DEPENDENCE INPUT MODELING WITH THE HELP OF NON-GAUSSIAN AR MODELS AND GENETIC ALGORITHMS

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

ABSTRACT

Input modeling software tries to fit standard probability distributions to data assuming that the data are independent. However, the input environment can generate correlated data. Ignoring the correlations might lead to serious inaccuracy in the performance measures. In the past few years, several dependence modeling packages with different properties have been developed. In this paper, we explain how to fit non-Gaussian autoregressive models to correlated data and compare our approach with similar dependence modeling approaches that already exist.

INTRODUCTION

Data measured on many real-life systems might exhibit correlations (dependencies) among themselves. Ignoring the correlations might lead to serious inaccuracy in the performance measures. For example, important statistical properties of the Internet traffic are burstiness and self-similarity (Klemm, Lindemann and Lohmann 2002). To illustrate this, consider for example packets arriving at an Internet server. If the average number of packets in a single burst increases, while spacing the bursts farther, the arrival rate of packets can be kept constant. On the other hand, the waiting times for the packets will increase considerably. Not taking the burstiness into account results in predicting optimistic performance measures. The main reasons behind burstiness and self-similarity are the correlations and the heavy-tails present in the interarrival process.

Before we proceed in our introduction, we will define some terms. The *Correlation* of two random variables X and Y, denoted as $\rho(X,Y)$, measures how much one random variable depends linearly on the other. The correlation is defined as the *covariance* standardized to the range [-1,1]:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y},$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively. The covariance of two random variables X and Y is defined as:

$$cov(X,Y) = E[X - E(X)][Y - E(Y)],$$

where E is the expectation. A series of random variables at successive times is called a *time-series.* The *Autocorrelation* is the correlation between two random variables of a time-series. The autocorrelations can be modeled using the autoregressive (AR) models of (Box and Jenkins 1976) (see below.) The autocorrelation between the two random variables in a time-series X_t which are lag h apart is denoted as $\rho_X(h)$. The autocorrelations in a stationary time-series depend only on the lag h, not on the time when the random variables are generated.

A stochastic process is said to be strongly stationary if all random variables of the process have the same distribution. On the other hand, a weakly stationary stochastic process is a process, in which the first and second moments (the mean and the variance) exist and do not change over time. In non-Gaussian processes (see below), neither strong stationarity follows from weak stationarity, nor weak stationarity follows from strong stationarity (Chatfield 1996). To be able to fit an AR model to a time-series, the time-series must be either strongly or weakly stationary. We refer to a strongly and/or weakly stationary time-series simply as a stationary time-series.

ARTA, NORTA, and VARTA processes, described below, have some similarities to our approach, but still have different advantages and disadvantages. The abbreviations stands for autoregressive to anything, normal to anything, and vector autoregressive to anything, respectively. From now on, we will call these processes as ARTA-like processes. The approaches of ARTA and VARTA try to model the dependencies in a time-series by transforming a Gaussian AR process to a non-Gaussian process. The later processes has similar statistical properties as the time-series. The NORTA approach in turn depends on transforming Gaussian random variables to "any" non-Gaussian random variables. The later random variables have some desired statistical properties (distribution and correlation. Unlike the approach of the ARTA-like processes, our approach depends on fitting non-Gaussian AR models (see below) to dependent time-series, by first transforming it to (nearly) independent data. After the transformation, a probability distribution is fitted to the data using any input modeling approach.

In comparison with the other dependence modeling approaches, Our approach gives the modeler more flexibilities. For example, it enables the modeler to fit heavy and power tailed processes and non-linearly correlated processes to time-series. Moreover, our approach can be easily integrated with already existing input modeling software.

GAUSSIAN AND NON-GAUSSIAN AR PROCESSES

The autocorrelations and underlying distribution of a time-series can be modeled with the help of AR models. A parameterization of a univariate linear Gaussian AR model of order p is:

$$Z_t = \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + ... + \alpha_p Z_{t-p} + Y_t, \qquad (1)$$

where p is the longest lag, and the Y_t are IID normal (Gaussian) random variables with mean zero and variance ${\sigma_Y}^2$ carefully chosen such that the Z_t are standard Gaussian (Law and Kelton 2000). The AR coefficients α_h, $h = 1, 2, ..., p$, uniquely determine the autocorrelations of the Z_t, $\rho_Z(h)$. The α_h are chosen such that the AR process is stationary.

Our parameterization of an AR model differs from that given in (1). We allow the Y_t to be drawn from *any* distribution. The Y_t can be even drawn from power-tailed distributions like the Pareto distribution. We call such models *non-Gaussian AR (nGAR) models.* nGAR models apply the same stationarity condition as the Gaussian AR models.

The nGAR models have the property that the distribution of the Z_t differs from that of the Y_t. This means that the input modeler, who does not have a sample of a time-series, can not generate an nGAR process which has a specific distribution and autocorrelation structure. However, if a modeler has a sample of a time-series, an nGAR process can be fitted. The fitted process will have similar statistical properties as the original time-series. In reality, it is common to have only a sample of a time-series which should be fitted to a distribution or to a stochastic process.

ARTA, NORTA, AND VARTA PROCESSES

ARTA processes use as a base process a standard Gaussian AR process described above. It then uses Z_t to generate a series of autocorrelated uniform random variables, U_t, by using the probability-integral transformation, $U_t = \Phi(Z_t)$, where Φ is the standard normal distribution. ARTA applies then the inverse transformation method, $X_t = F_X^{-1}[U_t]$, to generate random variables having a specific distribution, F_X. Please note that the Gaussian property of the Y_t in (1) ensures not only that the Z_t are standard Gaussian, but also that the autocorrelation coefficients of the base process, $\rho_Z(h)$, determined by the AR coefficients α_h, $h = 1, 2, ..., p$, uniquely determine the autocorrelation coefficients of the X_t, $\rho_X(h)$.

ARTA processes of (Cario and Nelson 1996) are able to generate random variables having a specific distribution and autocorrelation structure, which should in turn be given explicitly. A complementary work to that is the work described by (Biller and Nelson 2002). They describe how to fit ARTA processes to univariate time-series. This will enable the user to provide a time-series and to get as a result a fitted ARTA process that has similar statistical properties as the original time-series.

Another research in this area are the NORTA processes of (Cario and Nelson 1997). NORTA processes can be used to generate IID finite vectors of random variables. The random variables within the vectors can have arbitrary marginal distributions and correlation matrix. The idea behind this work is to transform a standard multivariate normal vector $\mathbf{Z} = (Z_1, Z_2, ..., Z_d)'$ into a vector $\mathbf{X} = (X_1, X_2, ..., X_d)'$, where $X_i = F_i^{-1}[\Phi(Z_i)]$. F_i, $i = 1, 2, ..., d$, may be different distributions. Moreover, X_i, $i = 1, 2, ..., d$, can exhibit correlations among themselves.

A generalization of ARTA and NORTA processes are the VARTA processes of (Biller and Nelson 2003). VARTA can be fitted to multivariate time-series by considering the AR base process as the standard Gaussian vector AR process of order p. Similar to the case of ARTA, the autocorrelation structure of the base process, determined by the AR coefficients, specifies uniquely the target autocorrelation structure of the resulted VARTA process.

ARTA-like processes depend on a transformation of a base process into a specific process. Let us consider for example the ARTA processes. The autocorrelations in the base process of ARTA, $\rho_Z(h)$, do not match the autocorrelations of the ARTA process, $\rho_X(h)$. However, (Cario and Nelson 1996) have shown that $\rho_X(h)$ is a continuous non-decreasing function of $\rho_Z(h)$:

$$\frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_X^{-1}[\Phi(z_t)]F_X^{-1}[\Phi(z_{t-h})]\vartheta_{\rho_Z(h)}(z_t, z_{t-h})dz_t dz_{t-h} - \mu^2}{\sigma^2},$$

where h is the current base process autocorrelation lag, and ϑ is the bivariate normal pdf.

NORTA and VARTA depend also on transformations similar to the above one. Current researches show that this kind of transformations has a drawback, that is, there are some random vectors, X_t, with feasible covariance matrices, Cov_X, which are transformed to non-feasible base process covariance matrices, Cov_Z. In other words, for some desired Cov_X matrices, the transformation results in non-positive definite Cov_Z matrices. Non-positive definite covariance matrices are invalid covariance matrices (Fishman 1978). These Cov_X matrices, that are transformed into non-positive definite Cov_Z matrices, are called *defective* matrices.

This drawback is discussed by (Ghosh and Henderson 2001), (Ghosh and Henderson 2002a), and (Ghosh and Henderson 2002b) in detail for the NORTA processes. The papers provide an example of a defective covariance matrix Cov_X. They suggest a modified NORTA process that can detect such defective matrices, and generate Cov_Z matrices that are positive definite and "close" to the desired ones. VARTA processes, which are generalizations of the NORTA processes, are supposed to have the same drawback. ARTA is not yet proved to suffer from the defective matrices problem, as the defective matrix given by (Ghosh and Henderson 2002a) and (Ghosh and Henderson 2002b) is not a valid ARTA covariance matrix. However, (Biller and Ghosh 2004) suggest that ARTA can also generate defective matrices, but they do not provide detailed information. Our method does not apply the kind of transformations mentioned above, and thus can not generate defective matrices.

Another drawback of the above transformation shows up when trying to fit ARTA-like processes to time-series. Let us consider ARTA for example. Fitting an ARTA process to a time-series, which have the distribution F_X with the parameters $\mathbf{p}$ and the autocorrelation $\rho_X(h)$, requires estimating F_X along with $\mathbf{p}$ and $\rho_Z(h)$ *in parallel*. In other words, the fitting procedure assumes a distribution F_X having the parameters $\mathbf{p}$, and try to estimate $\rho_Z(h)$ using an optimization procedure. Having $\rho_Z(h)$ estimated for specified F_X and $\mathbf{p}$, $\mathbf{p}$ and maybe F_X must be estimated using an optimization procedure. The procedure iterates until "convergence". This results generally in a relatively time consuming fitting procedure. Our approach does not perform such kind of parallel fitting. It handles the correlations and distributions separately.

The procedures of (Biller and Nelson 2002) and (Biller and Nelson 2003) fit ARTA and VARTA processes to time-series. The distributions considered in these two papers are only those from the Johnson translation system (Johnson 1987). This means that the current ARTA and VARTA approaches can not generate heavy-tailed ARTA and VARTA processes. Moreover, ARTA-like processes can not fit non-linear AR models to time-series. An example of non-linear AR models is

$$Z_t = \alpha_1 {Z_{t-1}}^p + \alpha_2 {Z_{t-2}}^{p-2} + ... + \alpha_p Z_{t-p} + Y_t. \quad (2)$$

In our paper, we explain how heavy-tailed and non-linear nGAR processes can be fitted to time-series.

THE GENETIC ALGORITHM

The genetic algorithms of (Chipperfield et al. 1994) and the programs of (Strelen 2003) are applied to help fitting distributions to IID samples, and to estimating their parameters. We also use the genetic algorithm for the purpose of optimizing some objective function in our independence method.

The genetic algorithm is a stochastic global search method that mimics the natural biological evolution. It operates on populations of individuals applying the principle of the survival of the fittest. The genetic algorithm uses operators borrowed from the natural genetics like the recombination, selection, and mutation.

The first step of a genetic algorithm procedure is to initialize a population randomly from a pre-specified range. The population consists of individuals who are assigned fitnesses according to an objective function. Fitter individuals have higher probability to propagate to the next generation and higher probability to be selected to produce the individuals of the next generation.

The individuals of a population can represent the parameters of a distribution. The objective function might then depend on the *mean absolute distance* principle and established as follows: Having samples $\mathbf{y} = (y_1, y_2, ..., y_n)$, one sorts $\mathbf{y}$ to get the order statistics $\mathbf{y}^r = (y^{(1)}, y^{(2)}, ..., y^{(n)})$. The piecewise-constant empirical distribution function is then built as $F_Y = r/n$, and the objective function for the genetic algorithm is

$$Z(\hat{\mathbf{p}}) = (\sum_{r=1}^{n} \Big| F_Y - F_{\hat{Y}}(\hat{\mathbf{p}}, y^{(r)}) \Big|)/n, \quad (3)$$

where $F_{\hat{Y}}(\hat{\mathbf{p}}, y^{(r)})$ is the value of the selected distribution function with the parameter(s) $\hat{\mathbf{p}}$ at the point $y^{(r)}$. The objective function $Z(\hat{\mathbf{p}})$ measures how accurately $F_{\hat{Y}}(\hat{\mathbf{p}})$ fits $\mathbf{y}$.

THE MODEL AND THE FITTING PROCEDURE

Our goal is to approximate a stationary (multivariate) time-series $\mathbf{X}_t$, $t = 1, 2, .., n$, by a (multivariate) nGAR process, $\mathbf{Z}_t$, specified by

$$\mathbf{Z}_t = \psi(\mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, ..., \mathbf{Z}_{t-p}, \mathbf{\Sigma}_\alpha) + \mathbf{Y}_t,$$

where $\mathbf{Z}_t = (Z_{1,t}, Z_{2,t}, ..., Z_{d,t})'$ is a d-vector of random variables observed at time t. $\mathbf{Y}_t = (Y_{1,t}, Y_{2,t}, ..., Y_{d,t})'$ is a d-vector of independent random variables. For each $i = 1, 2, ..., d$, the $Y_{i,t}$, $t = 1, 2, ..., n$ are independent and have the probability distribution F_{Y_i}. $\mathbf{\Sigma}_\alpha$ is a set of numerical parameters, e.g. a $d \times p$ coefficient matrix, and p is the longest lag. $\mathbf{\Sigma}_\alpha$, p, and ψ are assumed such that $\mathbf{Z}_t$ is stationary for any $t = 1, 2, ..n$. The function ψ, the parameters $\mathbf{\Sigma}_\alpha$, and the longest lag p, determine the autocorrelation structure of the $\mathbf{Z}_t$. We assume that the nGAR process $\mathbf{Z}_t$ is stationary. Therefore, we ignore the first elements generated by the model.

In the case of linear nGAR models, ψ is simply a matrix-multiplication:

$$\mathbf{Z}_t = [\mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, ..., \mathbf{Z}_{t-p}] \times {\mathbf{\Sigma}_\alpha}'. \quad (4)$$

In the case of non-linear nGAR models, ψ is more general. An example of non-linear nGAR models is shown in (2).

Fitting an nGAR model to a time-series corresponds to estimating the parameters $\Omega = \{\psi, \mathbf{\Sigma}_\alpha,$ p, $F_{Y_i}, i = 1, 2, ..., d\}$. Here, the function ψ is chosen out of a finite set of given functions. In real world problems, the value of d is usually 1 or 2 and the value of p is ≤ 5.

As a first main step of our approach, the numerical parameters $\mathbf{\Sigma}_\alpha$ are estimated for one or several ψ functions, and one or several p values. When ψ is linear in the $\mathbf{Z}_t$ as in (1), this step can be accomplished by means of the Yule-Walker or Burg method (parametric methods) or by the independence method explained below. If non-linear functions ψ are considered, only the independence method can be used, as the parametric methods work only when the correlations are linear in the random variables $\mathbf{Z}_t$.

In the second main step of our approach, the $\hat{\mathbf{Y}}_t$ are estimated for each (ψ, p) pair and their corresponding estimated $\hat{\mathbf{\Sigma}}_\alpha$:

$$\hat{\mathbf{Y}}_t = \mathbf{Z}_t - \psi(\mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, ..., \mathbf{Z}_{t-p}, \hat{\mathbf{\Sigma}}_\alpha), \qquad (5)$$

and then the independence of the $\hat{\mathbf{Y}}_t$ is postulated. The pair (ψ, p) which results in independent $\hat{\mathbf{Y}}_t$ is considered to be the "optimal" one. However, more than one pair can be considered. Next, for each $i = 1, 2, ..., d$, a distribution is fitted to the $Y_{i,t}$, $t = 1, 2, ..., n$, which resulted from the best (ψ, p) pari(s). This can be accomplished with a tool like ExpertFit, Arena input analyzer, or with the technique described in (Strelen 2003). At the end of step 2, estimated $\hat{\mathbf{\Sigma}}_\alpha$ and $\hat{F}_i$ will be available for each (ψ, p) pair. This means that different estimated nGAR model parameters $\hat{\Omega}$ will be available. In the third main step of the procedure, the best set of model parameters is chosen according to one or several statistical tests. More detail about the procedure is given in the following two subsections.

Fitting Linear nGAR Processes

Let us consider fitting a linear univariate nGAR Model similar to that given in (1), where the Y_t are drawn from a specific probability distribution. As mentioned, a sample of a time-series should be available. Moreover, an nGAR model order, p, should be assumed. The known methods for estimating the AR orders like the *Akaike* or *Schwarz information criterion* do not work in this case, as the provided samples are usually not normally distributed. Instead, an order, $\hat{p}_{test}$, is chosen, and a parametric method is used to estimate the AR coefficients $\hat{\alpha}_h$, $h = 1, 2, .., \hat{p}_{test}$.

If $\hat{p}_{test}$ is higher than the actual order, p, the estimated $\hat{\alpha}_h$ will contain "small" AR coefficients for lags higher than p. In the case that the sample is highly correlated and $\hat{p}_{test}$ is smaller than p, $\hat{\alpha}_{\hat{p}_{test}}$ will not be small. The term small AR coefficient in linearly correlated models like those given in (1) might mean any value smaller than 5%.

In general, the actual order, p is unknown. In this case, one can test whether $\hat{p}_{test}$ is large enough by applying an independence test on the $\hat{Y}_t$ estimated by (6). The dependent sample of the time-series is transformed to (nearly) independent one using:

$$\hat{Y}_t = Z_t - \alpha_1 Z_{t-1} - \alpha_2 Z_{t-2} - ... - \alpha_p Z_{t-\hat{p}_{test}}. \qquad (6)$$

Having the independent $\hat{Y}_t$, a distribution can be fitted. At this point, the parameters $\hat{\Omega}$ of the nGAR model (1) are estimated, an nGAR process can be generated, and statistical goodness-of-fit tests can be applied to compare the original time-series with the nGAR process.

The Independence Method

Another way to fit nGAR processes to time-series can be accomplished with the help of the Chi-square independence test. This procedure is used if the above described procedure is not applicable due to non-linear correlations in the time-series. In our independence method, the set of nGAR model parameters $\hat{\Omega} = \{\hat{\psi}, \hat{\mathbf{\Sigma}}_\alpha, \hat{p}, \hat{F}_i\}$ are estimated by first estimating the parameters $\hat{\mathbf{\Sigma}}_\alpha$ for each (ψ, p) pair.

The estimation of $\hat{\mathbf{\Sigma}}_\alpha$ is accomplished for one pair (ψ, p) by building the random variables $\hat{Y}_t$ using (5). Next, the independence of the Y_t is tested. If the Y_t are dependent, the parameters $\hat{\mathbf{\Sigma}}_\alpha$ must be adjusted. This results in an optimization procedure according to an objective function of independence.

Two vectors of realizations $\mathbf{X}$ and $\mathbf{Y}$ can be tested for independence using the chi-square test. The test requires in addition to the vectors a degree of freedom ν and a significance level (of rejection) α. If the test statistics calculated exceeds a value specified in the chi-square table under the selected ν and α, the vectors are said to be dependent.

The Test statistics are calculated as follows: Having $\mathbf{X}$ and $\mathbf{Y}$ vectors, the corresponding pairs (x, y)'s are sorted in different regions (u, v), $u = 1, 2, ..., k$, $v = 1, 2, ..., l$. The number of (x, y) pairs in each region (u, v) is then denoted as N_{uv}. $N_{u\bullet}$ denotes the number of pairs in the regions (u, v) for all $v = 1, 2, ..., l$. $N_{\bullet v}$ denotes the number of pairs in the regions (u, v) for all $u = 1, 2, ..., k$. The Chi-square test (7) is then applied to get a positive test value, Q. The smaller the test value is, the more independent $\mathbf{X}$ and $\mathbf{Y}$ are:

$$Q = n[(\sum_{u=1}^{k} \sum_{v=1}^{l} \frac{{N_{uv}}^2}{N_{u\bullet} N_{\bullet v}}) - 1] \qquad (7)$$

In the case of univariate time-series Z_t, the independence of the lag-h apart random variables $\hat{Y}_t$ must be postulated. This means that the vectors of random variables ($\hat{Y}_t$, $\hat{Y}_{t-h}$), $h = 1, 2, .., p$ must be tested for independence. This requires applying (7) p times, one time for each h. The values Q_h, $h = 1, 2, ..., p$, are then averaged and considered as the objective function value for independence.

For the purpose of testing a d-variate $\hat{\mathbf{Y}}_t$ with a target lag p for independence, each process in the $\hat{\mathbf{Y}}_t$, and the different processes in the $\hat{\mathbf{Y}}_t$, must be tested for independence. This means that for all $t = 1, 2, ..., n$, the pairs $(\hat{Y}_{i,t}, \hat{Y}_{j,t+h})$ are tested for independence for each $h = 1, 2, ..., p$, and $i, j = 1, 2, ..., d$. Let $Q_{i,j,h}(\hat{\mathbf{\Sigma}}_\alpha)$ denote the test statistics corresponding to i, j, and h having specific values, then, the objective function of independence might be defined as

$$Q(\hat{\mathbf{\Sigma}}_\alpha) = \sum_{i,j=1}^{d} \sum_{h=1}^{p} Q_{i,j,s}(\hat{\mathbf{\Sigma}}_\alpha)/pd^2.$$

Different $\hat{\mathbf{\Sigma}}_\alpha$ values result in different values for $Q(\hat{\mathbf{\Sigma}}_\alpha)$. The "optimal" $\hat{\mathbf{\Sigma}}_\alpha$ is calculated as:

$$\hat{\mathbf{\Sigma}}_{\alpha,\text{best}} = \arg\min_{\hat{\mathbf{\Sigma}}_\alpha} \; Q(\hat{\mathbf{\Sigma}}_\alpha).$$

If more than one ψ function or more than one p value are considered, $\hat{\mathbf{\Sigma}}_{\alpha,\text{best}}$ and the distributions $\hat{F}_{Y_i}$, $i = 1, 2, ..., d$ are searched for each (ψ, p) pair. $\hat{Y}_i$ are then built and tested for independence. The best (ψ, p) pair(s) are considered and different sets of nGAR parameters $\hat{\Omega}$ will be available. The best set of parameters $\hat{\Omega}_{\text{best}}$ is then chosen with the help of goodness-of-fit tests. In other words, the original time-series is tested against the fitted nGAR processes that have the different parameters $\hat{\Omega}$. The set of parameters $\hat{\Omega}$ which delivers the best test statistics (TS) is chosen to be the "optimal" one. More than one test can also be considered.

EXAMPLES

In this section, we give examples that show results of our fitting procedure. The realizations used in the first three examples are generated artificially from nGAR processes with known parameters. We call these realizations as empirical time-series. Our goal is to find out how well the fitting procedure recovers the original parameters (Ω) of the true processes. The samples used in the four example are real measurements done on an Internet server by (Klemm, Lindemann and Lohmann 2002).

Fitting Linear Univariate nGAR Processes

We consider an nGAR process with the following parameters Ω: The function ψ is the matrix-multiplication shown in (4). $\mathbf{\Sigma}_\alpha$ is a 1×2 matrix $[\alpha_1, \alpha_2] = [0.4, 0.2]$. The random variables Y_t are Pareto distributed having p_1 (shape parameter) = 1.7 and p_2 (scale parameter) = 3.7. This specifies $F_{\hat{Y}}(\hat{\mathbf{p}})$.

An empirical time-series with 10000 realizations from the above described nGAR process is generated. The large sample size is considered due to the property of the Pareto distribution under the specified shape parameter p_1. Pareto distributions with a shape parameter $p_1 < 2$ have infinite variance, which results in that the generated random variables are disperse along wide range. Therefore, a relatively large sample size is needed to capture enough information about the true process.

nGAR processes are fitted to the empirical time-series as follows: A function ψ (linear multiplication) and an nGAR order $p = 3$ are considered. Next, $\hat{\mathbf{\Sigma}}_\alpha = [\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3]$ is estimated using the independence method. Noticing that $\hat{\alpha}_3$ is small and that the $\hat{Y}_t$ are (nearly) independent, we suggest that an order $\hat{p} = 2$ is suitable. Having $[\hat{\alpha}_1, \hat{\alpha}_2]$ and $\hat{p}$ estimated for ψ, $\hat{Y}_t$ is built using

$$\hat{Y}_t = Z_t - \hat{\alpha}_1 Z_{t-1} - \hat{\alpha}_1 Z_{t-2}. \qquad (8)$$

Next, the distribution, $F_{\hat{Y}}$, and its parameters, $\hat{\mathbf{p}}$, that fit $\hat{Y}_t$ best are estimated using the techniques of (Strelen 2003). The test statistics used to select the best fitted distribution $F_{\hat{Y}}$ and its parameters $\hat{\mathbf{p}}$ is the *mean absolute distance* shown in (3). We call this statistic as MAD($F_{\hat{Y}}$). The three distributions that fit $\hat{Y}_t$ best and their corresponding MAD($F_{\hat{Y}}$) statistics are shown in table 1.

Having specific sets of parameters $\hat{\Omega}$, nGAR processes can be generated and compared with the empirical time-series. MAD(process) test is similar to the MAD($F_{\hat{Y}}$) test. MAD(process) and the Kolmogorov-Smirnov (KS) tests compare realization from the fitted nGAR processes and from the empirical time series. KSS(process) is the Kolmogorov-Smirnov test statistics. Table 1 summarizes the results of the fitting procedure considering the three distributions $F_{\hat{Y}}$ that delivers the best (smalles) MAD($F_{\hat{Y}}$) values. [$\hat{\alpha}_1$,$\hat{\alpha}_2$] are the fitted AR coefficients. [$\hat{p_1}$, $\hat{p_2}$] are the best fitted parameters of the Pareto, Weibull, and Lognormal distributions. We notice that all test statistics tend to be smaller (better) in the case of choosing $F_{\hat{Y}}$ to be the Pareto distribution.

Table 1: Results summary of fitting univariate linear nGAR processes to empirical time-series

[$\hat{\alpha}_1$,$\hat{\alpha}_2$]	[0.401, 0.198]		
$F_{\hat{Y}}$	**Pareto**	**Weibull**	**Lognormal**
[$\hat{p_1}$, $\hat{p_2}$]	[1.69, 3.68]	[2.59, 6.99]	[1.75, 0.45]
MAD ($F_{\hat{Y}}$)	0.007	0.0618	0.037
MAD (process)	1.4	6.8	6.0
KSS (process)	0.047	0.334	0.206

The scatter plots show whether the fitted nGAR processes and the empirical time-series have similar patterns. Fig. 1, Fig. 2, and Fig. 3 show the scatter plots

(Z_t, Z_{t+1}), (Z_t, Z_{t+2}) from the empirical time-series and two fitted nGAR processes. We notice that realizations from the fitted nGAR process with $F_{\hat{Y}}$ of Pareto have similar patterns as the empirical time-series.

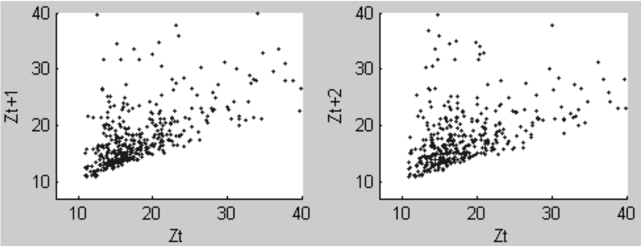

Figure 1: Plots from the empirical time-series

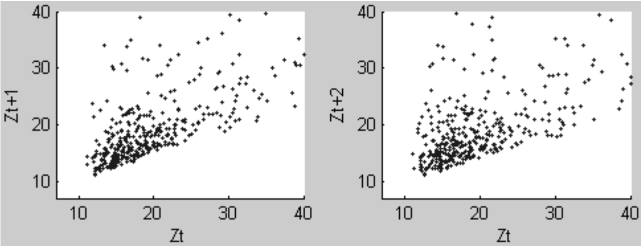

Figure 2: Plots from the fitted nGAR process with Pareto $F_{\hat{Y}}$

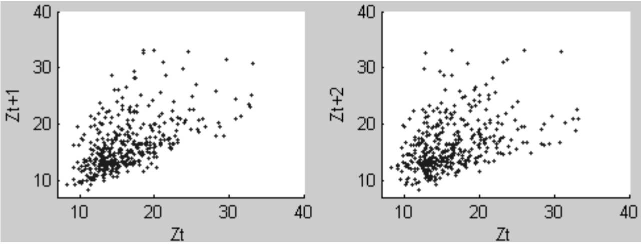

Figure 3: Plots from the fitted nGAR process with Lognormal $F_{\hat{Y}}$

Fitting a Linear Bivariate nGAR Process

We Generate a bivariate empirical time-series of size 1000 from the following linear bivariate nGAR process:

$$A_t = \alpha_1 A_{t-1} + \alpha_2 A_{t-2} + \alpha_3 B_{t-1} + \alpha_4 B_{t-2} + Y_{At}$$
$$B_t = \beta_1 B_{t-1} + \beta_2 B_{t-2} + \beta_3 A_{t-1} + \beta_4 A_{t-2} + Y_{Bt}$$

where $\underline{\alpha}$=[0.2,0.15,-0.15,0.10], $\underline{\beta}$=[0.2,-0.10,0.15,-0.1]. The Y_{A_t} are Weibull distributed with the parameters a_1 = 2 (shape), and b_1 = 10 (scale). The Y_{B_t} are Weibull distributed with a_2 = 1 and b_2 = 6.

A bivariate nGAR process is fitted as follows: A function ψ (linear multiplication) and nGAR order $p = 4$ are assumed. Next, $\hat{\mathbf{\Sigma}}_\alpha = [\hat{\underline{\alpha}};\ \hat{\underline{\beta}}]$ is estimated using the Yule-Walker method. Having $\hat{\mathbf{\Sigma}}_\alpha$ estimated for ψ and p, (nearly) independent bivariate time-series $(\hat{Y}_{A_t}, \hat{Y}_{B_t})'$ are built using (5). Next, the distributions and the parameters that fit $\hat{Y}_{A_t}$ and $\hat{Y}_{B_t}$ best are estimated. For the purpose of comparison, we fit the empirical time-series directly to theoretical distributions neglecting the fact that the empirical time-series is correlated. The best fitted distribution and parameters are shown in table 2. We notice from table 2 that an empirical time-series with relatively small size can recover the parameters of the true process with satisfying accuracy. This is due to the fact that Y_{A_t} and Y_{B_t} are not heavy-tailed. $\hat{\underline{\alpha}}$ and $\hat{\underline{\beta}}$ are the fitted AR coefficients ($\hat{\mathbf{\Sigma}}_\alpha$). The fitted distributions to the $\hat{Y}_{A_t}$ and $\hat{Y}_{B_t}$ are Weibull. They are marked with * in the table to distinguish then from the distributions fitted to the whole process. [$\hat{a}_1$, $\hat{b}_1$] and [$\hat{a}_2$, $\hat{b}_2$] are the fitted parameters. $MAD1$ and $MAD2$ are the mean absolute distance between the bivariate empirical time-series and the fitted bivariate processes, while (KSS1, KSS2) are the Kolmogorov-Smirnov statistics.

The test statistics MAD and KSS are only somewhat smaller in the case of fitting an nGAR process to the empirical time series. This show that the fitted theoretical distribution fit the *distribution* of the empirical time-series relatively well. However, the theoretical distributions can generate only independent data and the correlations of the time-series can not be modeled. This is notices by the correlation coefficients $\rho_A(1)$ and $\rho_A(2)$ and AR coefficients $\hat{\underline{\alpha}}$ and $\hat{\underline{\beta}}$.

Table 2: Results of fitting bivariate nGAR process and theoretical distributions to bivariate empirical time-series

	AR Process	**Theoretical Distr.**
$\hat{\underline{\alpha}}$	[0.19, 0.15, -0.13, 0.08]	[0, 0, 0, 0]
$\hat{\underline{\beta}}$	[0.19, -0.10, 0.13, -0.10]	[0, 0, 0, 0]
$F(\hat{a}_1,\ \hat{b}_1)$	$Weibull^*$ (2.02, 10.17)	LogN (2.5, 0.45)
$F(\hat{a}_2,\ \hat{b}_2)$	$Weibull^*$ (.99, 6.06)	Weibull (1.36, 7.43)
$\rho_A(1)$, $\rho_A(2)$	[0.22, 0.185]	[-0.010, 0.011]
(MAD1, MAD2)	(0.20, 0.18)	(0.84, 0.34)
(KSS1, KSS2)	(0.024, 0.019)	(0.040, 0.031)

Fig. 4, Fig. 5, and Fig. 6 show the plots (A_t, A_{t+1}) and (A_t, A_{t+2}) from the empirical time-series, the fitted AR processes, and the fitted independent process (distribution), respectively. We notice that the plots of the fitted nGAR process is more similar to the empirical time-series than those from the independent process.

Fitting a non-Linear nGAR Process

An empirical time-series of size 3000 is generated from a non-linear nGAR process:

$$Z_t = \alpha_1 {Z_{t-1}}^2 + \alpha_2 Z_{t-2} + Y_t \tag{9}$$

where $\underline{\alpha}$ =[0.034, 0.2]. Y_t are Weibull distributed with the parameters $[p_1, p_2]$ = [2.7, 4]. The linear correlations of the empirical time-series, $\rho_Z(1)$ and $\rho_Z(2)$ have the values 0.56 and 0.46, respectively.

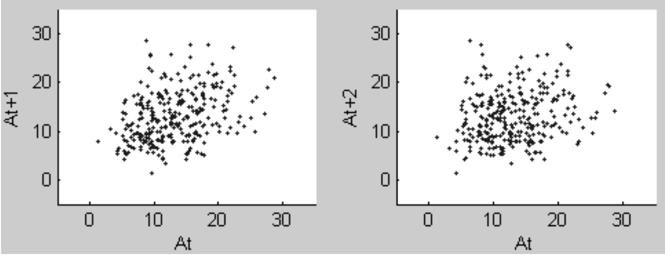

Figure 4: Plots from the empirical time-series

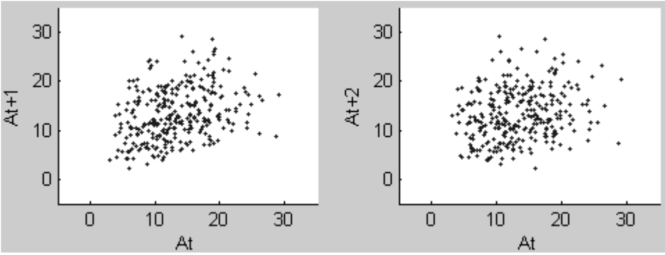

Figure 5: Plots from from the fitted nGAR process

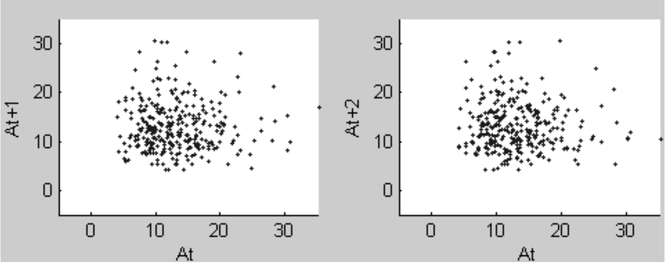

Figure 6: Plots from the fitted theoretical distribution

We consider fitting linear and non-linear nGAR processes. The assumed maximum lag $p = 2$. The assumed functions ψ are as follows:

$$\psi_1(Z_{t+1}, Zt+2) = \alpha_1 Z_{t+1} + \alpha_2 Z_{t+2},$$

and

$$\psi_2(Z_{t+1}, Zt+2) = \alpha_1 {Z_{t+1}}^2 + \alpha_2 Z_{t+2}.$$

In both cases of ψ, the parameters α_1 and α_2 are first estimated with the independence method. Next, the (nearly) independent $\hat{Y}_t$ are estimated and the distribution of the $\hat{Y}_t$, $F_{\hat{Y}}$, is determined. Table 3 summarizes some results of the fitting procedure.

Table 3: Results summary of fitting linear and non-linear nGAR process to empirical time-series

	Linear nGAR	**Non-linear nGAR**
[$\hat{\alpha}_1$,$\hat{\alpha}_2$]	[0.42 , 0.20]	[0.033, 0.202]
$F_{\hat{Y}}(\hat{p}_1, \hat{p}_2)$	Weibull(1.8 , 2.7)	*Weibull**(2.8, 3.9)
$\rho_A(1)$, $\rho_A(2)$	[0.59, 0.49]	[0.55, 0.45]
MAD(process)	0.51	0.2
KSS(process)	0.11	0.03

The entries of the table are similar to those described in the previous examples. We notice that the correlations of the linear and non-linear nGAR processes are similar to those of the empirical time-series. An example of these correlations are $\rho_A(1)$ and $\rho_A(2)$. However, the test statistics MAD (process) and KSS (process) of the non-linear nGAR process are better than the linear one. Hence, one would select the non-linear nGAR process as the process which fits better. We also noticed that the plots of the fitted non-linear nGAR process and the plots of the empirical time-series are alike. This is not the case considering plots of the linear nGAR process. The plots are not shown because of space limitations.

Fitting Models to Real Measurements

In this example, we fit theoretical distributions and a linear nGAR process to 20000 measurements taken by (Klemm, Lindemann and Lohmann 2002). The measurements describe the interarrival times of packets arriving at an Internet server. The measurements are correlated with AR coefficients $\underline{\alpha}$ of [0.07, 0.05, 0.04] and correlations $\underline{\rho}$ of [0.08, 0.07, 0.07].

For the fitting procedure, a function ψ of linear multiplication and an nGAR order $p = 3$ are considered. The AR coefficients $\underline{\alpha}$ are estimated using the Burg method. Next, the $\hat{Y}_t$ are estimated using (6). Noticing that the $\hat{Y}_t$ are (nearly) independent, we suggest that an order $p = 3$ is suitable. For the purpose of comparison, the real measurements are fitted directly to the theoretical distributions, neglecting the fact that the measurements are correlated. The statistical results of the two fittings are summarized in table 4.

Table 4: Results summary of fitting a theoretical distribution and a linear nGAR process to real measurements

	Theoretical distr.	**Linear nGAR**
[$\underline{\hat{\alpha}}$]	[0, 0, 0]	[0.07 , 0.05, 0.055]
$F(\hat{p}_1, \hat{p}_2)$	Weibull (0.64, 0.01)	*Weibull** (2.14, 0.02)
$\underline{\hat{\rho}}$	[0.006, -0.002, 0.002]	[.08, 0.065, 0.07]
MAD(process)	1.27	1.046
KSS(process)	0.2	0.288

$\underline{\hat{\alpha}}$ are the fitted AR coefficients. $F(\hat{p}_1, \hat{p}_2)$ are the fitted distributions and parameters. The best fitted distribution to the real measurements is Weibull, whereas the best fitted distribution to the estimated independent data $\hat{Y}_t$ when we fit an nGAR process is $Weibull^*$. The MAD test statistics and the KSS statistics of the fitted theoretical distribution and of the nGAR process are of the same order. This imply that this statistics gives no information about the model that fits the real measurements better. However the realizations from the nGAR process has similar correlations to that of the real measurements, whereas realizations from the fitted distribution are not correlated. This implies that the (correlations of) the nGAR process fits the real measurements better than the theoretical distribution.

CONCLUSION

Most dependence input modeling packages are based on Gaussian AR processes and random variables, because the behavior of Gaussian processes and random variables is mathematically well understood. However, most of the statistical estimators (e.g. covariance and correlation) and statistical methods (e.g. Yule-Walker method and test of independence of random variables) can be applied to Gaussian as well as to non-Gaussian AR (nGAR) processes. Unlike the other dependence modeling approaches, our approach can be easily integrated with the already existing input modeling tools for independent data. The approach also eliminates the Gaussian non-Gaussian transformations and provides higher flexibilities for the input modeler. This enables fitting heavy-tailed or non-linear nGAR processes to time-series.
The use of other optimization algorithms instead of the genetic algorithms and more detailed study of non-linear nGAR models are topics of future research.

REFERENCES

Biller B. and S. Ghosh, 2004, "Dependence modeling for stochastic simulation," in *Winter Simulation Conference*, pp. 153–161.

Biller B. and B. L. Nelson, 2002, "Advanced input modeling: Parameter estimation for arta processes," in *Winter Simulation Conference*, pp. 255–262.

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.

Box G. E. P. and G. M. Jenkins, 1976, *Time Series Analysis, Forecasting and Control.* Holden-Day, Inc.

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.

Chatfield C., 1996, *The Analysis of Time Series: An Introduction, 5th Edition.* New York: Chapman and Hall.

Chipperfield H. P. A., P. Fleming and C. Fonseca, 1994, "Genetic algorithm toolbox user's guide," ACSE Research Report No. 512, University of Sheffield.

Fishman G. S., 1978, *Principles of Discrete Event Simulation.* John Wiley and Sons, Inc.

Ghosh S. and S. G. Henderson, 2001, "Chessboard distributions," in *Winter Simulation Conference*, pp. 385–393.

Ghosh S. and S. G. Henderson, 2002a, "Chessboard distributions and random vectors with specified marginals and covariance matrix," *Operations Research Letters*, vol. 50, no. 5, pp. 820–834.

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

Johnson M. E., 1987, *Multivariate Statistical Simulation.* John Wiley and Sons, Inc.

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.

Strelen J. C., 2003, "The genetic algorithm is useful to fitting input probability distributions for simulation models," in *Business and Industry Symposium - ASTC*, pp. 8–13.