Proceedings of the 2001 Winter Simulation Conference
B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, eds.

MODELING AND GENERATING MULTIVARIATE TIME SERIES WITH ARBITRARY MARGINALS AND AUTOCORRELATION STRUCTURES

Bahar Deler
Barry L. Nelson

Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL, 60208-3119, U.S.A.

ABSTRACT

Providing accurate and automated input modeling support is one of the challenging problems in the application of computer simulation. In this paper, we present a general-purpose input-modeling tool for representing, fitting, and generating random variates from multivariate input processes to drive computer simulations. We explain the theory underlying the suggested data fitting and data generation techniques, and demonstrate that our framework fits models accurately to both univariate and multivariate input processes.

1 INTRODUCTION

Building a large-scale discrete-event stochastic simulation model may require the development of a large number of, possibly multivariate, input models. Development of these models is facilitated by accurate and automated (or nearly automated) input modeling support. Typical examples from manufacturing and service applications include the processing times of a workpiece across several work-centers, the medical characteristics of organ-transplant donors and recipients, and the times between arrivals of calls to a call center. We believe that the ability of an input model to represent the uncertainty in these phenomena is essential because even the most detailed logical model combined with a sound experimental design and thorough output analysis cannot compensate for inaccurate or irrelevant input models.

Interest among researchers and practitioners in modeling and generating input processes for stochastic simulation has led to commercial development of a number of input modeling packages, including ExpertFit (Averill M. Law and Associates, Inc.), the Arena Input Processor (Rockwell Software Inc.), Stat::Fit (Geer Mountain Software Corporation), and BestFit (Palisade Corporation). The input models incorporated in these packages sometimes fall short of what is needed because they emphasize good representations for the marginal distributions of independent and identically distributed (i.i.d.) processes. However, dependent and multivariate time-series input processes occur naturally in the simulation of many service, communications, and manufacturing systems (e.g., Melamed, Hill, and Goldsman 1992, Ware, Page, and Nelson 1998). Ignoring dependence can lead to performance measures that are seriously in error and a significant distortion of the simulated system.

The approach that the input modeling packages typically take for modeling the marginal distribution of i.i.d. data is to exhaustively fit and evaluate the fit of standard families of distributions (e.g., beta, Erlang, exponential, gamma, lognormal, normal, Poisson, triangular, uniform, Weibull, etc.), and then recommend as the input model the one with the best summary measures. However, the limited shapes represented by these distributions may not be flexible enough to represent some of the characteristics of the observed data or some known properties of the process that generated the data. Consequently, these input modeling packages are improved by expanding the list of distributions. Unfortunately, if the same philosophy is applied to modeling dependence, then the list of candidate *multivariate* distributions quickly explodes as we consider all possible combinations of the available marginal distributions.

The classical evaluation of a distribution fit is based on the hypothesis that there is a true, correct model among the list of candidates. In most simulation applications, the mechanisms generating real data of interest do not yield samples from *any* of the theoretical distributions under consideration, so the basic premise is false. A perfect goodness-of-fit test would, correctly, reject all candidates on the list. Therefore, we adopt the position that searching among a list of input models for the “true, correct" model is neither a theoretically supportable nor practically useful paradigm upon which to base general-purpose input modeling tools. Instead, we view input modeling as customizing a highly flexible model that can capture the important features present in data, while being easy to use, adjust, and understand. Thus, we propose to develop a single, but very general, input model, rather than a long list of more spe-

cialized models, by using a comprehensive input-modeling framework that can accomplish the following:

- Represent stationary multivariate time-series processes in such a way that univariate i.i.d. processes, stationary univariate time-series processes, and finite-dimensional random vectors are special cases of our model.
- Fit an input model to dependent and multivariate data via automated and statistically valid algorithms.
- Generate realizations of these input processes quickly and accurately in order to drive computer simulations.
- Develop a stand-alone, PC-based program that implements this framework for fitting and simulating input processes and generates random variates that can be read into any simulation.

Currently, we have a model for representing stationary multivariate time-series input processes with arbitrary autocorrelation structures and marginal distributions from the Johnson family. In Section 2, we give a brief overview of this model and refer the reader to Deler and Nelson (2001a) for details of the work completed to date. In the remainder of the paper, we address the problem of fitting input models to dependent and multivariate data. Section 3 provides a procedure for fitting models to univariate time-series data, and then discusses how to extend this procedure to fit models to multivariate time-series data. Section 4 presents computational results demonstrating the suggested algorithms. We give concluding remarks, together with the expected impact of this study, in Section 5.

2 OVERVIEW OF THE VARTA FRAMEWORK

We provide an input modeling framework for multivariate time-series processes with continuous marginal distributions by using a highly flexible model to capture the important features present in data. We achieve flexibility by combining Gaussian vector autoregressive processes and the Johnson family of distributions to characterize the process dependence and marginal distributions, respectively. Specifically, our framework is based on the ability to represent, fit, and generate random variates from a stationary k-variate vector time series $\{\mathbf{X}_t;\ t = 1, 2, \ldots\}$, where $\mathbf{X}_t = (X_{1,t}, X_{2,t}, \ldots, X_{k,t})'$ is a $(k \times 1)$ random vector of the observations recorded at time t. We do this by constructing a standard Gaussian vector autoregressive base process $\mathbf{Z}_t$ and transforming it to the desired multivariate input process $\mathbf{X}_t$. To achieve the target autocorrelation structure of the input process we adjust the autocorrelation structure of the base process. The ability of the Gaussian vector autoregressive base process to characterize dependencies in time sequence and with respect to other component series in the input process brings a significant flexibility to our framework.

In this paper, we are particularly interested in input modeling problems in which data are plentiful and nearly automated input modeling is required. Consequently, we use a member of the Johnson translation system of distributions to characterize the marginal distribution of each component series (Johnson 1949 and the Appendix). Our motivation for using the Johnson system is practical, rather than theoretical: In many applications, simulation output performance measures are insensitive to the specific input distribution chosen provided that enough moments of the distribution are correct (see, for instance, Gross and Juttijudata 1997). The Johnson system can match any feasible first four moments, while the standard families incorporated in existing software packages and simulation languages often match only one or two moments. Thus, the Johnson system enables us to represent key features of the data at hand, as opposed to finding the "true" distribution that was the source of the data.

To define our framework, let $\{Z_{i,t};\ t = 1, 2, \ldots\}$ be the ith component series of $\mathbf{Z}_t$, the k-variate Gaussian autoregressive base process of order p (denoted $\text{VAR}_k(p)$) with the representation

$$\mathbf{Z}_t = \sum_{h=1}^{p} \boldsymbol{\alpha}_h \mathbf{Z}_{t-h} + \mathbf{u}_t$$

(Lutkepohl 1993). The $\boldsymbol{\alpha}_h$, $h = 1, 2, \ldots, p$, are fixed $(k \times k)$ autoregressive coefficient matrices and $\mathbf{u}_t = (u_{1,t}, u_{2,t}, \ldots, u_{k,t})'$ is a k-dimensional white noise vector, representing that part of $\mathbf{Z}_t$ that is not linearly dependent on past observations. The structure of $\mathbf{u}_t$ is such that

$$\text{E}[\mathbf{u}_t] = \mathbf{0}_{(k\times 1)} \text{ and } \text{E}[\mathbf{u}_t \mathbf{u}'_{t+h}] = \begin{cases} \Sigma_u & \text{if } h = 0, \\ \mathbf{0}_{(k\times k)} & \text{otherwise.} \end{cases}$$

Choosing Σ_u appropriately ensures that each $Z_{i,t}$ is marginally standard normal.

Now let $\{X_{i,t};\ t = 1, 2, \ldots\}$ denote the ith component time series of the desired input process, for $i = 1, 2, \ldots, k$. In our framework, each of these univariate series has a Johnson marginal distribution. We reflect the desired dependence structure within and across series via Pearson product-moment correlations, denoted as $\rho_{\mathbf{X}}(i, j, h) \equiv \text{Corr}[X_{i,t}, X_{j,t+h}]$, for $h = 0, 1, 2, \ldots, p$. The ith time series is obtained via the transformation $X_{i,t} = F_{X_i}^{-1}[\Phi(Z_{i,t})]$, where F_{X_i} is the Johnson-type cumulative distribution function (cdf) suggested for the ith component series of the input process and $\Phi(\cdot)$ is the standard normal cdf. Notice that if $k = 1$, then this representation defines a univariate time-

series process; and if $k > 1$ but $p = 0$, then it defines a finite-dimensional random vector.

We refer to processes constructed in this way as having a VARTA (Vector-Autoregressive-To-Anything) distribution. To employ VARTA distributions we are required to match the desired correlation structure of the input process by manipulating the correlation structure of the Gaussian vector autoregressive base process. Deler and Nelson (2001a) suggest a way of selecting the correlation structure of the base process, $\rho_{\mathbf{Z}}(i, j, h)$, $i, j = 1, 2, \ldots, k$ and $h = 0, 1, 2, \ldots, p$ (except the case $i = j$ and $h = 0$), by solving correlation matching problems of the following form: Let $\rho = \rho_{\mathbf{Z}}(i, j, h)$ for convenience; then we need to find ρ such that

$$c_{ijh}[\rho] = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_{X_i}^{-1}[\Phi(z_i)]F_{X_j}^{-1}[\Phi(z_j)]\vartheta_\rho(z_i,z_j)dz_i dz_j - \mu_i\mu_j}{\sigma_i\sigma_j} = \rho_{\mathbf{X}}(i, j, h)$$

where $\vartheta_\rho(\cdot)$ is the standard bivariate normal probability density function with correlation ρ, $\mu_i = \mathrm{E}[X_{i,t}]$, $\mu_j = \mathrm{E}[X_{j,t+h}]$, $\sigma_i^2 = \mathrm{Var}[X_{i,t}]$, and $\sigma_j^2 = \mathrm{Var}[X_{j,t+h}]$. The reason that solving these correlation matching problems is sufficient is that the correlation between $X_{i,t}$ and $X_{j,t+h}$ is a function only of the correlation between $Z_{i,t}$ and $Z_{j,t+h}$, which appears in the expression for $\vartheta_\rho(\cdot)$. Thus, the problem of adjusting the correlation structure of the $\mathrm{VAR}_k(p)$ base process decomposes into $pk^2 + k(k-1)/2$ individual correlation matching problems in which we try to find the value $\rho_{\mathbf{Z}}(i, j, h)$ that makes $c_{ijh}[\rho_{\mathbf{Z}}(i, j, h)] = \rho_{\mathbf{X}}(i, j, h)$.

To summarize, the development of the VARTA framework has two major challenges. The first one is solving $pk^2 + k(k-1)/2$ correlation matching problems, for which Deler and Nelson (2001a) suggest a computationally feasible method. The second challenge is fitting k Johnson marginals to k-variate time-series data; we address this problem in the remainder of the paper.

3 FITTING VARTA MODELS

Input modeling packages, whether they are targeted for the simulation community or not, contain data fitting routines for the standard families of distributions. However, these routines typically assume i.i.d. data and they often use maximum likelihood estimators (MLEs) for the parameters of the distributions they fit. Unfortunately, these estimators are no longer the MLEs when the data are dependent; the true MLEs depend on the specification of the entire joint distribution of the process, a specification that is usually difficult to supply.

A robust method for fitting target distributions from Johnson's translation system to i.i.d. data is suggested by Swain, Venkatraman, and Wilson (1988) and implemented in software called FITTR1. They demonstrate the robustness and computational efficiency of least-squares, minimum L_1 norm, and minimum L_∞ norm techniques for estimating Johnson marginals, suggesting that similar techniques can be effectively adapted to fitting VARTA models to dependent and multivariate data. We outline our adaptation below.

Let $\{x_{i,t};\ i = 1, 2, \ldots, k;\ t = 1, 2, \ldots, n\}$ denote a sample from a stationary multivariate time-series input process. Our objective is to estimate the parameters of the Johnson marginals, $\lambda_i, \delta_i, \gamma_i, \xi_i$, $i = 1, 2, \ldots, k$, and $\mathrm{VAR}_k(p)$ base process, $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p$, and $\boldsymbol{\Sigma}_u$, so that a VARTA process provides an accurate representation of the input process. For ease of presenting the data fitting procedure below, we assume that the order of the underlying base process, p, and the types of Johnson marginals, F_{X_i}, $i = 1, 2, \ldots, k$, are known. Clearly, these also need to be determined in general. First we present a two-stage algorithm developed particularly for a univariate time-series input process and then discuss how to extend it to a multivariate time-series input process, which is more general and difficult.

3.1 The Univariate Case

Given a sample $\{x_t;\ t = 1, 2, \ldots, n\}$ from a univariate input process, we would like to determine the parameters of the Johnson marginal F, $\lambda, \delta, \gamma, \xi$, and the system parameters of the base process, $\alpha_1, \alpha_2, \ldots, \alpha_p$ and σ_u, such that the following model is a good fit:

$$\begin{aligned} X_t &= F^{-1}\left[\Phi^{-1}(Z_t)\right] \\ &= \xi + \lambda f^{-1}\left[\frac{Z_t - \gamma}{\delta}\right] \end{aligned}$$

where the base process is given by

$$Z_t = \sum_{h=1}^{p} \alpha_h Z_{t-h} + u_t \ \text{ with } \ u_t/\sigma_u \overset{i.i.d.}{\sim} N(0, 1).$$

The central idea is that if we have all of the parameter values correct, then

$$u_t = z_t - \sum_{h=1}^{p} \alpha_h z_{t-h} = \gamma + \delta f\left[\tfrac{x_t-\xi}{\lambda}\right] - \textstyle\sum_{h=1}^{p} \alpha_h \left(\gamma + \delta f\left[\tfrac{x_{t-h}-\xi}{\lambda}\right]\right) \quad (1)$$

for $t = p + 1, p + 2, \ldots, n$, will appear to be independent and identically distributed $\mathrm{N}(0, \sigma_u^2)$ random variables. We work iteratively between improving the estimates of $(\alpha_1, \alpha_2, \ldots, \alpha_p, \sigma_u)$ and of $(\gamma, \delta, \lambda, \xi)$, as follows:

Stage 0 The objective is to obtain starting values for the parameters of the Johnson marginal. We let $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ denote the order statistics based on the given sample and solve the following diagonally-weighted least-squares (DWLS) problem suggested by Swain, Venkatraman, and Wilson (1988):

$$\begin{aligned} &\underset{\gamma,\delta,\lambda,\xi}{\text{minimize}} \\ &\sum_{t=1}^{n} \frac{\left(\Phi\left\{\gamma+\delta f\left[\frac{x_{(t)}-\xi}{\lambda}\right]\right\}-\frac{t}{n+1}\right)^2}{\frac{t(n+1-t)}{(n+1)^2(n+2)}} \\ &\text{subject to} \quad \delta > 0 \\ &\qquad \lambda \begin{cases} > 0, & \text{for } S_U, \\ > x_{(n)}-\xi, & \text{for } S_B, \\ = 1, & \text{for } S_L \text{ and } S_N. \end{cases} \\ &\qquad \xi \begin{cases} < x_{(1)}, & \text{for } S_L \text{ and } S_B, \\ = 0, & \text{for } S_N. \end{cases} \end{aligned} \tag{2}$$

Stage 1 Keeping the estimates for the parameters of the Johnson marginal, $\widehat{\gamma}, \widehat{\delta}, \widehat{\lambda}, \widehat{\xi}$, fixed, we find the conditional least-squares estimators for the autoregressive coefficients of the base process, $\alpha_1, \ldots, \alpha_p$, and the residual variance, σ_u^2, by minimizing $\sum_{t=p+1}^{n} u_t^2$, where u_t is defined in (1). In order to ensure a stationary base process, we require the roots of $1 - \alpha_1 B - \alpha_2 B^2 - \cdots - \alpha_p B^p = 0$ to lie outside the unit circle (Lutkepohl 1993), where B is the backshift operator.

Stage 2 Keeping the estimates for the system parameters of the base process, $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_p, \widehat{\sigma}_u$, fixed, we solve the following conditional DWLS problem for the parameters of the Johnson marginal distribution:

$$\begin{aligned} &\underset{\gamma,\delta,\lambda,\xi}{\text{minimize}} \\ &\sum_{t=p+1}^{n} \frac{\left(\Phi\left\{\frac{u_{(t)}}{\widehat{\sigma}_u}\right\}-\frac{t}{n+1}\right)^2}{\frac{t(n+1-t)}{(n+1)^2(n+2)}} \\ &\text{subject to} \quad (2) \end{aligned} \tag{3}$$

where $u_{(p+1)}, u_{(p+2)}, \ldots, u_{(n)}$ are the order statistics of the residuals (1). If the fit has improved significantly since the last iteration, then go to Stage 1; else, stop the procedure and report the result.

We now briefly explain each stage of the procedure: Stage 0 performs least-squares fitting, as suggested by Swain, Venkatraman, and Wilson (1988), by treating the given sample points as independent. If the model is correct, then the transformed random variate $\Phi\left\{\gamma + \delta f[(x_{(t)} - \xi)/\lambda]\right\}$, which is equivalent to $F[x_{(t)}]$, has the distribution of the tth uniform order statistic—that is, the smallest observation in a random sample of size n from the uniform distribution on the unit interval (0,1); thus, its mean and variance are given by $t/(n+1)$ and $t(n+1-t)/\left((n+1)^2(n+2)\right)$, respectively (Kendall and Stuart 1979). The fitting is based on minimizing the quadratic distance between the parametric approximation of the transformed random variate to the uniformized order statistics and the corresponding expected value. We also incorporate a diagonal weight matrix, whose diagonal entries are reciprocals of the variances of the uniform order statistics, into the corresponding objective function. Motivation comes from the discussion in Kuhl and Wilson (1999) on the performance of the weighted least-squares and the ordinary least-squares procedures, and also the empirical evidence supporting the superiority of the DWLS fitting procedure (particularly for the Johnson translation system) to the fits based on the conventional weighted least-squares procedures (Swain, Venkatraman, and Wilson 1988).

Stage 1 estimates the autoregressive coefficients, $\alpha_1, \alpha_2, \ldots, \alpha_p$, and the residual variance, σ_u^2, which imply a stationary autoregressive process. Asymptotically, the least-squares estimators fall into the stability region of the corresponding base process. However, we have observed that whether or not the least-squares estimators correspond to a stationary process depends on the sample size, n. Therefore, we carry out this stage in such a way that it always ensures stationarity of the underlying base process.

In Stage 2, we solve the DWLS procedure to estimate the parameters of the Johnson distribution of the input process. The formulation (3) is based on the fact that if the model is correct, then u_t/σ_u forms a sequence of independent and identically distributed standard normal random variables.

3.2 The Multivariate Case

With multivariate time-series data, the fitting process is complicated by the need to compute not only the serial dependence within the component series, but also the interdependence among the component series. Below, we briefly discuss how to extend the procedure in Section 3.1 to the multivariate case:

Stage 0 We obtain starting values for the parameters of the Johnson marginals by implementing Stage 0 of the univariate procedure for each component series. For $i = 1, 2, \ldots, k$, we let $x_{i,(1)} \le x_{i,(2)} \le \cdots \le x_{i,(n)}$ denote the order statistics based on the sample given for the ith component series and

solve the following DWLS problem:

$$\underset{\gamma_i,\,\delta_i,\,\lambda_i,\,\xi_i}{\text{minimize}} \quad \sum_{t=1}^{n} \frac{\left(\Phi\left\{\gamma_i + \delta_i f\left[\frac{x_{i,(t)}-\xi_i}{\lambda_i}\right]\right\} - \frac{t}{n+1}\right)^2}{\frac{t(n+1-t)}{(n+1)^2(n+2)}}$$

$$\text{subject to} \quad \delta_i > 0 \tag{4}$$

$$\lambda_i \begin{cases} > 0, & \text{for } S_U, \\ > x_{i,(n)} - \xi_i, & \text{for } S_B, \\ = 1, & \text{for } S_L \text{ and } S_N. \end{cases}$$

$$\xi_i \begin{cases} < x_{i,(1)}, & \text{for } S_L \text{ and } S_B, \\ = 0, & \text{for } S_N. \end{cases}$$

Stage 1 Keeping the estimates for the parameters of the Johnson marginals, $\widehat{\gamma}_i, \widehat{\delta}_i, \widehat{\lambda}_i, \widehat{\xi}_i$, for $i = 1, 2, \ldots, k$, fixed, we find the multivariate least-squares estimators of $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p$, and $\boldsymbol{\Sigma}_u$ by minimizing $\sum_{t=p+1}^{n} \mathbf{u}_t \mathbf{u}_t'$. In order to ensure a stationary base process, we require the roots of $|\mathbf{I}_{(k\times k)} - \boldsymbol{\alpha}_1 B - \boldsymbol{\alpha}_2 B^2 - \cdots - \boldsymbol{\alpha}_p B^p| = 0$ to lie outside the unit circle ($\mathbf{I}_{(k\times k)}$ is the $(k \times k)$ identity matrix).
Stage 2 Keeping the estimates for the system parameters of the VAR$_k(p)$ base process, $\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\alpha}}_2, \ldots, \widehat{\boldsymbol{\alpha}}_p, \widehat{\boldsymbol{\Sigma}}_u$, fixed, we modify the Johnson parameters for each component series by implementing Stage 2 of the univariate procedure. For $i = 1, 2, \ldots, k$, we solve the following conditional DWLS problem:

$$\underset{\gamma_i,\,\delta_i,\,\lambda_i,\,\xi_i}{\text{minimize}} \quad \sum_{t=p+1}^{n} \frac{\left(\Phi\left\{\frac{u_{i,(t)}}{\sqrt{\widehat{\boldsymbol{\Sigma}}_u(i,i)}}\right\} - \frac{t}{n+1}\right)^2}{\frac{t(n+1-t)}{(n+1)^2(n+2)}}$$

$$\text{subject to} \quad (4)$$

where $\widehat{\boldsymbol{\Sigma}}_u(i,i)$ corresponds to the estimate for the residual variance of the ith component series. If the fit has improved significantly since the last iteration, then go to Stage 1; else, stop the procedure and report the result.

4 IMPLEMENTATION

In this section, we present computational results obtained from the implementation of the fitting procedures. In these preliminary results, we assume autoregressive base processes of orders 1 and 2 for the univariate case, and of order 1 for the multivariate case. In all of these experiments, we generate 1000 observations from a stationary VARTA process, then see how well our procedures recover the true parameters of the process.

Table 1: Absolute Difference and Relative Percent Difference between the Estimates and the True Parameters when $p = 1$

ρ_X	0.90		0.55	
	E_1	E_2	E_1	E_2
$\widehat{\rho}_Z(1)$	4.129×10^{-3}	0.453	3.249×10^{-3}	0.499
$\widehat{\gamma}$	14.686×10^{-3}	2.084	13.314×10^{-3}	1.889
$\widehat{\delta}$	19.407×10^{-3}	1.778	16.567×10^{-3}	1.518
$\widehat{\lambda}$	4.708×10^{-3}	0.898	4.338×10^{-3}	0.828
$\widehat{\xi}$	10.215×10^{-3}	1.851	3.605×10^{-3}	0.653
ρ_X	-0.55		-0.30	
	E_1	E_2	E_1	E_2
$\widehat{\rho}_Z(1)$	4.875×10^{-3}	0.636	3.832×10^{-3}	0.909
$\widehat{\gamma}$	18.922×10^{-3}	2.685	14.954×10^{-3}	2.122
$\widehat{\delta}$	41.247×10^{-3}	3.779	16.615×10^{-3}	1.523
$\widehat{\lambda}$	14.138×10^{-3}	2.698	8.731×10^{-3}	1.666
$\widehat{\xi}$	21.645×10^{-3}	3.923	15.268×10^{-3}	2.767

Table 2: KS and PM Tests when $p = 1$

	$\boldsymbol{X}_t$	$\boldsymbol{Z}_t$	
ρ_X	KS	PM	KS
0.90	15.643×10^{-3}	6.544	0.638
0.55	2.021×10^{-3}	7.031	0.664
-0.55	11.142×10^{-3}	7.095	0.443
-0.30	13.300×10^{-3}	7.105	0.452

4.1 The Univariate Case

Using the procedure in Section 3.1, we fit a VARTA model to a sample of 1000 observations of a univariate VARTA process of order $p = 1$ or 2 with Johnson unbounded (S_U) marginal distribution having parameters $(\gamma, \delta, \lambda, \xi) = (-0.705, 1.091, 0.524, -0.552)$. The eight experiments reported here differ in their correlation structures. Tables 1 and 3 report the absolute difference (E_1) and relative percent difference (E_2) between the fitted Johnson parameters/base correlation structures and the true Johnson parameters/base correlation structures used to generate the sample data. Tables 2 and 4 present the results of the Kolmogorov-Smirnov (KS) test and the diagnostic checking on the residuals. We use the KS test to check whether the fitted distribution $\Phi\left\{\widehat{\gamma} + \widehat{\delta} f[(x - \widehat{\xi})/\widehat{\lambda}]\right\}$ differs from the true Johnson S_U distribution that generated the data, and to check whether the $\widehat{u}_t$'s are normally distributed; we use the Portmanteau (PM) test for checking whether the $\widehat{u}_t$'s are uncorrelated random shocks with zero mean and constant variance, that is, $\{\widehat{u}_t; t = p+1, p+2, \ldots, n\}$ is white noise.

Table 3: Absolute Difference and Relative Percent Difference between the Estimates and the True Parameters when $p = 2$

$\boldsymbol{\rho}_X$	$(0.50, 0.10)'$		$(0.55, -0.10)'$	
	E_1	E_2	E_1	E_2
$\widehat{\rho}_Z(1)$	6.164×10^{-3}	1.028	9.799×10^{-3}	1.508
$\widehat{\rho}_Z(2)$	0.990×10^{-3}	0.739	1.688×10^{-3}	1.227
$\widehat{\gamma}$	10.256×10^{-3}	1.455	23.524×10^{-3}	3.338
$\widehat{\delta}$	21.477×10^{-3}	1.968	36.507×10^{-3}	3.346
$\widehat{\lambda}$	11.476×10^{-3}	2.189	11.278×10^{-3}	2.152
$\widehat{\xi}$	12.265×10^{-3}	2.223	10.744×10^{-3}	1.947
$\boldsymbol{\rho}_X$	$(-0.25, -0.20)'$		$(-0.40, -0.10)'$	
	E_1	E_2	E_1	E_2
$\widehat{\rho}_Z(1)$	10.596×10^{-3}	3.032	9.159×10^{-3}	1.625
$\widehat{\rho}_Z(2)$	10.804×10^{-3}	3.885	2.341×10^{-3}	1.701
$\widehat{\gamma}$	34.952×10^{-3}	4.959	35.354×10^{-3}	5.016
$\widehat{\delta}$	66.617×10^{-3}	6.105	31.403×10^{-3}	2.878
$\widehat{\lambda}$	29.088×10^{-3}	5.549	24.158×10^{-3}	4.609
$\widehat{\xi}$	23.045×10^{-3}	4.177	32.614×10^{-3}	5.911

Table 4: KS and PM Tests when $p = 2$

	$\boldsymbol{X}_t$	$\boldsymbol{Z}_t$	
$\boldsymbol{\rho}_X$	KS	PM	KS
$(0.50, 0.10)'$	12.841×10^{-3}	8.922	0.772
$(0.55, -0.10)'$	5.200×10^{-3}	53.856	0.593
$(-0.25, -0.20)'$	29.204×10^{-3}	8.055	0.825
$(-0.40, -0.10)'$	2.569×10^{-3}	7.276	0.713

The results in Tables 1 and 2 and in Tables 3 and 4 are obtained by performing three and five iterations of the procedure, respectively. During the execution of the procedure, the estimates for the parameters of the marginal distribution were observed to be sensitive to changes in the characterization of the base process. On the other hand, the estimated correlation structure was found to be relatively insensitive to adjustments in the parameters of the marginal distribution. This observation suggests that the procedure gives pretty robust fits for the correlation structure of the base process quickly, while the success of getting the right marginals in the second stage is slower to converge.

In the second columns of Tables 2 and 4, we report the KS test statistics indicating the maximum absolute differences between the cdfs of the fitted and the true Johnson distributions. Comparison of these statistics to the distribution-free critical value, 1.358, at a significance level of 5%, suggests that the fitted Johnson marginals are good representations of the true distributions. Although the critical value of 1.358 is for a test based on i.i.d. data, it still provides a rough guide for judging the adequacy of our fit.

The last columns of Tables 2 and 4 give KS test statistics for the normality of the residuals. At a significance level of 5%, we expect them to be less than 0.895 and, in all cases, we find the residuals of the base process statistically normal.

Table 5: Estimated Residual Correlations for Lags $h = 1, 2, \ldots, 10$

h	$\widehat{\rho}_u(h)$	h	$\widehat{\rho}_u(h)$
1	0.067	6	0.026
2	−0.033	7	0.096
3	−0.087	8	0.079
4	−0.123	9	0.037
5	−0.083	10	−0.027

Finally, in the third columns of Tables 2 and 4, we present the PM test statistics that measure departures of the residuals from being white noise. At a significance level of 5%, the test statistics are expected to be less than 16.919 for $p = 1$ and 15.507 for $p = 2$ to support the white noise hypothesis for the residuals. We find that the residuals are statistically white noise in all cases except when $\boldsymbol{\rho}_X = (0.55, -0.10)'$. However, a close look at the estimated residual correlations in Table 5 indicates that we still have a reasonably good fit for the corresponding base process. Thus, the algorithm suggested for the univariate case can successfully capture the underlying VARTA framework incorporated into the sample data.

4.2 The Multivariate Case

In this section, we fit a VARTA model to a sample of 1000 trivariate ($k = 3$) VARTA order 1 ($p = 1$) observations with Johnson marginals that are lognormal ($\gamma_1 = -1.462, \delta_1 = 2.236, \lambda_1 = 1, \xi_1 = -2.125$); unbounded ($\gamma_2 = -0.730, \delta_2 = 1.905, \lambda_2 = 1.521, \xi_2 = -0.686$); and bounded ($\gamma_3 = 1.258, \delta_3 = 0.426, \lambda_3 = 4.421, \xi_3 = -0.694$). Further, the base correlation matrices are specified at lags 0 and 1 as

$$\Sigma_Z(0) = \begin{pmatrix} 1.00000 & 0.37671 & 0.46049 \\ 0.37671 & 1.00000 & 0.29906 \\ 0.46049 & 0.29906 & 1.00000 \end{pmatrix}$$

and

$$\Sigma_Z(1) = \begin{pmatrix} 0.30689 & 0.24408 & 0.12449 \\ 0.13745 & 0.28600 & 0.33600 \\ 0.14104 & 0.30177 & 0.21864 \end{pmatrix}.$$

This example is taken from Deler and Nelson (2001a). As in the previous section, we report the absolute difference and relative percent difference between the estimates for the Johnson parameters and the base correlation structure and the true values used to generate the sample data; see Tables 6 and 8. These results are obtained at the end of the 7th iteration.

Table 6: Absolute Difference and Relative Percent Difference between the Estimates and the True Parameters for the Johnson Marginals

	$\boldsymbol{X}_1$		$\boldsymbol{X}_2$		$\boldsymbol{X}_3$	
	E_1	E_2	E_1	E_2	E_1	E_2
$\widehat{\gamma}$	0.016	1.094	0.017	2.329	0.060	4.769
$\widehat{\delta}$	0.065	2.907	0.111	5.827	0.024	5.634
$\widehat{\lambda}$	0.000	0.000	0.059	3.879	0.002	0.045
$\widehat{\xi}$	0.076	3.576	0.029	4.227	0.000	0.000

Table 7: KS Tests for each Component Series

	$\text{KS}_{\mathbf{X}}$	$\text{KS}_{\mathbf{Z}}$
$\boldsymbol{X}_1$	2.177×10^{-2}	0.662
$\boldsymbol{X}_2$	4.039×10^{-2}	0.423
$\boldsymbol{X}_3$	1.562×10^{-2}	0.656

Table 8: Absolute Difference and Relative Percent Difference between the Estimates and the True Parameters for the Base Correlation Structure

$\rho_{\mathbf{Z}}(i,j,h)$	E_1	E_2	$\rho_{\mathbf{Z}}(i,j,h)$	E_1	E_2
$\rho_{\mathbf{Z}}(1,2,0)$	0.022	5.836	$\rho_{\mathbf{Z}}(2,1,1)$	0.022	16.058
$\rho_{\mathbf{Z}}(1,3,0)$	0.011	2.391	$\rho_{\mathbf{Z}}(2,2,1)$	0.004	1.399
$\rho_{\mathbf{Z}}(2,3,0)$	0.001	0.334	$\rho_{\mathbf{Z}}(2,3,1)$	0.016	4.762
$\rho_{\mathbf{Z}}(1,1,1)$	0.004	1.303	$\rho_{\mathbf{Z}}(3,1,1)$	0.024	17.021
$\rho_{\mathbf{Z}}(1,2,1)$	0.003	1.229	$\rho_{\mathbf{Z}}(3,2,1)$	0.007	2.318
$\rho_{\mathbf{Z}}(1,3,1)$	0.004	3.226	$\rho_{\mathbf{Z}}(3,3,1)$	0.002	0.913

The results of the KS tests in Table 7 indicate that the residuals of the base process are statistically normal and the characterizations of the components using the estimated Johnson marginals provide adequate representations of the true marginals, while the PM test does not support the hypothesis that the residuals are white noise. However, the estimated residual correlation matrices for lags 1 through 4 in Table 9 show that the fitted $\text{VAR}_3(1)$ model might still adequately represent the underlying base process.

Fitting an input model to multivariate data by our framework requires the validation of a Gaussian base process. The two-stage algorithm suggested in Section 3.2 ensures the normality and the independence of the residuals of the base process at the component level in Stage 2, while the cross-correlations are incorporated into the model at Stage 1. This approach results in robust fits for the base process, leading to statistically acceptable marginal fits; thus, we get a reasonably good VARTA fit to the trivariate system of this section. However, in a $\text{VAR}_k(p)$ process, individually each component series follows a univariate mixed autoregressive/moving-average model up to a maximum order $(kp, (k-1)p)$, while the procedure assumes that the underlying base process follows an autoregressive model of order p. Therefore, assessing the significance of the normality and the independence of the residuals at the component level, rather than the multivariate level, is an issue that we look into for further validation of the theoretical framework on which our fitting procedure is based. Exploration of alternative procedures for the multivariate case is the major focus of our ongoing research (Deler and Nelson 2001b).

Table 9: Estimated Residual Correlation Matrices for Lags $h = 1, 2, 3, 4$

h	$\widehat{\rho}_{\mathbf{u}}(h)$			h	$\widehat{\rho}_{\mathbf{u}}(h)$		
1	−0.07	0.03	−0.13	3	0.00	0.02	0.01
	−0.16	0.04	−0.36		−0.01	0.00	−0.01
	−0.04	0.14	−0.05		0.00	0.01	−0.02
2	0.00	0.00	−0.01	4	0.01	−0.04	0.02
	−0.07	−0.02	−0.14		−0.01	−0.03	0.01
	0.00	0.01	−0.09		0.02	0.03	0.02

5 CONCLUSION

In this paper, we focus on developing automated and statistically valid algorithms to fit stochastic input models to multivariate input processes. In order to demonstrate the validity of our algorithms, we fit input models to data, which are simply generated by true VARTA distributions. However, it is essential that we use real data and measure the (possible) systematic deviation from the VARTA distribution, which we propose as a flexible alternative to capturing the true distribution. This is one of the major issues that we aim to explore through a comprehensive numerical analysis on the suggested procedure in Deler and Nelson (2001b).

At the end of our ongoing research, we will have developed a stand-alone, PC-based program that implements the VARTA framework with the suggested data fitting and data generation techniques for simulating input processes. The key computational components of the software are written in portable C code in such a way that we can make them available individually for incorporation into commercial products. This way, we expect the product of this research to take reliable input modeling out of the domain of statistical specialists and put it into the hands of everyday simulation users.

ACKNOWLEDGMENTS

This research was partially supported by National Science Foundation Grant numbers DMI-9821011 and DMI-9900164.

APPENDIX: JOHNSON FAMILY OF DISTRIBUTIONS

The Johnson translation system for a random variable X, whose range depends on the family of interest, is defined

by a cumulative distribution function (cdf) of the form $F_X(x) = \Phi\{\gamma + \delta f[(x-\xi)/\lambda]\}$, where $\Phi(\cdot)$ is the cdf of the standard normal distribution, γ and δ are shape parameters, ξ is a location parameter, λ is a scale parameter, and $f(\cdot)$ is one of the following transformations:

$$f(y) = \begin{cases} \log(y) & \text{for the } S_L \text{ (lognormal) family,} \\ \sinh^{-1}(y) & \text{for the } S_U \text{ (unbounded) family,} \\ \log\left(\frac{y}{1-y}\right) & \text{for the } S_B \text{ (bounded) family,} \\ y & \text{for the } S_N \text{ (normal) family.} \end{cases}$$

There is a unique family (choice of f) for each feasible combination of the skewness and the kurtosis, which determine the parameters γ and δ. Any mean and (positive) variance can be attained by any one of the families by manipulation of the parameters λ and ξ. Within each family, a distribution is completely specified by the values of the parameters γ, δ, λ, and ξ.

In our framework, the characterization of the input process using the Johnson system simplifies the evaluation of the composite function $F_X^{-1}[\Phi(z)]$ significantly because $F_X^{-1}[\Phi(z)] = \xi + \lambda f^{-1}[(z-\gamma)/\delta]$, where

$$f^{-1}(y) = \begin{cases} e^y & \text{for the } S_L \text{ (lognormal) family,} \\ \frac{e^y - e^{-y}}{2} & \text{for the } S_U \text{ (unbounded) family,} \\ \frac{1}{1+e^{-y}} & \text{for the } S_B \text{ (bounded) family,} \\ y & \text{for the } S_N \text{ (normal) family.} \end{cases}$$

REFERENCES

Deler, B. and B. L. Nelson. 2001a. Modeling and generating multivariate time series with arbitrary marginals using a vector autoregressive technique. Working paper, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Deler, B. and B. L. Nelson. 2001b. Fitting dependent multivariate time series with arbitrary marginals and correlation structure. Working paper, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Gross, D. and M. Juttijudata. 1997. Sensitivity of output performance measures to input distributions in queueing simulation modeling. In *Proceedings of the 1997 Winter Simulation Conference*, ed. S. Andradottir, K. J. Healy, D. H. Withers, and B. L. Nelson, pp. 296–302. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Johnson, N. L. 1949. Systems of frequency curves generated by methods of translation. *Biometrika* 36 (1): 297-304.

Kendall, M. G. and A. Stuart. 1979. *The Advanced Theory of Statistics*. New York: Macmillan.

Kuhl, M. E. and J. R. Wilson. 1999. Least-squares estimation of non-homogeneous Poisson processes. *Journal of Statistical Computation and Simulation* 67: 75-108.

Lutkepohl, H. 1993. *Introduction to Multiple Time Series Analysis*. New York: Springer-Verlag.

Melamed, B., J. R. Hill, and D. Goldsman. 1992. The TES methodology: Modeling empirical stationary time series. In *Proceedings of the 1992 Winter Simulation Conference*, ed. J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson, pp. 135–144. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Swain, J. J., S. Venkatraman, and J. R. Wilson. 1988. Least-squares estimation of distribution functions in Johnson's translation system. *Journal of Statistical Computation and Simulation* 29: 271-297.

Ware, P. P., T. W. Page, and B. L. Nelson. 1998. Automatic modeling of file system workloads using two-level arrival processes. *ACM Transactions on Modeling and Computer Simulation* 8 (3): 305-330.

AUTHOR BIOGRAPHIES

BAHAR DELER is a Ph.D. candidate in the Department of Industrial Engineering and Management Sciences at Northwestern University. She received her B.S. and M.S. in industrial engineering from Bilkent University, Turkey. Her research interests include computer simulation of stochastic systems and stochastic input modeling. Her e-mail and web addresses are <bahar@iems.nwu.edu> and <www.iems.northwestern.edu/~bahar>.

BARRY L. NELSON is a Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University, and is Director of the Master of Engineering Management Program there. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems. He has published numerous papers and two books. Nelson has served the profession as the Simulation Area Editor of *Operations Research* and President of the INFORMS (then TIMS) College on Simulation. He has held many positions for the Winter Simulation Conference, including Program Chair in 1997 and current membership on the Board of Directors. His e-mail and web addresses are <nelsonb@northwestern.edu> and <www.iems.northwestern.edu/ñelsonb>.