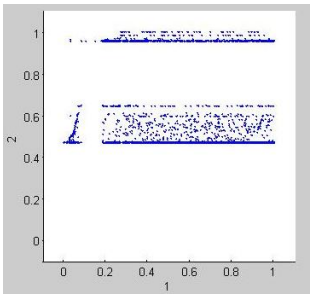

Figure 6: TRANSFORMED SAMPLE POINTS, $(u_{1,i}, u_{2,i})$

ize the dimension 1 and 2 of the sub-cubes $\mathcal{S}_{\mathbf{j}}$ for $K_1 = K_2 = K_3 = K_4 = 40$ and $K_1 = K_2 = K_3 = K_4 = 60$. Every point indicates one or more sub-cubes with sample points. Obviously, the higher accuracy separates better the impossible regions.

CONCLUSION

Copulas seem to be useful for the analysis of multivariate samples and for the generation of multivariate random numbers and time series. In contrast to other approaches, this approach is able to approximate any type

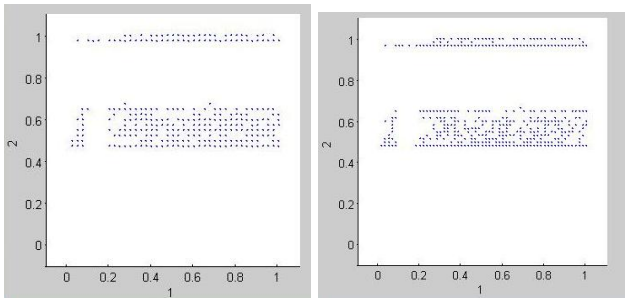

Figure 7: DIMENSIONS 1 AND 2 OF THE SUB-CUBES, DIFFERENT APPROXIMATION ACCURACY

of models (linear or nonlinear regression, multidimensional time-series, ...).
For future work, more goodness-of-fit tests can be done, and variations of the proposed method should be considered, for example:

- More flexible sub-cubes.
- Other marginal distribution functions, e.g. fitted standard distributions.
- Other kinds of estimated copulas.

Acknowledgement We gratefully appreciate the recommendation of our colleague Dr. H.J. Kühn to consider copulas.

REFERENCES

Biller B. and B. L. Nelson, 2003, "Modeling and generating multivariate time-series input processes using a vector autoregressive technique," *ACM Transactions on Modeling and Computer Simulation*, vol. 13, no. 3, pp. 211–237.

P. Blum and A. Dias and P. Embrechts, 2002, "The art of dependence modelling: the latest advances in correlation analysis," *Alternative Risk Strategies*, London. 339-356.

Cario M. C. and B. L. Nelson, 1996, "Autoregressive to anything—time-series input processes for simulation," *Operations Research Letters*, vol. 19, no. 2, pp. 51–58.

Cario M. C. and B. L. Nelson, 1997, "Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix", *Department of Industrial Engineering and Management Sciences*, Evanston, Ill.

F. Nassaj and J. Ch. Strelen, 2005, "Dependence input modeling with the help of non-Gaussian AR models and genetic algorithms," *Modelling and Simulation 2005, Proceedings of the European Simulation and Modelling Conference, Porto, 2005*, pp. 146-153.

Ghosh S. and S. G. Henderson, 2002b, "Properties of the norta method in higher dimensions," in *Winter Simulation Conference Proceedings*, Piscataway, N.J., pp. 263–269.

Klemm A., C. Lindemann, and M. Lohmann, 2002, "Traffic Modeling of IP Networks Using the Batch Markovian Arrival Process", *12th Int. Conf. on Modelling Tools and Techniques for Computer and Communication System Performance Evaluation*. London. UK. Lecture Notes in Computer Science, 2324, 92-110.

Law A. M. and D. W. Kelton, 2000, *Simulation Modeling and Analysis, 3rd edition*. New York: McGraw-Hill.

B. Melamed, 1997, "The empirical TES methodology: Modeling empirical time series," *J. of Applied Mathematics and Stochastic Analysis*, vol. 10, no. 4, pp. 333-353.

R.B. Nelsen, 1998, *An introduction to copulas*, New York: Springer.

D. Pfeifer and J. Neslehova, 2003, "Modeling dependence in finance and insurance: the copula approach," *Blätter der DGVFM*, vol. 26, no. 2, pp. 177-191.

Rob W.J. van den Goorbergh, Christian Genest, Bas J.M. Werker, 2005, "Bivariate option pricing using dynamic copula models," *Insurance: Mathematics and Economics*, vol. 37, no. 1, pp. 101-114.

APPENDIX

Algorithm for the subsets of the chi-square test
$l := 0$;
$sum := 0$;
$sum' := 0$;
for all $\mathbf{j} \in \mathcal{K}$
  sum **plus** $n'H_{\mathbf{j}}$;
  sum' **plus** $N'_{\mathbf{j}}$;
  if $sum \geq 5$
    l **plus** 1; // next subset $\mathcal{R}_l$
    $n'p_l := sum$;
    $y_l := sum'$;
    $sum := 0$;
    $sum' := 0$;
  fi
end
$n'p_1$ **plus** sum;
y_1 **plus** sum';
$r := l$; //number of subsets