

FITTING TIME-SERIES INPUT PROCESSES FOR SIMULATION

Bahar Biller

*Graduate School of Industrial Administration
Carnegie Mellon University*

Barry L. Nelson

*Department of Industrial Engineering & Management Sciences
Northwestern University*

October 2002

Abstract

Providing accurate and automated input modeling support is one of the challenging problems in the application of computer simulation. The models incorporated in current input-modeling software packages often fall short because they assume independent and identically distributed processes, even though dependent time-series processes occur naturally in many real-life systems. This paper introduces a statistical methodology for fitting stochastic models to dependent time-series input processes. Specifically, an automated and statistically valid algorithm is presented to fit autoregressive-to-anything processes with marginal distributions from the Johnson translation system to stationary univariate time-series data. The use of this algorithm is illustrated with examples.

Key Words: Correlation, estimation, time series

1 INTRODUCTION

Dependent time-series input processes occur naturally in the simulation of many service, communications, and manufacturing systems. For example, Melamed, Hill, and Goldsman (1992) observe autocorrelation in sequences of compressed video frame bitrates, while Ware, Page, and Nelson (1998) report that the times between file accesses on a computer network frequently exhibit burstiness, as characterized by a sequence of short interaccess times followed by one or more long ones. Later in this article, we model a pressure variable of a continuous-flow production line and sales of a large vehicle manufacturer that are recorded at fixed time intervals; both exhibit strong series dependence. Ignoring these dependencies can lead to performance measures that are seriously in error and a significant distortion of the simulated system. An illustration is given by Livny, Melamed, and Tsiolis (1993) who examine the impact of autocorrelation on queueing systems.

Much of the past work on time-series input processes is based on linear models, such as the autoregressive moving average class or those that underlie Kalman filtering and related methods (Chatfield 1999). Mallows (1967) shows that the linearity of these models imply normal marginal distributions, but there are many physical situations in which the marginals of the time series are non-normal. Motivated by this, there has been considerable research on modeling time series with marginals from specific families, such as exponential, gamma, geometric, or general discrete marginal distributions (see, for example, Lewis, McKenzie, and Hugus 1989). However, these models often allow only limited control of the dependence structure and a different model is required for each type of marginal distribution.

A way to overcome these limitations is to construct the desired process by a monotone transform of a Gaussian linear process. For example, Cario and Nelson (1996, 1998) take this approach to develop models for representing and generating stationary univariate time-series processes. The central idea is to transform a Gaussian autoregressive process into the desired univariate time-series input process that they presume as having an ARTA (Autoregressive-To-Anything) distribution. The authors manipulate the correlations of the corresponding Gaussian process so that they achieve the desired correlations for the simulation input process. They assume — as is common in the simulation input-modeling literature — that the desired marginal distribution and dependence structure (specified via correlations) are given. However, there is no rigorously justified method

for fitting the input model when only raw data generated by an unknown process are available. To fill this gap, we solve the problem of fitting stochastic input models to stationary univariate time-series data and, specifically, fit ARTA processes with marginal distributions from the Johnson translation system.

When simple models, which assume a sequence of independent and identically distributed (i.i.d.) random variables and standard marginal distributions, do apply, there are a number of software packages, packages that support automated (or nearly automated) input modeling, including ExpertFit (Averill M. Law and Associates, Inc.), the Arena Input Analyzer (Rockwell Software Inc.), Stat::Fit (Geer Mountain Software Corporation), and BestFit (Palisade Corporation). Because of the lack of a widely available, general-purpose method for fitting time-series input processes, simulation practitioners fit the marginal distributions using these input-modeling software packages that often use maximum likelihood estimators (MLEs) for the parameters of the distributions under the assumption of i.i.d. data. Unfortunately, when the data are dependent, these estimators are no longer the MLEs; the true MLEs depend on the specification of the entire joint distribution of the process, a specification that is usually difficult to produce (Kotz, Balakrishnan, and Johnson 2000).

A number of researchers have explored the extent to which large-sample properties such as consistency, asymptotic sufficiency, efficiency, and normality of MLEs for i.i.d. data carry over to more general stochastic processes. The literature for generally dependent observations includes Wald (1949), Silvey (1961), Hartley and Rao (1967), Bar-Shalom (1971), Weiss (1971, 1973), Bhat (1974), Crowder (1976), Basawa, Feigin, and Heyde (1976), Basawa and Rao (1980), Sweeting (1980), Heijmans and Magnus (1986), and Sarma (1986). While these papers are important contributions to the MLE literature, they impose conditions that are either very restrictive or difficult to verify. Further, all of these theoretical results are based on the assumption that the MLEs exist, which is not always the case.

We conclude that neither existing research nor available software packages provide a general-purpose method for fitting time-series input processes despite the fact that ignoring dependence can lead to erroneous decisions. Therefore, the purpose of this paper is to solve the problem of fitting stochastic input models to stationary univariate time-series data and develop a general purpose data-fitting algorithm. We focus on marginal distributions from the Johnson translation system.

While this might seem restrictive, it is less so than it first appears: 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 translation system can match any feasible finite 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 translation system enables us to estimate key features of the data at hand, as opposed to finding the “true” distribution that was the source of the data. The Johnson translation system is particularly compatible with the ARTA approach (see Section 2), but all of our results can be easily extended to any other continuous distribution.

In summary, we develop an algorithm to fit stochastic input models to univariate time-series data without having to rely on any hypothesis about the physical mechanism generating them. We focus on processes having marginals from the Johnson translation system, but discuss how our approach generalizes to other continuous marginal distributions. To facilitate a detailed discussion of the data-fitting problem, we first review the essential ideas involved in ARTA processes introduced by Cario and Nelson (1996). The rest of the paper is organized around the presentation of the three key levels of solving the data-fitting problem. In Section 3, we provide the iterative fitting algorithm, prove its convergence to a stationary solution, and show the consistency properties of the resulting estimators as the sample size approaches infinity. Section 4 presents the numerical methods used to implement the suggested algorithm and Section 5 provides illustrative examples demonstrating the use of the algorithm. We conclude with directions for future research and the expected impact of this new statistical methodology on the development of stochastic input models for simulation in Section 6.

2 OVERVIEW OF THE ARTA FRAMEWORK

In this section, we introduce the notation we will use and provide a brief review of the Johnson translation system; we then describe ARTA processes.

2.1 Notation

We let the generic univariate input random variable be denoted by X , with marginal cumulative distribution function (cdf) F_X . The cdf of the standard normal distribution is denoted by Φ and its

probability density function by ϕ . The mean of a random variable is denoted by μ and its variance by σ^2 .

A stationary univariate time-series input process is denoted by $\{X_t; t = 1, 2, \dots\}$. The term “time series” means that the random variables may be dependent in sequence, such as the month-to-month demands for a product by a customer. We denote any realization of length n from the input process X_t by $\{x_t; t = 1, 2, \dots, n\}$. Boldface type is used to denote column vectors, e.g., $\mathbf{x} = (x_1, x_2, \dots, x_n)'$.

We account for dependence between random variables that are lag- h apart, say X_t and X_{t-h} , via their product-moment correlation defined as $\rho_X(h) = E[\sigma^{-2}(X_t - \mu)(X_{t-h} - \mu)]$, where X_t has mean μ and variance σ^2 for all t due to the assumption of a stationary input process. Representation of dependence by product-moment correlation is a practical compromise we make in this paper. Many other measures of dependence have been defined (see Nelsen 1998) and they are arguably more informative than the product-moment correlation for some distribution pairs. However, product-moment correlation is the only measure of dependence that is widely used and understood in engineering applications. We believe that making it possible for simulation users to incorporate dependence by product-moment correlation, while limited, is substantially better than ignoring dependence. Further, the ARTA model is flexible enough to incorporate dependence measures that remain unchanged under strictly increasing transformations of the random variables, such as Spearman’s rank correlation and Kendall’s τ , should those measures be desired.

2.2 The Johnson Translation System

The Johnson translation system for a random variable X is defined by a cdf of the form

$$F_X(x) = \Phi \left\{ \gamma + \delta f \left[\frac{x - \xi}{\lambda} \right] \right\},$$

where 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:

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

There is a unique family (choice of f) for each feasible combination of finite skewness and kurtosis that determine the parameters γ and δ . Any mean and (positive) variance can be attained by any one of the families by the manipulation of the parameters λ and ξ . Within each family, a distribution is completely specified by the values of the parameters $(\gamma, \delta, \lambda, \xi)$; the range of X depends on the family of interest (Johnson 1949).

The Johnson translation system provides good representations for unimodal distributions and can represent certain bimodal shapes, but not three or more modes. Illustrations of the shapes of the Johnson-type probability density functions can be found in Johnson (1987). The first four moments of all distributions in the families S_L , S_B , S_U , and S_N are finite. Nevertheless, the ability to match any (finite) first four moments provides a great deal of flexibility that is sufficient for many practical problems.

2.3 ARTA Processes

An ARTA process is a time series with arbitrary marginal distribution and autocorrelation structure specified through finite lag p . It is based on the construction of a Gaussian standard time-series $\{Z_t; t = 1, 2, \dots, n\}$ as a base process, from which we obtain a series of autocorrelated $(0, 1)$ uniform random variables $\{U_t; t = 1, 2, \dots, n\}$ by using the probability-integral transformation $U_t = \Phi(Z_t)$. The transformation $X_t = F_X^{-1}[U_t]$ is then applied, ensuring that the input time-series process $\{X_t; t = 1, 2, \dots, n\}$ has the desired marginal distribution F_X . This approach works for *any* marginal distribution, although F_X^{-1} may have to be evaluated by an approximate numerical method when there is no exact closed-form expression. The inverse cdf method is an essential ingredient of the framework described in the remainder of this section.

In the ARTA framework, the base process $\{Z_t; t = 1, 2, \dots, n\}$ is a stationary, standard Gaussian

autoregressive process of order p (denoted by $\text{AR}(p)$) with the representation

$$Z_t = \sum_{h=1}^p \alpha_h Z_{t-h} + Y_t, \quad t = 1, 2, \dots, n.$$

The α_h , $h = 1, 2, \dots, p$, are fixed autoregressive coefficients and Y_t is white noise, representing that part of Z_t that is not linearly dependent on past observations. The structure of Y_t is such that

$$\text{E}[Y_t] = 0 \quad \text{and} \quad \text{E}[Y_t Y_{t-h}] = \begin{cases} \sigma_Y^2 & \text{if } h = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Choosing σ_Y^2 appropriately ensures that each Z_t is marginally standard normal and the autocorrelation structure of the base process, $\rho_Z(h)$, $h = 1, 2, \dots, p$, is uniquely determined by the autoregressive coefficients α_h , $h = 1, 2, \dots, p$.

Unfortunately, the autocorrelations of the input process X_t are not the same as the autocorrelations of the base process Z_t , so the dependence in the Z_t process must be adjusted to yield the desired dependence in the X_t process. Cario and Nelson (1996) have shown that the lag- h input autocorrelation $\rho_X(h)$ of the time-series input process $\{X_t; t = 1, 2, \dots, n\}$ is a continuous, nondecreasing function of the lag- h autocorrelation $\rho_Z(h)$ of the base process. Therefore, if the desired input autocorrelations are known, then the problem decomposes into p independent correlation-matching problems of determining the value $\rho_Z(h)$, $h = 1, 2, \dots, p$, that map into the desired autocorrelations $\rho_X(h)$, $h = 1, 2, \dots, p$.

Given the input lag- h autocorrelation $\rho_X(h)$, $h = 1, 2, \dots, p$, each correlation-matching problem corresponds to solving the equation

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

for $\rho_Z(h)$, $h = 1, 2, \dots, p$, where $\vartheta_{\rho_Z(h)}$ is the standard bivariate normal probability density function with correlation $\rho_Z(h)$. Unfortunately, it is not possible to solve the correlation-matching problems analytically except in special cases (e.g., Li and Hammond 1975), but there exist efficient numerical methods to solve these problems (see Song, Hsiao, and Chen 1996, Cario and Nelson 1998, and Chen 2001). Finally, variate generation is accomplished by generating a stationary, standard $\text{AR}(p)$

process $\{Z_t; t = 1, 2, \dots, n\}$ by any method and applying the equation $X_t = F_X^{-1}(U_t) = F_X^{-1}[\Phi(Z_t)]$ for $t = 1, 2, \dots, n$.

To summarize, the development of an ARTA process in the case where both F_X and $\rho_X(h)$, $h = 1, 2, \dots, p$, are given becomes solving p correlation-matching problems, for which a number of researchers have suggested computationally feasible methods. The problem that has not been addressed is estimating the parameters of an ARTA process when only raw data produced by an unknown process are available. We propose a way of solving this problem in the remainder of the paper.

3 FITTING ARTA MODELS

In this section, we develop the theory to determine “optimal” fits of dependent time series, present the data-fitting algorithm, and prove the consistency properties of the resulting estimators.

3.1 The Data-Fitting Model

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 to characterize the marginal distribution of the input process. A robust method for fitting target distributions from the Johnson 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-type marginals. We believe that similar techniques can be effectively adapted to fitting ARTA models to dependent univariate data. We outline our approach below.

Let $\{X_t; t = 1, 2, \dots, n\}$ denote a stationary univariate time-series input process. The goal is to approximate $\{X_t; t = 1, 2, \dots, n\}$ by an ARTA process whose complete specification is given by

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

where

$$Z_t = \sum_{h=1}^p \alpha_h Z_{t-h} + Y_t,$$

with Y_t , $t = p + 1, p + 2, \dots, n$, independent and identically distributed Gaussian random variables with mean zero and variance σ_Y^2 . We force the base process Z_t to have variance 1 by choosing $\sigma_Y^2 = 1 - \sum_{h=1}^p \alpha_h \rho_Z(h)$. This closed-form expression for variance σ_Y^2 is given by the Yule-Walker equation corresponding to lag 0. If we know the autoregressive coefficients $\alpha_1, \alpha_2, \dots, \alpha_p$, then we can obtain the autocorrelations $\rho_Z(1), \rho_Z(2), \dots, \rho_Z(p)$ by solving the Yule-Walker equations corresponding to lags $1, 2, \dots, p$. Therefore, the value of σ_Y^2 , which is required to force Z_t to have variance 1, is completely determined by $\alpha_1, \alpha_2, \dots, \alpha_p$; we will write $\sigma_Y \equiv g(p, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_p)$, and no longer consider σ_Y^2 as a parameter to be estimated. And from here on when we say ARTA process, we will mean an ARTA process with Johnson-type marginals.

Clearly, fitting an ARTA process to data corresponds to the estimation of f , γ , δ , λ , ξ , p , and α_h for $h = 1, 2, \dots, p$. For ease of presentation of our data-fitting algorithm, we assume that the order of the underlying base process p and the type of the Johnson transformation f are known. Clearly, these also need to be determined in general. We address this issue in Section 4.

Let $\boldsymbol{\psi}$ correspond to the vector of ARTA parameters, i.e., $\boldsymbol{\psi} = (\lambda, \delta, \gamma, \xi, \alpha_1, \alpha_2, \dots, \alpha_p)$, and consider the standardized white noise process

$$V_t(\boldsymbol{\psi}) = \frac{Y_t}{g(p, \boldsymbol{\alpha})} = \frac{Z_t - \sum_{h=1}^p \alpha_h Z_{t-h}}{g(p, \boldsymbol{\alpha})}.$$

If we further write the base random variable Z_t as a function of the input random variable X_t using (1), then we get the following expression for the standardized white-noise process:

$$V_t(\boldsymbol{\psi}) = \frac{\gamma + \delta f\left[\frac{X_t - \xi}{\lambda}\right] - \sum_{h=1}^p \alpha_h \left(\gamma + \delta f\left[\frac{X_{t-h} - \xi}{\lambda}\right]\right)}{g(p, \boldsymbol{\alpha})}. \quad (2)$$

Now, suppose X_t is actually an ARTA process with the parameter vector $\boldsymbol{\psi}^*$. If we have all of the parameter values correct, i.e., $\boldsymbol{\psi} = \boldsymbol{\psi}^*$, then $V_t(\boldsymbol{\psi}^*)$, $t = p + 1, p + 2, \dots, n$, are independent and identically distributed standard normal random variables. Thus, the fitting procedure we propose searches for parameters that make $V_t(\boldsymbol{\psi})$, $t = p + 1, p + 2, \dots, n$, appear to be such a sample.

Let $V_{(p+1)}(\boldsymbol{\psi}) \leq V_{(p+2)}(\boldsymbol{\psi}) \leq \dots \leq V_{(n)}(\boldsymbol{\psi})$ denote the order statistics corresponding to the random variables $V_t(\boldsymbol{\psi})$, $t = p+1, p+2, \dots, n$. If $\boldsymbol{\psi} = \boldsymbol{\psi}^*$, then the transformed variate $R_{(t)}(\boldsymbol{\psi}^*) = \Phi \left\{ V_{(t)}(\boldsymbol{\psi}^*) \right\}$ has the distribution of the t^{th} order statistic in a random sample of size $n-p$ from the uniform distribution on the unit interval $(0, 1)$. Since $R_{(t)}(\boldsymbol{\psi}^*)$ has mean $\rho_t = (t-p)/(n-p+1)$ (Kendall and Stuart 1979), we can write $R_{(t)}(\boldsymbol{\psi}^*) = \rho_t + \varepsilon_t(\boldsymbol{\psi}^*)$ so that the $\{\varepsilon_t(\boldsymbol{\psi}^*); t = p+1, p+2, \dots, n\}$ are translated uniform order statistics with mean zero and covariance

$$\text{Cov}(\varepsilon_j(\boldsymbol{\psi}^*), \varepsilon_k(\boldsymbol{\psi}^*)) = \frac{\rho_j(1-\rho_k)}{n-p+2}, \quad p+1 \leq j \leq k \leq n. \quad (3)$$

Let $\mathbf{R}_o(\boldsymbol{\psi}) \equiv (R_{(p+1)}(\boldsymbol{\psi}), R_{(p+2)}(\boldsymbol{\psi}), \dots, R_{(n)}(\boldsymbol{\psi}))'$, $\boldsymbol{\rho} \equiv (\rho_{p+1}, \rho_{p+2}, \dots, \rho_n)'$, and $\boldsymbol{\varepsilon}(\boldsymbol{\psi}) \equiv (\varepsilon_{p+1}(\boldsymbol{\psi}), \varepsilon_{p+2}(\boldsymbol{\psi}), \dots, \varepsilon_n(\boldsymbol{\psi}))'$, so that $\boldsymbol{\varepsilon}(\boldsymbol{\psi}^*) \equiv \mathbf{R}_o(\boldsymbol{\psi}^*) - \boldsymbol{\rho}$. Since the first and second moments of the uniformized order statistics are known and easily computed, we exploit this fact to develop a single, distribution-free formulation of the fitting problem. Specifically, we minimize the distance between $\boldsymbol{\rho}$ and $\mathbf{R}_o(\boldsymbol{\psi})$ as a function of $\boldsymbol{\psi}$ with respect to some metric defined by a quadratic form in the $(n-p)$ -dimensional Euclidean space. If \mathbf{W} is the $(n-p) \times (n-p)$ matrix associated with this quadratic form, then the parameter estimates can be obtained via least-squares fitting given by

$$\begin{aligned} \min_{\boldsymbol{\psi}} \quad & S_{\mathbf{W}}(\boldsymbol{\psi}) \equiv (\mathbf{R}_o(\boldsymbol{\psi}) - \boldsymbol{\rho})' \mathbf{W} (\mathbf{R}_o(\boldsymbol{\psi}) - \boldsymbol{\rho}) \equiv \boldsymbol{\varepsilon}(\boldsymbol{\psi})' \mathbf{W} \boldsymbol{\varepsilon}(\boldsymbol{\psi}) \\ \text{subject to} \quad & \boldsymbol{\psi} \in \boldsymbol{\Psi}. \end{aligned} \quad (4)$$

We define the feasible region Ψ as follows:

$$\Psi = \{(\gamma, \delta, \lambda, \xi, \alpha_1, \alpha_2, \dots, \alpha_p)' : \begin{cases} \delta & \begin{cases} > 0 & \text{for } f = S_U, f = S_B, f = S_L, \text{ and } f = S_N, \\ < \infty & \text{for } f = S_U \text{ and } f = S_B, \end{cases} \\ \lambda & \begin{cases} > 0 & \text{for } f = S_U, \\ > X_{(n)} - \xi & \text{for } f = S_B, \\ = 1 & \text{for } f = S_L \text{ and } f = S_N, \end{cases} \\ \xi & \begin{cases} < X_{(1)} & \text{for } f = S_L \text{ and } f = S_B, \\ = 0 & \text{for } f = S_N, \end{cases} \\ \left| \text{RootOf} \left(1 - \sum_{h=1}^p \alpha_h B^h = 0, B \right) \right| & > 1, \end{cases} \quad (5)$$

where the function ‘‘RootOf’’ is a place holder for representing all the roots of the equation $1 - \sum_{h=1}^p \alpha_h B^h = 0$ in the variable B . The first three of the constraints (5) ensure the feasibility of the Johnson parameters depending on the family of interest, and the last constraint ensures the stationarity of the autoregressive base process, and hence the stationarity of the input process.

The least-squares fitting problem gives rise to different estimators depending on the form of the weight matrix \mathbf{W} . When the weight matrix \mathbf{W} is the $(n-p) \times (n-p)$ identity matrix \mathbf{I} , we obtain the ordinary least-squares (OLS) estimators for $\boldsymbol{\psi}$; and when $\mathbf{W} \neq \mathbf{I}$, we obtain weighted least-squares (WLS) parameter estimators. WLS parameter estimators are of interest since $\{\varepsilon_t(\boldsymbol{\psi}^*); t = p+1, p+2, \dots, n\}$ are neither independent nor homoscedastic. In addition, (3) shows that the matrix $\mathbf{W} = [\text{Cov}(\varepsilon_j(\boldsymbol{\psi}^*), \varepsilon_k(\boldsymbol{\psi}^*))]_{(n-p) \times (n-p)}^{-1}$ is readily computed. For the estimation of a linear model, it is well known that such a weight matrix yields the minimum variance linear unbiased estimator of the vector of model parameters (Seber 1977); and this strongly suggests that we should also take $\mathbf{W} = [\text{Cov}(\varepsilon_j(\boldsymbol{\psi}^*), \varepsilon_k(\boldsymbol{\psi}^*))]_{(n-p) \times (n-p)}^{-1}$ while estimating the nonlinear model (4). However, we choose to use the diagonal weight matrix, $\mathbf{W} = \mathbf{D}$, defined as

$$\mathbf{D} = \text{diag} \{1/\text{Var}(\varepsilon_{p+1}(\boldsymbol{\psi}^*)), 1/\text{Var}(\varepsilon_{p+2}(\boldsymbol{\psi}^*)), \dots, 1/\text{Var}(\varepsilon_n(\boldsymbol{\psi}^*))\}, \quad (6)$$

giving us the diagonally-weighted least-squares (DWLS) parameter estimators. In this type of problem, fitting with DWLS is usually superior to WLS based on the experience of Swain, Venkatraman, and Wilson (1988) and Kuhl and Wilson (1999).

In expanded form, the objective function of the DWLS estimation problem, using (4) and (6), can be written as

$$\begin{aligned} S_{\mathbf{D}}(\boldsymbol{\psi}) &= \boldsymbol{\varepsilon}(\boldsymbol{\psi})' \mathbf{D} \boldsymbol{\varepsilon}(\boldsymbol{\psi}) \\ &= \frac{1}{(n-p)^2} \sum_{t=p+1}^n \frac{(n-p+1)^2(n-p+2)}{(t-p)(n+1-t)} \left(\Phi \{V_{(t)}(\boldsymbol{\psi})\} - \frac{t-p}{n-p+1} \right)^2, \end{aligned} \quad (7)$$

where $\{V_t(\boldsymbol{\psi}); t = p+1, p+2, \dots, n\}$ is given by (2). Notice that the use of the uniformized order statistics for fitting permits a single formulation for not only Johnson-type distributions, but all continuous distributions, because the necessary first and second moments of $\{\Phi\{V_{(t)}(\boldsymbol{\psi}^*)\}; t = p+1, p+2, \dots, n\}$, are known and easily computed.

Next, we present our data-fitting algorithm together with the statistical properties of the resulting estimators.

3.2 The Data-Fitting Algorithm

We minimize the objective function (7) subject to the constraints in (5) by using a general-purpose optimization algorithm. Unfortunately, many of these algorithms are dependent upon good initial estimates of the parameters. Further, the number of model parameters we need to estimate is $p+4$, which increases linearly with the order of dependence p , making it even less likely that we can obtain robust estimates that are independent of the quality of the initial solution as p gets larger. Fortunately, there is a natural decomposition of our optimization problem between determining the Johnson parameters $(\gamma, \delta, \lambda, \xi)$ and the base-process parameters $(\alpha_1, \alpha_2, \dots, \alpha_p)$. We have also empirically observed that solving $S_{\mathbf{D}}(\boldsymbol{\psi})$ for any given fixed feasible $\gamma, \delta, \lambda, \xi$ provides pretty robust estimates of $\alpha_1, \alpha_2, \dots, \alpha_p$. Therefore, we work iteratively between improving the estimates for $(\gamma, \delta, \lambda, \xi)$ and $(\alpha_1, \alpha_2, \dots, \alpha_p)$.

Before we give a complete statement of our data-fitting algorithm, we introduce the notation we will use in its presentation.

3.2.1 Notation

We define the feasible region $\mathbf{cl}\Psi$ by closing the open convex set Ψ defined in (5):

$$\mathbf{cl}\Psi = \{(\gamma, \delta, \lambda, \xi, \alpha_1, \alpha_2, \dots, \alpha_p)' : \begin{cases} \delta \begin{cases} \geq 0 & \text{for } f = S_U, f = S_B, f = S_L, \text{ and } f = S_N, \\ \leq \infty & \text{for } f = S_U \text{ and } f = S_B, \end{cases} \\ \lambda \begin{cases} \geq 0 & \text{for } f = S_U, \\ \geq X_{(n)} - \xi & \text{for } f = S_B, \\ = 1 & \text{for } f = S_L \text{ and } f = S_N. \end{cases} \\ \xi \begin{cases} \leq X_{(1)} & \text{for } f = S_L \text{ and } f = S_B, \\ = 0 & \text{for } f = S_N. \end{cases} \\ \left| \text{RootOf} \left(1 - \sum_{h=1}^p \alpha_h B^h = 0, B \right) \right| \geq 1 \end{cases} \quad (8)$$

Such a modification in the definition of the feasible region enables us to use Theorem 7.3.4 of Bazaraa, Sherali, and Shetty (1993) in order to prove the convergence of the data-fitting algorithm (see Theorem 1). Fortunately, this modification does not change the outcome of the algorithm, because we show that there are no local minimums of the objective function $S_{\mathbf{D}}(\psi)$ on the boundary of the feasible region $\mathbf{cl}\Psi$ (see Appendix, Proposition 1).

Let $\mathbf{x} = (x_1, x_2, \dots, x_n)'$ denote a vector of sample data in which ties occur with probability zero and define $S_{\mathbf{D}}(\psi|\mathbf{x})$ as the objective function dependent on the given sample. Let $\mathbf{C} : \mathbf{cl}\Psi \rightarrow \mathbf{cl}\Psi$ and $\mathbf{D} : \mathbf{cl}\Psi \rightarrow \mathbf{cl}\Psi$ be point-to-point maps given by

$$\begin{aligned} \mathbf{C}(\psi) \equiv & \operatorname{argmin}_{\gamma, \delta, \lambda, \xi} S_{\mathbf{D}}(\psi|\mathbf{x}) \\ & \text{subject to } \psi \in \Psi_{\mathbf{C}} \end{aligned} \quad (9)$$

and

$$\begin{aligned} \mathbf{D}(\psi) \equiv & \operatorname{argmin}_{\alpha_1, \alpha_2, \dots, \alpha_p} S_{\mathbf{D}}(\psi|\mathbf{x}) \\ & \text{subject to } \psi \in \Psi_{\mathbf{D}}, \end{aligned} \quad (10)$$

where $\Psi_{\mathbf{C}}$ and $\Psi_{\mathbf{D}}$ correspond to the constraints ensuring the feasibility of the Johnson parameters and the stationarity of the underlying base process, respectively, and satisfy $\mathbf{cl}\Psi = \Psi_{\mathbf{C}} \cup \Psi_{\mathbf{D}}$.

Finally, we define the solution set $\Omega(\mathbf{x}) \equiv \{\bar{\boldsymbol{\psi}} : \nabla_{\boldsymbol{\psi}} S_{\mathbf{D}}(\bar{\boldsymbol{\psi}}|\mathbf{x}) = \mathbf{0}\}$, corresponding to the collection of parameters at which all of the entries of the gradient of the objective function $S_{\mathbf{D}}(\boldsymbol{\psi}|\mathbf{x})$ attain the value of zero.

3.2.2 Statement of the Algorithm

Initialization Step

Let $k = 1$ and $\boldsymbol{\psi}_0$ be a starting parameter vector in the interior of $\mathbf{cl}\Psi$.

Main Step

1. Let $\boldsymbol{\psi}_k \in \mathbf{C}(\boldsymbol{\psi}_{k-1})$. Replace k by $k + 1$ and go to Step 2.
2. Let $\boldsymbol{\psi}_k \in \mathbf{D}(\boldsymbol{\psi}_{k-1})$. If $\boldsymbol{\psi}_k \in \Omega(\mathbf{x})$, then stop; otherwise, replace k by $k + 1$ and repeat Step 1.

Starting with a parameter vector in the interior of $\mathbf{cl}\Psi$, we first solve the least-squares fitting problem for the Johnson parameters $\gamma, \delta, \lambda, \xi$ by keeping the base process parameters $\alpha_1, \alpha_2, \dots, \alpha_p$ fixed. We call this Stage 1. Then, we solve the least-squares fitting problem for $\alpha_1, \alpha_2, \dots, \alpha_p$ by keeping $\gamma, \delta, \lambda, \xi$ fixed and we call this Stage 2.

Until the data-fitting algorithm reaches a point in the solution set $\Omega(\mathbf{x})$, we work iteratively between Stage 1 and Stage 2, converging to a stationary point.

Theorem 1 *Given a starting parameter vector $\boldsymbol{\psi}_0$ in the interior of $\mathbf{cl}\Psi$ and using the Levenberg-Marquardt algorithm to carry out Steps 1 and 2, the data-fitting algorithm either stops in a finite number of steps at a point in $\Omega(\mathbf{x})$ or generates an infinite sequence $\{\boldsymbol{\psi}_k\}$ such that all of its accumulation points belong to $\Omega(\mathbf{x})$.*

Proof. See the Appendix.

We terminate the data-fitting algorithm when we reach a point in the solution set $\Omega(\mathbf{x})$. In most cases, however, convergence to a point in the solution set occurs only in a limiting sense

and we must resort to some practical rules for terminating the iterative procedure. We terminate the algorithm when either the absolute or relative requested error tolerance has been attained, i.e., $|S_{\mathbf{D}}(\boldsymbol{\psi}_k|\mathbf{x}) - S_{\mathbf{D}}(\boldsymbol{\psi}_{k-1}|\mathbf{x})| \leq \text{AbsoluteErrorRequest}$ or $|S_{\mathbf{D}}(\boldsymbol{\psi}_k|\mathbf{x}) - S_{\mathbf{D}}(\boldsymbol{\psi}_{k-1}|\mathbf{x})| \leq S_{\mathbf{D}}(\boldsymbol{\psi}_{k-1}|\mathbf{x}) \times \text{RelativeErrorRequest}$. If either is attained, the algorithm stops, reporting success. We can force that one or the other of these criteria to be satisfied by specifying the requested error for the other as zero.

Notice that the parameters given by the data-fitting algorithm do not necessarily correspond to a local minimum solution of the problem (Bazaraa, Sherali, and Shetty 1993). However, if the starting parameters fall in a convex subregion that includes any local solution of the problem, then the data-fitting algorithm converges to a local minimum solution, because the objective function $S_{\mathbf{D}}(\boldsymbol{\psi}|\mathbf{x})$ is convex around any unconstrained local minimum of the problem (see Appendix, Corollary 1). Thus, using a general-purpose optimization algorithm with local convergence properties will ensure that we reach a local optimal point when we start in its convex region. Later, in Section 4, we describe the numerical methods used to implement the data-fitting algorithm.

We note that our approach bears some similarity to the forecasting technique of Block, Langberg, and Stoffer (1990). They also consider the observed data to be a transformation, via the inverse cdf, of an underlying Gaussian process. They propose using this transformation to provide a joint distribution for the observed data, but, unlike us, they solve for the unknown parameters of the marginal distribution and the underlying Gaussian process simultaneously via maximum likelihood estimation. However, the resulting likelihood function appears difficult to maximize except for the simplest models, and they provide no properties for the resulting estimators. The statistical properties of our estimators are established in the next section.

3.3 Properties of the ARTA Estimators

Suppose that X_1, X_2, \dots, X_n are identically distributed random variables with a joint ARTA distribution characterized by the parameter vector $\boldsymbol{\psi}^*$. Even if we assume that the type of the Johnson transformation f and the order of dependence p are known, the DWLS estimators $\hat{\boldsymbol{\psi}}_n$ are not necessarily consistent. Next, we explain why the ARTA estimators are not consistent: For the consistency of the ARTA estimators to hold, we need only that the empirical distribution of $R_t(\boldsymbol{\psi}) = \Phi\{V_t(\boldsymbol{\psi})\}$, $t = p + 1, p + 2, \dots, n$, converges to the uniform distribution on the unit

interval $(0,1)$ in the limit, i.e., $n \rightarrow \infty$. By construction, $R_t(\boldsymbol{\psi}^*)$, $t = p+1, p+2, \dots, n$, have i.i.d. uniform marginals. We also show that $R_t(\boldsymbol{\psi})$, $t = p+1, p+2, \dots, n$, are *i.i.d.* uniform random variables on the unit interval $(0,1)$ only at $\boldsymbol{\psi} = \boldsymbol{\psi}^*$ (see Appendix, Theorem 3). However, $R_t(\boldsymbol{\psi})$, $t = p+1, p+2, \dots, n$, can have uniform marginals at parameter settings other than $\boldsymbol{\psi}^*$. For instance, at $\gamma = \gamma^*, \delta = \delta^*, \lambda = \lambda^*, \xi = \xi^*$, and $\alpha_h = 0$ for $h = 1, 2, \dots, p$, although $R_t(\boldsymbol{\psi})$, $t = p+1, p+2, \dots, n$ are not independently distributed, they are uniform on the unit interval $(0,1)$ by the definition of the Johnson-type random variable. Since our goal is data modeling, rather than recovery of the true distribution, consistency is not a critical property. Nevertheless, it is desirable. Although the estimator $\hat{\boldsymbol{\psi}}_n$ is not consistent in general, it does have certain limited consistency properties that are of interest. These are summarized in the next theorem:

Theorem 2 *Let X_1, X_2, \dots, X_n be identically distributed random variables with a joint ARTA distribution characterized by the parameter vector $\boldsymbol{\psi}^*$ and assume that the type of the Johnson transformation f and the order of dependence p are known. We solve the diagonally weighted least squares problem given by (7) and obtain $\hat{\boldsymbol{\psi}}_n$ as the vector of estimates for a sample of size n . Then as $n \rightarrow \infty$, the following properties hold:*

1. $\Pr [\hat{\lambda}_n \rightarrow \lambda^*] = 1$ and $\Pr [\hat{\xi}_n \rightarrow \xi^*] = 1$.
2. If $\alpha_h = \alpha_h^*$, $h = 1, 2, \dots, p$, then $\Pr [\hat{\boldsymbol{\psi}}_n \rightarrow \boldsymbol{\psi}^*] = 1$.

Proof. See the Appendix.

The first result in Theorem 2 is of limited practical value. However, the second result is helpful in two ways: 1) We proposed decomposing the algorithm for solving the least-squares problem in two steps — improving the estimates of $(\gamma, \delta, \lambda, \xi)$ by keeping $(\alpha_1, \alpha_2, \dots, \alpha_p)$ fixed and improving the estimates of $(\alpha_1, \alpha_2, \dots, \alpha_p)$ by keeping $(\gamma, \delta, \lambda, \xi)$ fixed — because we observed that the estimates of $(\alpha_1, \alpha_2, \dots, \alpha_p)$ are robust to poor estimates of $(\gamma, \delta, \lambda, \xi)$. Theorem 2 shows that when we get the base process parameters right, the least-squares estimators of the remaining Johnson parameters are strongly consistent. 2) Notice also that when $p = 0$, the first stage of the data-fitting algorithm reduces to the one suggested by Swain, Venkatraman, and Wilson (1988) and it performs least-squares fitting by treating the given sample points as independent. If $\{X_t; t = 1, 2, \dots, n\}$ were i.i.d. Johnson-type random variables with the parameter set

$\boldsymbol{\psi}^* = (\gamma^*, \delta^*, \lambda^*, \xi^*)$, then the model with $p = 0$ would be correct and the transformed random variate $R_{(t)}(\boldsymbol{\psi}^*) = \Phi \left\{ \gamma^* + \delta^* f[(X_{(t)} - \xi^*)/\lambda^*] \right\}$ would have the distribution of the t^{th} uniform order statistic on the unit interval $(0, 1)$. Although Swain, Venkatraman, and Wilson (1988) show empirically that the suggested least-squares estimators provide a convenient computational method for fitting any member of the Johnson system when $p = 0$, they do not present any statistical properties of these estimators. Theorem 2 indicates that the fitting procedure of Swain, Venkatraman, and Wilson gives strongly consistent estimators of the Johnson parameters.

Following from the second result, a natural question to ask is whether the following property holds:

If $\hat{f}_n \rightarrow f^*$, $\hat{\delta}_n \rightarrow \delta^*$, $\hat{\gamma}_n \rightarrow \gamma^*$, $\hat{\lambda}_n \rightarrow \lambda^*$, and $\hat{\xi}_n \rightarrow \xi^*$, then $\Pr[\hat{\alpha}_h \rightarrow \alpha_h^*] = 1$ for $h = 1, 2, \dots, p$.

It is straightforward to establish the result for $p = 1$, but it cannot be extended to higher orders of dependence. However, if we had modified the second stage of the data-fitting algorithm in such a way that it would find the parameters of a Gaussian AR(p) process, then the property would hold because strongly consistent estimators of the parameters of Gaussian AR(p) processes are well known. Our motivation for not taking this approach, but instead using the formulation (7), is to characterize the joint estimation of $(\gamma, \delta, \lambda, \xi, \alpha_1, \alpha_2, \dots, \alpha_p)$ by a single objective that does not favor either the Johnson or the base process parameters, while leading to a direct proof of the convergence of the numerical algorithm. We consider alternative formulations that lead to strongly consistent estimators of the entire vector of parameters $\boldsymbol{\psi}^*$ as a subject of future research.

3.4 Extension to Other Continuous Distributions

Although the development in this paper and the results above are based on the Johnson translation system, the standardized white noise process can be written as

$$V_t(\boldsymbol{\psi}) = \frac{\Phi^{-1}\{F_X(X_t)\} - \sum_{h=1}^p \alpha_h \Phi^{-1}\{F_X(X_{t-h})\}}{g(p, \boldsymbol{\alpha})}, \quad t = p+1, p+2, \dots, n,$$

for any continuous marginal distribution F_X . Therefore, our approach is not limited to Johnson-type distributions and could be extended to any continuous F_X with the vector of parameters $\boldsymbol{\psi}$, defined in a feasible region $\boldsymbol{\Psi}$, with the following properties:

P(1). If the boundary of $\mathbf{cl}\Psi$ does not include any local minimum solutions, then every possible solution ψ to the optimization problem $\min_{\psi} S_{\mathbf{D}}(\psi|\mathbf{x})$ lies in the interior of $\mathbf{cl}\Psi$.

P(2). The function $R_t(\psi)$ is a three-times continuously differentiable function for every $\psi \in \Psi$.

If all of the items in P(1)-P(2) are satisfied for the choice of the functional form of the target cdf, then the convergence of the data-fitting algorithm and its limited consistency properties remain valid.

4 IMPLEMENTATION

We have developed a stand-alone, PC-based program that implements the suggested algorithm for fitting stochastic input models to raw data. The key computational components of the software are written in portable C++ code and are available at www.iems.northwestern.edu/~nelsonb/ARTAFIT/artafit.html. In this section, we discuss the numerical methods used in the implementation. The issues of interest are the determination of the starting values for the parameters of the Johnson-type marginal distribution and the autoregressive base process, the optimization algorithms used to solve the maps \mathbf{C} (9) and \mathbf{D} (10), the assurance of the stationarity of the input process, and the positive definiteness of the autocorrelation structure of the base process.

4.1 Selection of the Johnson Transformation f and Initial Values of $\gamma, \delta, \lambda, \xi$

A common procedure for identifying the type of transformation to use from the Johnson translation system is to compute the sample skewness and kurtosis, and then pick the family associated with that point on the (skewness, kurtosis) plane. Algorithm AS 99 developed by Hill, Hill, and Holder (1976) does this, for instance. We have observed that, for large enough sample sizes, this procedure identifies the true Johnson transformation f when the autocorrelations are relatively weak. However, as the strength of the autocorrelations increases, we observe significant bias and variability in the higher sample moments, increasing the likelihood of identifying the wrong transformation. Therefore, our approach is to fit *all* of the families and compare the goodness of the fits.

After the type of the Johnson transformation is identified, we enter the sample mean, variance, skewness, and kurtosis as input into Algorithm AS 99. As output, we get such $\gamma, \delta, \lambda,$ and ξ that the distance between the first four moments of the suggested Johnson distribution and the sample data

is minimized. We finish the initialization step of the fitting procedure by implementing a feasibility check on the Johnson parameters. Unfortunately, the feasibility check might fail in the following two cases: 1) When the type of the Johnson transformation is lognormal, the location parameter might be equal to or larger than the smallest observation of the sample data, i.e., $x_{(1)} \leq \xi$. 2) When the type of the Johnson transformation is bounded, either the location parameter might be equal to or larger than the smallest observation of the sample data, i.e., $x_{(1)} \leq \xi$, or the summation of the location and scale parameters might be equal to or smaller than the largest observation of the sample data, i.e., $x_{(n)} \geq \lambda + \xi$. If either of these cases takes place, then we modify the Johnson parameters to be feasible for the prespecified Johnson transformation and the given sample data. To do this, we implement feasibility-constrained moment matching with the Nelder-Mead algorithm (Olsson 1974). We first modify λ and ξ to find the location and scale parameters that are feasible for the sample data and then, for the modified values of λ and ξ , we find such γ and δ that the squared distance between the third and fourth moments of the Johnson marginal and the sample data is minimized by the Nelder-Mead algorithm. Finally, we modify λ and ξ so that they both match the sample mean and sample variance as closely as possible and pass the feasibility check. This routine is very similar to the one implemented in the FITTR1 software (Swain, Venkatraman, and Wilson 1988).

4.2 Selection of the Order of the Underlying Base Process

Although there is no upper bound on the order of the underlying autoregressive process p , in practice we expect p to be less than or equal to 5. For example, Wei (1990) states that usually the order of dependence is less than or equal to 3. Therefore, we could relax the assumption that p is known through complete enumeration. However, there is a well developed literature on AR order selection that has been shown to choose the correct order with probability one as the sample size goes to infinity. We refer the reader to Chatfield (1999) for an extensive review. We chose to use the Schwarz criterion as it is asymptotically consistent and has been quite popular in recent applied work.

Before the execution of the ARTA fitting algorithm, we get initial estimates for the Johnson parameters, $\hat{\gamma}$, $\hat{\delta}$, $\hat{\lambda}$, $\hat{\xi}$, as described in Section 4.1. Using $\hat{\gamma}$, $\hat{\delta}$, $\hat{\lambda}$, and $\hat{\xi}$, we transform the input data x_1, x_2, \dots, x_n to $\hat{z}_t = \hat{\xi} + \hat{\lambda}f\left[(x_t - \hat{\gamma})/\hat{\delta}\right]$, $t = 1, 2, \dots, n$. By treating the transformed data \hat{z}_t as a

sample of length n from a Gaussian $\text{AR}(p)$ process, we fit an autoregressive model by least-squares estimation. We take the resulting estimates as the starting parameter vector for the underlying base process. Unfortunately, getting good starting solutions for the Johnson distribution is not as easy as for the base process due to the bias involved in the estimation of sample third and fourth order moments. However, the robustness of the base process parameters allows us to improve the marginal distribution parameters pretty quickly in the upcoming iterations.

Application of the Schwarz criterion to the transformed data \hat{z}_t , $t = 1, 2, \dots, n$, gives us the order of dependence p that we keep constant throughout the execution of the algorithm. When we decide to terminate the algorithm, as described in Section 3.2, we update the transformed data \hat{z}_t , $t = 1, 2, \dots, n$, using the most recent estimates of the Johnson parameters and, similarly, determine the order of dependence implied by the Schwarz criterion. If we find the same order of dependence, then we stop the algorithm. Otherwise, we rerun the algorithm using the most recent underlying base process and the new order estimate.

4.3 Optimization Algorithms

The objective function (7) can be minimized subject to the constraints in (8) by using a general-purpose optimization algorithm. We choose to perform the DWLS estimation using a Levenberg-Marquardt (LM) optimization algorithm (Marquardt 1963), whose many implementations have proven to be very successful in practice. In addition, the convergence properties of the LM algorithm ensure the convergence of our data-fitting algorithm. However, if the termination criterion has not been satisfied in a prespecified number of iterations, then, for practical purposes, we resort to the Nelder-Mead algorithm. On a suite of typical least-squares test problems, Kuhl and Wilson (1999) found that the Nelder-Mead algorithm is faster than the LM algorithm while yielding solutions with virtually the same accuracy. These considerations motivate the use of the Nelder-Mead algorithm as a practical addition to the implementation of the DWLS estimation procedure. Despite the lack of a convergence proof, we find the Nelder-Mead algorithm very useful in the implementation of our data-fitting algorithm.

4.4 Stationarity and Positive Definiteness of the Autocorrelation Structure

Our data-fitting algorithm approximates the input process by a stationary ARTA model with a positive definite base autocorrelation matrix. It maintains stationarity by enforcing the constraint $\left| \text{RootOf} \left(1 - \sum_{h=1}^p \alpha_h B^h = 0, B \right) \right| \geq 1$. To do this, it assigns a very large value, e.g., 10^{20} , to the objective function when the roots of the reverse characteristic polynomial, $1 - \sum_{h=1}^p \alpha_h B^h = 0$, falls in or onto the complex unit circle. This eliminates the consideration of unstable base processes, fitting a stationary ARTA model to any sample data. Further, the autocorrelation matrix of the fitted base process is always positive definite, because the autocovariance function of a covariance stationary sequence with an autoregressive representation is positive definite (Fishman 1973).

5 ILLUSTRATIVE EXAMPLES

In this section, we illustrate the use of the suggested data-fitting algorithm with three different examples. The data used in the first example are generated artificially from a true ARTA process, while the other two data sets come from real processes representing a pressure variable of a continuous-flow production line and sales of a large vehicle manufacturer. A comprehensive empirical comparison/analysis on our data-fitting algorithm is provided in Biller (2002).

5.1 Source of Data: A Known ARTA Process

In this section, we test the performance of the data-fitting algorithm against data that come from a stationary ARTA process with Johnson unbounded (S_U) marginal distribution having parameters $(\gamma, \delta, \lambda, \xi) = (-0.54, 1.54, 1.14, -0.51)$. The factors considered in the experiments are the sample size n and the order of dependence p . The goal is to see how well the data-fitting algorithm recovers the true parameters of the process. The results of this section are representative of a larger study in Biller (2002).

The experimental results are reported in Table 1 for different autocorrelation structures that are $\rho_X = 0.35$, $\rho_X = (0.6, 0.2)$, and $\rho_X = (-0.45, 0.20, -0.10)$. The results are the summary statistics whose averages are taken over a minimum of 30 replications and a maximum of 62 replications. We start each experiment by carrying out 30 replications and increase the number of replications whenever necessary to ensure an absolute error of no more than 0.1 on the Kolmogorov-Smirnov

Table 1: Goodness-of-fit test statistics of a selected set of experiments

$\rho_X = 0.35$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$
KS_X	0.282	0.201	0.153	0.040	0.000
AS_{ρ_X} (2.22)	2.401	2.045	2.145	2.224	2.225
KS_{ρ_X}	0.245	0.238	0.241	0.139	0.132
KS^s_X	1.270	0.941	0.641	0.674	0.666
$AS^s_{\rho_X}$	1.185	1.671	1.948	2.195	2.201
$KS^s_{\rho_X}$	0.374	0.164	0.157	0.160	0.148
$\rho_X = (0.6, 0.2)$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$
KS_X	0.388	0.293	0.201	0.110	0.030
AS_{ρ_X} (3.29)	3.721	3.675	3.426	3.302	3.308
KS_{ρ_X}	0.485	0.385	0.213	0.194	0.104
KS^s_X	1.235	1.183	1.059	1.008	1.000
$AS^s_{\rho_X}$	2.547	2.897	2.975	2.978	2.976
$KS^s_{\rho_X}$	0.356	0.397	0.412	0.403	0.401
$\rho_X = (-0.45, 0.20, -0.10)$	$n = 50$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$
KS_X	0.516	0.387	0.246	0.133	0.021
AS_{ρ_X} (5.32)	5.622	5.498	5.304	5.310	5.308
KS_{ρ_X}	0.393	0.321	0.252	0.230	0.241
KS^s_X	1.521	1.756	1.542	1.387	1.306
$AS^s_{\rho_X}$	3.572	3.814	3.648	3.652	3.649
$KS^s_{\rho_X}$	0.405	0.348	0.297	0.268	0.271

test statistic comparing the fitted cdf against the empirical cdf.

Each column of Table 1 shows the results of the goodness-of-fit tests for a different sample size; e.g., the sample size is 50 in the second column. We compare the fitted ARTA model to both the true ARTA model and the empirical model, whose marginal distribution and the autocorrelation structure are given by the empirical cumulative distribution function and the sample autocorrelation function. We distinguish the test statistics corresponding to the comparison with the empirical model using the subscript s .

We evaluate the goodness of the fit of the marginal distribution using the Kolmogorov-Smirnov test statistic KS_X , while we evaluate the goodness of the fit of the autocorrelation structure using the asymptotic variance constant, AS_{ρ_X} , and the comparison of the spectral distribution functions via the Kolmogorov-Smirnov criterion KS_{ρ_X} (Anderson 1993). KS_X measures the maximum absolute difference between the fitted Johnson-type marginal and the true ARTA marginal, while KS^s_X corresponds to the scaled KS test statistic obtained as a result of comparing the fitted Johnson-

type marginal against the empirical cdf. AS_{ρ_X} and $AS_{\rho_X}^s$ are the asymptotic variance constants of the fitted ARTA process and the empirical input process, respectively. The true asymptotic variance constant $AS_{\rho_X}^*$ for each experiment setting is given in parenthesis on the 3rd, 10th, and 17th rows of the first column. Finally, we report the KS criterion comparing the spectral distribution functions of the fitted and true ARTA processes (denoted by KS_{ρ_X}) and the spectral distribution functions of the fitted ARTA process and the empirical process (denoted by $KS_{\rho_X}^s$).

The test statistics KS_X , AS_{ρ_X} , and KS_{ρ_X} , which are written in bold in Table 1, indicate that we approximately recover the true ARTA cdf and the autocorrelation structure as the sample size approaches 5000. Although that we have $KS_X = 0$ and $AS_{\rho_X} = AS_{\rho_X}^*$ does not necessarily imply that $f = f^*$, $\gamma = \gamma^*$, $\delta = \delta^*$, $\lambda = \lambda^*$, $\xi = \xi^*$, and $\rho_X(h) = \rho_X^*(h)$ for $h = 1, 2, \dots, p$, we observe convergence both in the first four moments of the marginals and the individual input autocorrelations in this particular experimental setting. Notice the significant deviation of the fitted ARTA process from the true process when the sample size we experiment with is not as large as 5000. This is not surprising at all, because when the sample size is small, most of the time the sample itself does not contain enough information to identify the true system. However, in all of the cases considered, the ARTA fit has been observed to be successful in capturing most of the characteristics of the sample data, even though the true ARTA process has remained unidentified. For example, see the test statistics KS_X^s , $AS_{\rho_X}^s$, and $KS_{\rho_X}^s$ for different values of n . Despite the limited consistency properties of our ARTA estimators, they provide plausible models for most of the data samples and this can be explained by the impact of the joint distribution of the order statistics in the minimization of the objective function (7) in finite samples.

5.2 Source of Data: Real Processes

It is essential that we experiment with realistic — as opposed to artificially generated — input-modeling problems to stress the proposed methodology, because real problems violate many or all of the assumptions of the simple input models in current use, including the existence of a true, underlying distribution with a simple, independent structure. Therefore, in this section, we approximate two physical processes using our data-fitting algorithm.

5.2.1 Modeling Pressure Measurements on a Continuous-Flow Production Line

Continuous-flow production lines, such as those used to extrude plastics, are common in the chemical industry. Process variables, including temperatures and pressures, are often key parameters of this type of production line and understanding their effects on the manufacturing system is critical because these measurements exhibit strong series dependence. System simulation is sometimes used to model new and existing lines, as well as to train new operators in proper responses to process changes. We therefore use our software to approximate this input process. In particular, we fit 519 data points recorded at fixed time increments on a pressure variable of a continuous-flow production line at a large chemical manufacturing plant.

Cario and Nelson (1998), who previously tried to fit an input model to these data, chose the Weibull marginal distribution function. Since their software ARTAFACTS has no capabilities for fitting marginal distributions, they determined the parameters of the Weibull distribution with the aid of the Arena Input Analyzer (Rockwell Software) that assumes i.i.d. data and uses maximum likelihood estimation. In addition, they approximated the input autocorrelation structure using the estimated autocorrelation function of the raw data and assumed an order of dependence $p = 3$. We call this model “artafact” and denote it by “AF” in Table 2.

Our software fit a Johnson unbounded distribution and an autocorrelation structure with $p = 2$ characterized by $(\gamma, \delta, \lambda, \xi) = (2.046, 3.151, 0.457, 1.217)$ and $(\alpha_1, \alpha_2) = (1.050, -0.342)$. We call this model “artafit,” denote it by “ARF(p)” in Table 2, and use p to denote its order. The quantile-quantile (q-q) plots comparing the empirical pressure data with the artafact and artafit data are given in Figure 1. Observe the difference between the left tails of the distribution functions. Comparing Figure 1a with 1b, it is visually obvious that our fit is not only superior to Cario’s and Nelson’s, but also very close to the probability density function of the physical process (Figure 1b). To substantiate our visual observation, we use goodness-of-fit tests to compare our results to Cario’s and Nelson’s.

In the first two rows of Table 2, we report the scaled Kolmogorov-Smirnov (KS_X^s) and Anderson-Darling (AD_X^s) test statistics comparing the empirical distribution function with the fitted distributions. The second column (AF) corresponds to the fit suggested by Cario and Nelson (1998), the third column (ARF(0)) corresponds to the Johnson fit under the assumption of independence, and the other columns correspond to the Johnson fits under the assumptions of orders of dependence 1,

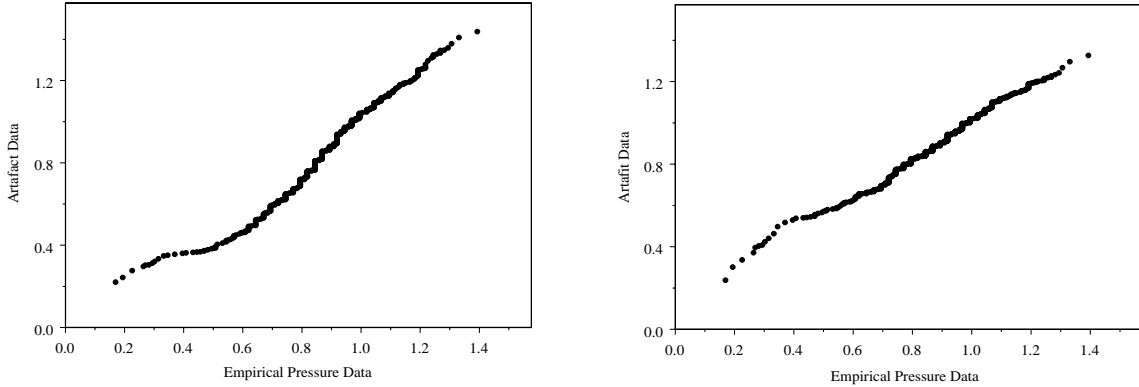


Figure 1: (a) Q-Q Plot Comparing the Empirical Pressure Data and the Artafit Data (b) Q-Q Plot Comparing the Empirical Pressure Data and the Artafit Data

Table 2: Comparison of Kolmogorov-Smirnov, Anderson-Darling, and KS Spectral Test Statistics

	AF	ARF(0)	ARF(1)	ARF(2)	ARF(3)
KS_X^s	1.764	1.038	0.841	0.841	1.538
AD_X^s	3.295	0.758	0.758	0.755	2.563
$KS_{\rho_X}^s$	0.004	0.509	0.204	0.095	0.099

2, and 3, respectively. Comparison of these statistics to the critical values 0.895 and 0.870, respectively, at a significance level of 5%, suggests that the fitted Johnson-type marginal distributions when $p = 1$ and $p = 2$ are statistically superior to the one suggested by Cario and Nelson (1998), particularly in capturing the tail behavior as indicated by the AD test statistics.

In order to check the goodness of the fit of the autocorrelation structure, we choose to compare the spectral distribution functions using the Kolmogorov-Smirnov criterion. The corresponding test statistics are provided on the last row of Table 2. Although the Johnson fit under the assumption of independence has approximately the same AD test statistic as the Johnson fits with $p = 1$ and $p = 2$, it has significantly the largest spectral test statistic and clearly falls short in providing a good fit for the autocorrelation structure of the physical process. Between the Johnson-type fits with $p = 1$ and $p = 2$, the one with $p = 2$ has a significantly smaller spectral test statistic, providing a better fit for the autocorrelation structure of the process. At the same time, the artafit autocorrelations give significantly better fits for the sample autocorrelations than the autocorrelations of our artafit(2) model. However, a pure correlation match is not the only thing that matters while choosing a good

representation for the underlying system. This will be clear in the visual analysis of the time-series plots in Figures 2, 3, and 4 and scatter plots in Figures 5, 6, and 7.

Next, using the `artafit` and `artafact` models, we generate 519 data points. Figures 2, 3, and 4 display the time-series plots, while Figures 5, 6, and 7 provide the scatter plots of (x_t, x_{t+1}) , (x_t, x_{t+2}) , (x_t, x_{t+3}) for the empirical pressure data and the data of the fitted `artafact` and `artafit` models. The sample paths in Figures 2, 3, and 4 are qualitatively similar, although we observe differences that can be partly attributed to the sampling error. In the empirical time series, there appear spikes that cannot be captured by the `artafact` model which varies more consistently about its mean. In addition, comparison of Figures 5 and 6 shows that the `artafact` data appear to be more scattered or random than the empirical data. Thus, the marginal distribution and autocorrelation structure of the `artafact` process do not perform well in capturing the characteristics of the time-series process. However, our `artafit(2)` model appears to be more successful in representing the characteristics of the empirical data. Comparison of Figures 2 and 4 shows that our `artafit(2)` model captures the height of the spikes reasonably well, while comparison of Figures 5 and 7 shows that our `artafit(2)` process captures the autocorrelation structure of the empirical time-series process. Overall, our `artafit(2)` process provides a very plausible model for the empirical time series.

5.2.2 Modeling Sales of a Large Vehicle Manufacturer

In this section, we analyze the sales data of a large vehicle manufacturer to capture the underlying demand process. Before the application of our fitting algorithm, we remove any trends and seasonality from the sales data using standard methods, because our framework currently applies only to stationary processes. The challenge with fitting an ARTA model to this particular process is its *small* sample size, $n = 90$.

The approach taken by most of the research in supply chain management is to assume a demand process that is either i.i.d. over time or a p^{th} -order autoregressive process. For example, Lee, So, and Tang (2000) assume a first-order autoregressive demand process while analyzing the value of information sharing in a two-level supply chain. In this section, we fit models to the sales data assuming that the underlying demand process is (i) i.i.d. over time, (ii) an autoregressive process, and (iii) an ARTA process.

For model (ii), we fit a Gaussian autoregressive model to the sales data using the software

Figure 2: Time-series Plot of the Empirical Pressure Data

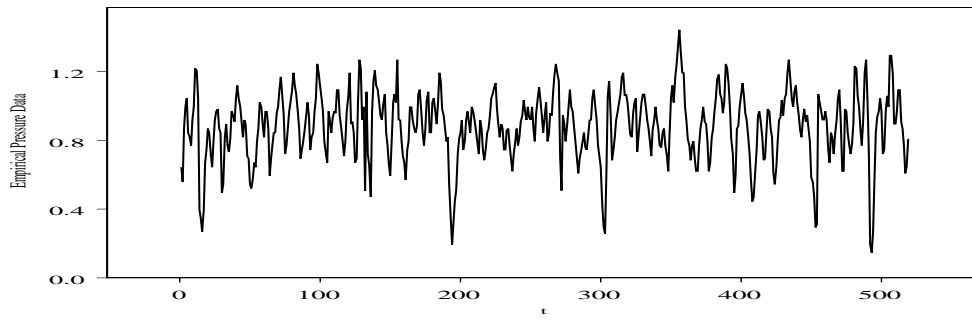


Figure 3: Time-series Plot of the Artefact Data

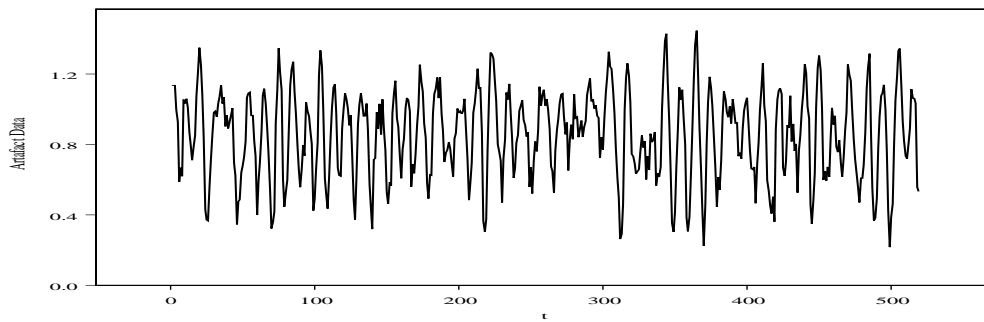


Figure 4: Time-series Plot of the Artefit Data

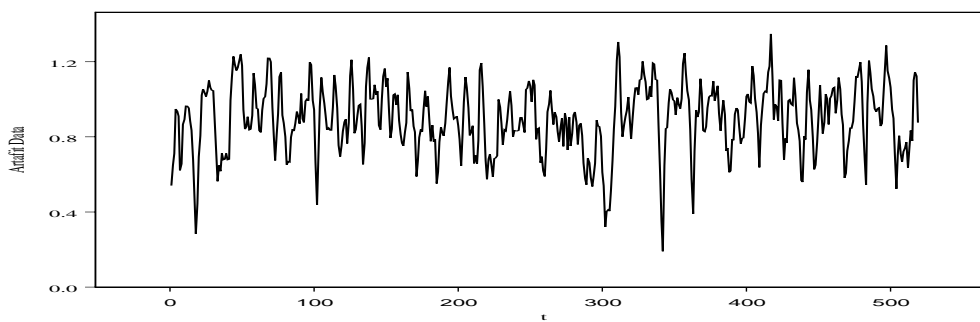


Figure 5: Scatter Plots for the Empirical Pressure Data

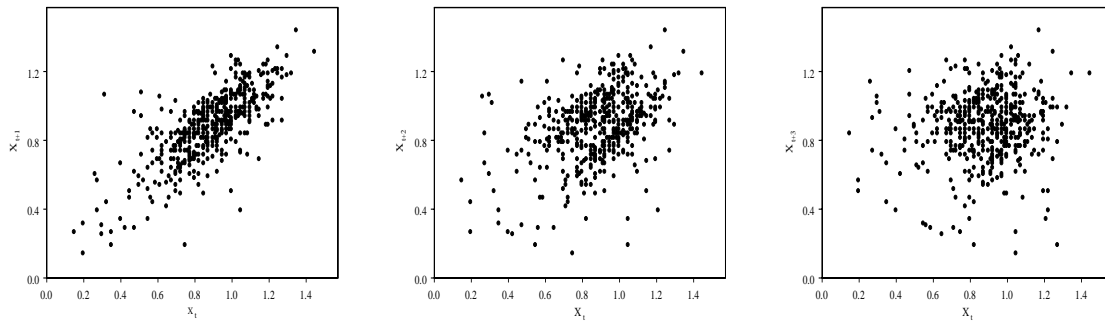


Figure 6: Scatter Plots for the Artefact Data

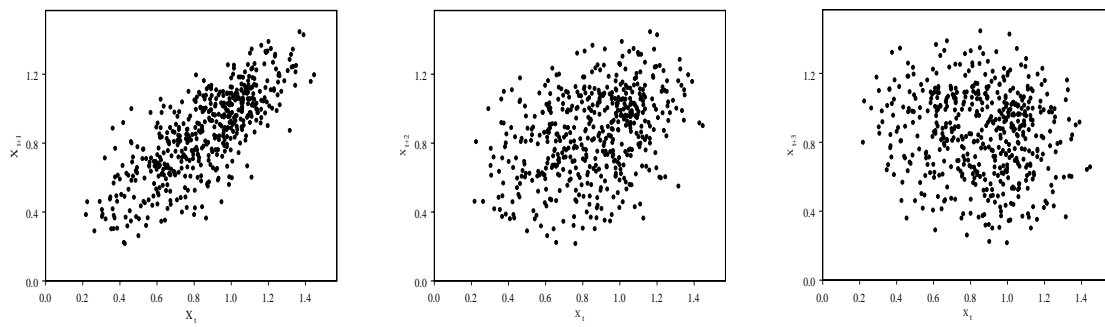
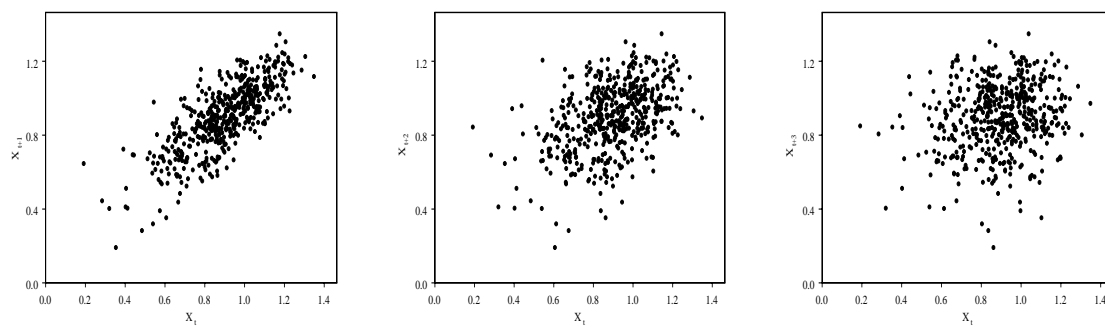


Figure 7: Scatter Plots for the Artefit Data



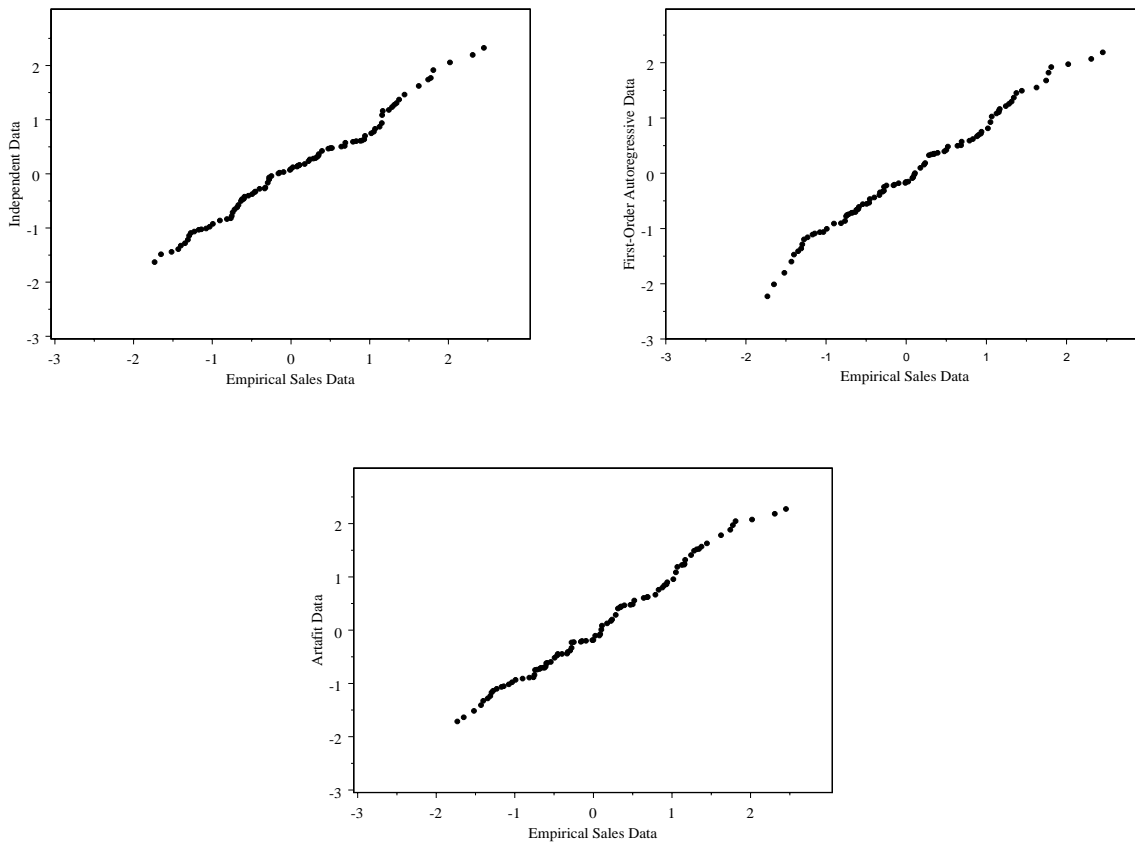


Figure 8: (a) Q-Q Plot Comparing the Empirical Sales Data and the Independent Data (b) Q-Q Plot Comparing the Empirical Sales Data and the Autoregressive Data (c) Q-Q Plot Comparing the Empirical Sales Data and the Artafit Data

SPLUS that suggested an order of dependence $p = 1$ and for model (iii), our software fit a Johnson bounded distribution and an autocorrelation structure with $p = 1$ characterized by $(\gamma, \delta, \lambda, \xi) = (0.412, 1.099, 5.295, -2.156)$ and $\rho_X(1) = 0.433$. The q-q plots comparing the empirical sales data with the i.i.d. data, the autoregressive data, and the artafit data are given in Figure 8. It is visually obvious that both the i.i.d. fit and the artafit are superior to the autoregressive fit. That i.i.d. fit is marginally superior to the autoregressive fit can be explained by selecting a distribution from the Johnson translation system in the case of independence, while the autoregressive fit assumes the distribution to be normal. Later, we will see which one gives a better fit for the autocorrelation structure of the input process. Between the q-q plots of the i.i.d. fit and the artafit, the artafit appears to perform better on the right tail. Next, we use the goodness-of-fit tests to substantiate our visual observation.

Table 3: Comparison of Kolmogorov-Smirnov, Anderson-Darling, and KS Spectral Test Statistics

	AR(1)	ARF(0)	ARF(1)
KS_X^s	0.528	0.387	0.433
AD_X^s	0.384	0.190	0.158
$KS_{\rho_X}^s$	0.171	0.532	0.139

We report the KS_X^s and AD_X^s test statistics in the first two rows of Table 3. The second column (AR(1)) corresponds to the fit of a first-order autoregressive process, the third column (ARF(0)) corresponds to the fit under the assumption of independence, and the last column (ARF(1)) corresponds to our Johnson fit under the assumption of order 1. Although all of the KS_X^s and AD_X^s test statistics are less than the critical values 0.895 and 0.870 at a significance level of 5%, our fit is statistically superior to the one suggested by the first-order autoregressive process, particularly in capturing the tail behavior as indicated by the AD test statistics. We also observe that the KS spectral test statistic of our fit takes a value that is smaller than the values of the AR(1) and ARF(0) fits, suggesting that the ARTA distribution provides a better fit for the autocorrelation structure of the empirical process than the i.i.d. and AR distributions.

6 CONCLUSION

In this paper, we proposed an automated and statistically valid algorithm to fit stochastic input models to dependent univariate time-series input processes. We illustrated the algorithm using data generated by real-world processes and observed that we were able to develop plausible input models in both examples. Since simulation inputs form the core of every stochastic simulation model, the product of this research is expected to improve the fidelity of practical simulation models, leading to more accurate results and better decisions.

Recently, we suggested a more comprehensive model for representing and generating stationary *multivariate* time-series input processes with arbitrary autocorrelation structures and specifically considered the case of marginal distributions from the Johnson translation system (Billar and Nelson 2002). Our approach is very similar to the one in Cario and Nelson (1996), but we use a vector autoregressive Gaussian process that allows the modeling and generation of multivariate time-series processes. A natural extension of the work presented in this paper is to fit stochastic models to

dependent, multivariate time-series input processes. This is a subject of future research.

APPENDIX

This appendix includes the proofs of the convergence of the data-fitting algorithm and the consistency properties of the resulting estimators.

Proposition 1 *Let $\mathbf{x} = (x_1, x_2, \dots, x_n)'$ denote a vector of data in which ties occur with probability zero. Then, with probability one, there are no local minimum solutions to the objective function $S_{\mathbf{D}}(\psi|\mathbf{x})$ (7) on the boundary of the feasible region $\mathbf{cl}\Psi$.*

Proof. The boundary of the feasible region Ψ is reached when any of the following six cases occur: (i) $\delta = 0$; (ii) $\delta = \infty$; (iii) for $h = 1, 2, \dots, p$, values of α_h , say $\bar{\alpha}_h$, occur such that the absolute value of at least one of the roots of the reverse characteristic polynomial is one, i.e.,

$$\left| \text{RootOf} \left(1 - \sum_{h=1}^p \bar{\alpha}_h B^h = 0, B \right) \right| = 1; \quad (11)$$

(iv) $\lambda = 0$ when the Johnson unbounded family is of interest; (v) $\lambda = x_{(n)} + \xi$ when the Johnson bounded family is of interest; and (vi) $\xi = x_{(1)}$ when either the Johnson lognormal family or the Johnson bounded family is of interest. In all of these cases, the objective function (7) simplifies to $\sum_{t=p+1}^n w(n, p, t) (c_{(t)} - \rho_t)^2$, where $c_{(t)}$, $t = p+1, p+2, \dots, n$, are all equal to the same constant in $(0, 1)$ when (i) and (ii) occur, or every $c_{(t)}$, $t = p+1, p+2, \dots, n$, takes a value from the set $\{0, 1/2, 1\}$ when (iii)-(vi) occur, for the following reasons:

- If case (i) occurs, then the objective function can be written as

$$\sum_{t=p+1}^n w(n, p, t) \left(\Phi \left\{ \frac{\gamma (1 - \sum_{h=1}^p \alpha_h)}{g(p, \boldsymbol{\alpha})} \right\} - \rho_t \right)^2,$$

where $\Phi \{ \gamma (1 - \sum_{h=1}^p \alpha_h) / g(p, \boldsymbol{\alpha}) \}$ will be the same constant in $(0, 1)$ for any fixed γ and $\boldsymbol{\alpha}$, and $t = p+1, p+2, \dots, n$. Notice that $1 - \sum_{h=1}^p \alpha_h > 0$ and $g(p, \boldsymbol{\alpha}) > 0$ by the assumption of a stationary autoregressive base process.

- If case (ii) occurs, then the $c_{(t)}$, $t = p+1, p+2, \dots, n$, will assume only the values 0 and 1.

- If case (iii) occurs, then all values of the input process are identical with probability one. This follows from Theorem 5.2.2 of Anderson (1971). Therefore, when (11) holds, the underlying base process is deterministic with $g(p, \boldsymbol{\alpha}) = 0$ and depending on the sign of the function $V_t(\boldsymbol{\psi})$, the transformation $R_t(\boldsymbol{\psi})$, $t = p + 1, p + 2, \dots, n$, takes values only from the set $\{0, 1/2, 1\}$.
- For cases (iv)-(vi), the direct insertion of $\lambda = 0$, $\lambda = x_{(n)} + \xi$, and $\xi = x_{(1)}$ into the objective function (7) forces $c_{(t)}$, $t = p + 1, p + 2, \dots, n$, to assume only the values from the set $\{0, 1/2, 1\}$.

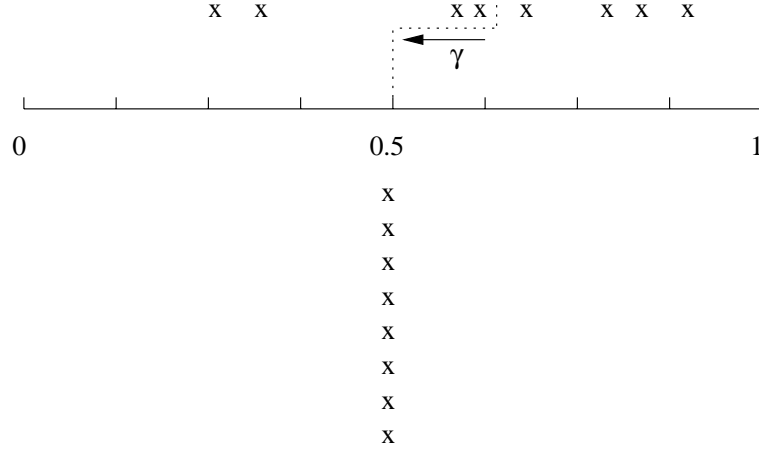
In all of the cases (i)-(vi), the objective function is minimized by letting $c_{(t)}$ be equal to $1/2$ for all values of t . In the remainder of the proof, we will show that we can always find a point in the interior of $\mathbf{cl}\Psi$ that makes the objective function value smaller than it is when $c_{(t)} = 1/2$ for $t = p + 1, p + 2, \dots, n$: First, notice that the function $V_t(\boldsymbol{\psi})$ can be written as

$$V_t(\boldsymbol{\psi}) = \frac{\gamma(1 - \sum_{h=1}^p \alpha_h) + \delta \left[f \left[\frac{X_t - \xi}{\lambda} \right] - \sum_{h=1}^p \alpha_h \left(\gamma + \delta f \left[\frac{X_{t-h} - \xi}{\lambda} \right] \right) \right]}{g(p, \boldsymbol{\alpha})}.$$

If we solve for γ that satisfies $\sum_{t=p+1}^n \Phi\{V_t(\boldsymbol{\psi})\} / (n - p) = 1/2$, where the parameters δ , λ , ξ , and α_h , $h = 1, 2, \dots, p$, take values from the interior of $\mathbf{cl}\Psi$, then we obtain the following representation for γ :

$$\gamma = \frac{-\delta / (n - p) \sum_{t=p+1}^n \left[f \left[\frac{X_t - \xi}{\lambda} \right] - \sum_{h=1}^p \alpha_h \left(\gamma + \delta f \left[\frac{X_{t-h} - \xi}{\lambda} \right] \right) \right]}{1 - \sum_{h=1}^p \alpha_h}. \quad (12)$$

This particular value of γ slides all the transformed data points, $\Phi\{V_t(\boldsymbol{\psi})\}$, $t = p + 1, p + 2, \dots, n$, in such a way that half of them lie below $1/2$, which is the middle of the $[0, 1]$ line, while the other half lies above $1/2$, bringing the transformed data points closer to their ideal order statistics. This is illustrated in Figure 13. All the transformed data points, marked by x , are shown to be lined up in the middle and below the $[0, 1]$ line for the case $c_{(t)} = 1/2$, $t = p + 1, p + 2, \dots, n$. Since the Johnson translation system imposes no restriction on γ and the transformation $\Phi(\cdot)$ ensures that any relocation of the data points result in values in the $[0, 1]$ interval, we can always find a value for γ in the interior of the feasible region that centers the transformed data. So, above the $[0, 1]$ line, sliding all the transformed data points by γ (12) to the left ensures that half of the transformed data points are in $[0, 1/2]$ and the other half in $(1/2, 1]$. This way, the transformed data points get

Figure 9: Effect of the change in γ on the location of the transformed points.

closer to the ideal order statistics, resulting in a lower objective function value.

Thus, when a solution occurs on the boundaries (i)-(vi), we can always decrease the objective function value by finding a value for γ that improves over $c_{(t)} = 1/2$, $t = p + 1, p + 2, \dots, n$, in the interior of $\mathbf{cl}\Psi$. Hence, there are no local minimum solutions to the objective function on the boundary of the feasible region $\mathbf{cl}\Psi$. ■

Proposition 2 For nonnegative integers n and p , the objective function $S_{\mathbf{D}}(\boldsymbol{\psi}|\mathbf{x})$ is a three-times continuously differentiable function for every $\boldsymbol{\psi} \in \Psi$.

Proof. It is straightforward to show that $R_t(\boldsymbol{\psi})$ is a three-times continuously differentiable function of $\boldsymbol{\psi}$. Now consider

$$S_{\mathbf{D}}(\boldsymbol{\psi}|\mathbf{x}) = \frac{1}{(n-p)^2} \sum_{t=p+1}^n \frac{(n-p+1)^2(n-p+2)}{(t-p)(n+1-t)} \left(R_{(t)}(\boldsymbol{\psi}) - \frac{t-p}{n-p+1} \right)^2$$

and let $\boldsymbol{\psi}' \in \Psi$ be a parameter setting such that $R_{p+1}(\boldsymbol{\psi}') \neq R_{p+2}(\boldsymbol{\psi}') \neq \dots \neq R_n(\boldsymbol{\psi}')$. The continuity of R_t implies that there is a neighborhood around $\boldsymbol{\psi}'$ in which the R_t will not change order. Thus, for all such points $S_{\mathbf{D}}(\boldsymbol{\psi}'|\mathbf{x})$ is easily shown to be three-times continuously differentiable.

Therefore, we only need to consider parameter settings $\boldsymbol{\psi}^\dagger$ at which some or all of the R_t are equal. Suppose that $R_{p+1}(\boldsymbol{\psi}^\dagger) = R_{p+2}(\boldsymbol{\psi}^\dagger) = \dots = R_n(\boldsymbol{\psi}^\dagger)$ (cases in which not all of the R_t are equal are a simple variation of the following analysis and will not be done separately). Assuming

that ties are broken arbitrarily, we can write

$$S_{\mathbf{D}}(\boldsymbol{\psi}^\dagger|\mathbf{x}) = \frac{1}{(n-p)^2} \sum_{t=p+1}^n \frac{(n-p+1)^2(n-p+2)}{(t-p)(n+1-t)} \left(R_t(\boldsymbol{\psi}^\dagger) - \frac{t-p}{n-p+1} \right)^2.$$

Notice that R_t replaces $R_{(t)}$. Thus, $S_{\mathbf{D}}(\boldsymbol{\psi}^\dagger|\mathbf{x})$ is three-times continuously differentiable at such points because $R_t(\boldsymbol{\psi}^\dagger)$ is. Hence, the proof is complete. ■

Corollary 1 *The objective function $S_{\mathbf{D}}(\boldsymbol{\psi}|\mathbf{x})$ is convex around any unconstrained local minimum.*

Proof. The objective function $S_{\mathbf{D}}(\boldsymbol{\psi}|\mathbf{x})$ has continuous first- and second-order derivatives (Proposition 2). Hence, the result follows from Theorem 4.4 of Simmons (1975). ■

Proof of Theorem 1 The convergence of the data-fitting algorithm follows from Theorem 7.3.4 of Bazaraa, Sherali, and Shetty (1993). First we present this theorem and then discuss its application to our data-fitting algorithm.

Theorem 7.3.4 *Let Ψ be a non-empty closed set in the m -dimensional real Euclidean space, composed of all real vectors of dimension m and denoted by \mathfrak{R}^m , and let $\Omega \subseteq \Psi$ be a non-empty solution set. Let $\mathbf{C} : \Psi \rightarrow \Psi$ and $\mathbf{D} : \Psi \rightarrow \Psi$ be point-to-set maps and β a continuous function in Ψ , with the following properties:*

- (i) *The map \mathbf{C} is closed over the complement of Ω and satisfies $\beta(\bar{\boldsymbol{\psi}}) < \beta(\boldsymbol{\psi})$ for each $\bar{\boldsymbol{\psi}} \in \mathbf{C}(\boldsymbol{\psi})$ if $\boldsymbol{\psi} \notin \Omega$.*
- (ii) *Given $\boldsymbol{\psi} \in \Psi$, then $\beta(\bar{\boldsymbol{\psi}}) \leq \beta(\boldsymbol{\psi})$ for $\bar{\boldsymbol{\psi}} \in \mathbf{D}(\boldsymbol{\psi})$.*

Now consider the algorithm defined by the composite map $\mathbf{E} = \mathbf{DC}$. Given $\boldsymbol{\psi}_0 \in \Psi$, suppose that the sequence $\{\boldsymbol{\psi}_k\}$ is generated as follows:

If $\boldsymbol{\psi}_k \in \Omega$, then stop; otherwise, let $\boldsymbol{\psi}_{k+1} \in \mathbf{E}(\boldsymbol{\psi}_k)$, replace k by $k+1$, and repeat.

Suppose that the set $\Lambda = \{\boldsymbol{\psi} : \beta(\boldsymbol{\psi}) \leq \beta(\boldsymbol{\psi}_0)\}$ is compact. Then, either the algorithm stops in a finite number of steps with a point in Ω or all accumulation points of $\{\boldsymbol{\psi}_k\}$ belong to Ω .

From here on, for $\mathbf{cl}\Psi$, $\Omega(\mathbf{x})$, \mathbf{C} , and \mathbf{D} , we will use the definitions provided in Section 2.1

Letting $m = p + 4$, we prove the convergence of our data-fitting algorithm showing that it satisfies all the requirements described in Theorem 7.3.4:

First, let the descent function β be the objective function $S_{\mathbf{D}}(\boldsymbol{\psi}|\mathbf{x})$. Notice that β is continuous in $\boldsymbol{\psi}$ (Proposition 2) and the point-to-point maps \mathbf{C} and \mathbf{D} defined in (9) and (10), respectively, satisfy the properties (i) and (ii). The former property follows from the construction of the Levenberg-Marquardt method we use to solve the point-to-point maps of the algorithm (Bazaraa, Sherali, and Shetty 1993, Section 8.7) and the latter property holds because we will never return a solution that increases β .

Next, we define the set $\mathbf{\Lambda} = \{\boldsymbol{\psi} \in \mathbf{cl}\Psi : \beta(\boldsymbol{\psi}) \leq \beta(\boldsymbol{\psi}_0)\}$, where $\boldsymbol{\psi}_0$ corresponds to the starting parameter vector in the interior of $\mathbf{cl}\Psi$. We will show that the set $\mathbf{\Lambda}$ is compact (bounded and closed) by following a discussion similar to the one in Proposition 1: Suppose the set $\mathbf{\Lambda}$ is not bounded. Since there are no local solutions to the objective function $S_{\mathbf{D}}(\boldsymbol{\psi}|\mathbf{x})$ on the boundary of the feasible region $\mathbf{cl}\Psi$, and the autoregressive coefficients lie in a bounded region due to the stationarity of the underlying base process, we will consider the cases when there exists (i) no upper bound on λ when either the Johnson unbounded family or the Johnson bounded family is of interest; (ii) no lower bound on ξ when either the Johnson unbounded family or the Johnson lognormal family is of interest; and (iii) neither a lower bound nor an upper bound on γ for any of the Johnson families.

Notice that the function $V_t(\boldsymbol{\psi})$, $t = p + 1, p + 2, \dots, n$, can be written as

$$\frac{\gamma(1 - \sum_{h=1}^p \alpha_h) + \delta\left(f\left[\frac{X_t - \xi}{\lambda}\right] - \sum_{h=1}^p \alpha_h f\left[\frac{X_{t-h} - \xi}{\lambda}\right]\right)}{g(p, \boldsymbol{\alpha})}.$$

Given any starting solution $\boldsymbol{\psi}_0$ in the interior of $\mathbf{cl}\Psi$ and letting the unrestricted parameters grow without bound, the $c_{(t)}$, $t = p + 1, p + 2, \dots, n$, all approach the same constant for any given values of the remaining parameters. This drives the transformed data points $\Phi\{V_t(\boldsymbol{\psi})\}$, $t = p + 1, p + 2, \dots, n$, away from their ideal order statistics. However, for $\boldsymbol{\psi} \in \mathbf{cl}\Psi$ to be contained in the set $\mathbf{\Lambda}$, it needs to satisfy $\beta(\boldsymbol{\psi}) \leq \beta(\boldsymbol{\psi}_0)$; that is, the parameter vector $\boldsymbol{\psi}$ takes such values for both the input and base process parameters that the transformed data points $\Phi\{V_t(\boldsymbol{\psi})\}$, $t = p + 1, p + 2, \dots, n$, get closer to their ideal order statistics. Clearly, allowing any of the unrestricted parameters to grow without bound will eventually lead to solutions with objective function values larger than $\beta(\boldsymbol{\psi}_0)$. Therefore, the set $\mathbf{\Lambda}$ is bounded.

Since $\mathbf{\Lambda}$ is bounded, for every sequence $\{\boldsymbol{\psi}_k\}$ in $\mathbf{\Lambda}$, there is a convergent subsequence $\{\boldsymbol{\psi}_{k_j}\} \rightarrow \bar{\boldsymbol{\psi}}$, with limit $\bar{\boldsymbol{\psi}} \in \mathbf{cl}\Psi$. Since the function β is continuous, then $\lim_{j \rightarrow \infty} \beta(\boldsymbol{\psi}_{k_j}) = \beta(\bar{\boldsymbol{\psi}})$. In order to complete the proof that the set $\mathbf{\Lambda}$ is compact, we need to show that $\bar{\boldsymbol{\psi}} \in \mathbf{\Lambda}$. Suppose $\bar{\boldsymbol{\psi}} \notin \mathbf{\Lambda}$. Since β is continuous, there exists a neighborhood, say $N_\epsilon(\bar{\boldsymbol{\psi}})$, around $\bar{\boldsymbol{\psi}}$ such that $\boldsymbol{\psi} \in N_\epsilon(\bar{\boldsymbol{\psi}})$ implies $\boldsymbol{\psi} \notin \mathbf{\Lambda}$. But since $\boldsymbol{\psi}_{k_j} \rightarrow \bar{\boldsymbol{\psi}}$, there exists an index, say \bar{k} , for which $\boldsymbol{\psi}_{k_j} \in N_\epsilon(\bar{\boldsymbol{\psi}}), \forall j \geq \bar{k}$ holds. Then, $\boldsymbol{\psi}_{k_j} \notin \mathbf{\Lambda} \forall j \geq \bar{k}$, resulting in a contradiction. Thus, $\bar{\boldsymbol{\psi}} \in \mathbf{\Lambda}$, indicating that for every sequence in $\mathbf{\Lambda}$, there is a convergent subsequence with a limit in $\mathbf{\Lambda}$ and the set $\mathbf{\Lambda}$ is compact.

Thus, we have shown that our data-fitting algorithm satisfies all the assumptions of Theorem 7.3.4. Therefore, either the algorithm stops in a finite number of steps at a point in $\mathbf{\Omega}(\mathbf{x})$ or it generates the infinite sequence $\{\boldsymbol{\psi}_k\}$ such that all of its accumulation points belong to $\mathbf{\Omega}(\mathbf{x})$. ■

Proof of Theorem 2 We first establish that, in the limit, the least-squares estimators will converge to a value of $\boldsymbol{\psi}$, denoted generically as $\boldsymbol{\psi}^\dagger$, at which $G(r) = \Pr\{R_t(\boldsymbol{\psi}^\dagger) \leq r\} = r$ for all t , and for which

$$\sqrt{n-p} \sup_{r \in (0,1)} \left| G(r) - \hat{G}_n(r) \right| \xrightarrow{\mathcal{L}} M \text{ as } n \rightarrow \infty,$$

where M is a proper random variable and $\hat{G}(r)$ is the empirical cdf of the $R_t(\boldsymbol{\psi}^\dagger)$. Stated differently, the least squares estimators will converge to a parameter setting $\boldsymbol{\psi}^\dagger$ at which $R_t(\boldsymbol{\psi}^\dagger)$ is marginally uniform. To do so, we give an alternative representation for the objective function (7) that does not use the order statistics.

By defining $\tilde{F}_n(v) \equiv \#\{V_t(\boldsymbol{\psi}) \leq v, t = p+1, p+2, \dots, n\} / (n-p+1)$, we can write the expanded form of the objective function of the DWLS least-squares estimation problem (7) as

$$\frac{n-p+2}{(n-p)^2} \sum_{t=p+1}^n \frac{\left(\Phi\{V_t(\boldsymbol{\psi})\} - \tilde{F}_n(V_t(\boldsymbol{\psi})) \right)^2}{\tilde{F}_n(V_t(\boldsymbol{\psi})) \left(1 - \tilde{F}_n(V_t(\boldsymbol{\psi})) \right)}. \quad (13)$$

Letting $R_t(\boldsymbol{\psi}) = \Phi\{V_t(\boldsymbol{\psi})\}$ and noting that $R_t(\boldsymbol{\psi})$ is nondecreasing in $V_t(\boldsymbol{\psi})$, we can further write (13) as

$$\frac{n-p+2}{(n-p)^2} \sum_{t=p+1}^n \frac{\left(R_t(\boldsymbol{\psi}) - \tilde{G}_n(R_t(\boldsymbol{\psi})) \right)^2}{\tilde{G}_n(R_t(\boldsymbol{\psi})) \left(1 - \tilde{G}_n(R_t(\boldsymbol{\psi})) \right)}, \quad (14)$$

where $\tilde{G}_n(r) \equiv \#\{R_t(\boldsymbol{\psi}) \leq r, t = p+1, p+2, \dots, n\} / (n-p+1)$.

Let $\hat{G}_n(r) \equiv \#\{R_t(\boldsymbol{\psi}) \leq r, t = p+1, p+2, \dots, n\} / (n-p)$ and let $G(r)$ correspond to the limiting cdf to which $\hat{G}_n(r)$ converges. Since $\tilde{G}_n(r) - \hat{G}_n(r) \rightarrow 0$ a.s. as $n \rightarrow \infty$, for sufficiently large n , we can write the objective function (14) as

$$\frac{1}{n-p} \sum_{t=p+1}^n \frac{\left(R_t(\boldsymbol{\psi}) - \hat{G}_n(R_t(\boldsymbol{\psi}))\right)^2}{\tilde{G}_n(R_t(\boldsymbol{\psi})) \left(1 - \tilde{G}_n(R_t(\boldsymbol{\psi}))\right)}.$$

We will show that

$$\lim_{n \rightarrow \infty} \frac{1}{n-p} \sum_{t=p+1}^n \frac{\left(R_t(\boldsymbol{\psi}) - \hat{G}_n(R_t(\boldsymbol{\psi}))\right)^2}{\tilde{G}_n(R_t(\boldsymbol{\psi})) \left(1 - \tilde{G}_n(R_t(\boldsymbol{\psi}))\right)} \xrightarrow{\text{a.s.}} \begin{cases} 0, & \text{if } \boldsymbol{\psi} = \boldsymbol{\psi}^\dagger, \\ > 0, & \text{otherwise.} \end{cases} \quad (15)$$

That is, there exists a parameter vector $\boldsymbol{\psi} = \boldsymbol{\psi}^\dagger$ for which the objective function goes to zero with probability one as the sample size approaches infinity and the objective function converges to a positive value when $\boldsymbol{\psi} \neq \boldsymbol{\psi}^\dagger$. However, we will also show that $\boldsymbol{\psi}^\dagger$ does not necessarily correspond to the true parameter vector $\boldsymbol{\psi}^*$.

First, suppose $\boldsymbol{\psi} = \boldsymbol{\psi}^\dagger$. We replace the notation $R_t(\boldsymbol{\psi})$ by R_t for ease of presentation in the remainder of the proof:

$$\begin{aligned} & \frac{1}{n-p} \sum_{t=p+1}^n \frac{\left(R_t - \hat{G}_n(R_t)\right)^2}{\tilde{G}_n(R_t) \left(1 - \tilde{G}_n(R_t)\right)} \\ &= \frac{1}{n-p} \sum_{t=p+1}^n \frac{\left|R_t - \hat{G}_n(R_t)\right|}{1 - \tilde{G}_n(R_t)} \frac{\left|R_t - \hat{G}_n(R_t)\right|}{\tilde{G}_n(R_t)} \\ &= \frac{1}{n-p} \sum_{t=p+1}^n \frac{1 + \tilde{G}_n(R_t)}{1 - \tilde{G}_n(R_t)} \frac{\left|R_t - \hat{G}_n(R_t)\right|}{1 + \tilde{G}_n(R_t)} \frac{\left|R_t - \hat{G}_n(R_t)\right|}{\tilde{G}_n(R_t)} \\ &\leq \frac{1}{n-p} \sum_{t=p+1}^n \frac{1 + \tilde{G}_n(R_t)}{1 - \tilde{G}_n(R_t)} \frac{\left|R_t - \hat{G}_n(R_t)\right|}{\tilde{G}_n(R_t)} \\ &= \frac{1}{n-p} \sum_{t=p+1}^n \frac{\left(1 + \tilde{G}_n(R_t)\right) \left|R_t - G(R_t) + G(R_t) - \hat{G}_n(R_t)\right|}{\left(1 - \tilde{G}_n(R_t)\right) \tilde{G}_n(R_t)} \\ &\leq \frac{1}{n-p} \sum_{t=p+1}^n \frac{1 + \tilde{G}_n(R_t)}{\left(1 - \tilde{G}_n(R_t)\right) \tilde{G}_n(R_t)} \left|R_t - G(R_t)\right| \end{aligned} \quad (16)$$

$$+ \frac{1}{n-p} \sum_{t=p+1}^n \frac{1 + \tilde{G}_n(R_t)}{(1 - \tilde{G}_n(R_t)) \tilde{G}_n(R_t)} |G(R_t) - \hat{G}_n(R_t)|. \quad (17)$$

Next we will show how (16) and (17) diminish: The first term is 0 if $G(R_t) = R_t$ for $t = p+1, p+2, \dots, n$, which will occur when $\boldsymbol{\psi} = \boldsymbol{\psi}^\dagger$. For the second term,

$$\begin{aligned} & \frac{1}{n-p} \sum_{t=p+1}^n \frac{1 + \tilde{G}_n(R_t)}{(1 - \tilde{G}_n(R_t)) \tilde{G}_n(R_t)} |G(R_t) - \hat{G}_n(R_t)| \\ & \leq \sup_{p+1 \leq t \leq n} |G(R_t) - \hat{G}_n(R_t)| \frac{1}{n-p} \sum_{t=p+1}^n \frac{1 + \tilde{G}_n(R_t)}{(1 - \tilde{G}_n(R_t)) \tilde{G}_n(R_t)} \\ & \leq \sqrt{n-p} \sup_{r \in (0,1)} |G(r) - \hat{G}_n(r)| \frac{1}{\sqrt{n-p}} \sum_{t=p+1}^n \frac{n+1+t}{t(n+1-t)}. \end{aligned}$$

Notice that $R_t(\boldsymbol{\psi}) \stackrel{\text{i.i.d.}}{\sim} U(0,1)$ for $t = p+1, p+2, \dots, n$ when $\boldsymbol{\psi} = \boldsymbol{\psi}^*$. Thus, when we have the true parameter setting, i.e., $\boldsymbol{\psi}^\dagger = \boldsymbol{\psi}^*$, $\sqrt{n-p} \sup_{r \in (0,1)} |G(r) - \hat{G}_n(r)|$ has a limiting distribution, i.e.,

$$\sqrt{n-p} \sup_{r \in (0,1)} |G(r) - \hat{G}_n(r)| \xrightarrow{\mathcal{L}} M \text{ as } n \rightarrow \infty,$$

where M is a proper random variable (Lehmann 1998). However, this is not the only parameter setting for which a limiting distribution exists, e.g., $\gamma = \gamma^*, \delta = \delta^*, \lambda = \lambda^*, \xi = \xi^*$, and $\alpha_h = 0$ for $h = 1, 2, \dots, p$. Since

$$\lim_{n \rightarrow \infty} \frac{1}{\sqrt{n-p}} \sum_{t=p+1}^n \frac{n+1+t}{t(n+1-t)} = 0,$$

it holds that as $n \rightarrow \infty$

$$\frac{1}{n-p} \sum_{t=p+1}^n \frac{1 + \tilde{G}_n(R_t)}{(1 - \tilde{G}_n(R_t)) \tilde{G}_n(R_t)} |G(R_t) - \hat{G}_n(R_t)| \xrightarrow{\text{a.s.}} M \times 0 = 0. \quad (18)$$

Thus, the discussion in the second paragraph of the previous page together with (18) proves the first line of (15). We have also shown that the objective function can take the value of 0 at a parameter setting that is different from the true one. Next, we will prove the second line of (15) by deriving a lower bound on the objective function value and showing that it approaches a positive

value as $n \rightarrow \infty$ when $\boldsymbol{\psi} \neq \boldsymbol{\psi}^\dagger$:

$$\begin{aligned}
\frac{1}{n-p} \sum_{t=p+1}^n \frac{\left(R_t - \hat{G}_n(R_t)\right)^2}{\tilde{G}_n(R_t) \left(1 - \tilde{G}_n(R_t)\right)} &\geq \frac{1}{n-p} \sum_{t=p+1}^n \left(R_t - \hat{G}_n(R_t)\right)^2 \\
&= \frac{1}{n-p} \sum_{t=p+1}^n \left(R_t - G(R_t) + G(R_t) - \hat{G}_n(R_t)\right)^2 \\
&= \frac{1}{n-p} \sum_{t=p+1}^n \left(R_t - G(R_t)\right)^2 \\
&\quad + \frac{2}{n-p} \sum_{t=p+1}^n \left(R_t - G(R_t)\right) \left(G(R_t) - \hat{G}_n(R_t)\right) \\
&\quad + \frac{1}{n-p} \sum_{t=p+1}^n \left(G(R_t) - \hat{G}_n(R_t)\right)^2 \\
&\geq \frac{1}{n-p} \sum_{t=p+1}^n \left(R_t - G(R_t)\right)^2 \\
&\quad - \frac{2}{n-p} \sum_{t=p+1}^n \left|R_t - G(R_t)\right| \left|G(R_t) - \hat{G}_n(R_t)\right| \\
&\quad + \frac{1}{n-p} \sum_{t=p+1}^n \left(G(R_t) - \hat{G}_n(R_t)\right)^2 \\
&\geq \frac{1}{n-p} \sum_{t=p+1}^n \left(R_t - G(R_t)\right)^2 \\
&\quad - 2 \sup_{r \in (0,1)} \left|r - G(r)\right| \left|G(r) - \hat{G}_n(r)\right|
\end{aligned} \tag{19}$$

$$+ \frac{1}{n-p} \sum_{t=p+1}^n \left(G(R_t) - \hat{G}_n(R_t)\right)^2. \tag{20}$$

$$+ \frac{1}{n-p} \sum_{t=p+1}^n \left(G(R_t) - \hat{G}_n(R_t)\right)^2. \tag{21}$$

Given the order of dependence p and Johnson transformation f , the function G is continuous in $\gamma, \delta, \lambda, \xi$, and $\alpha_h, h = 1, 2, \dots, p$; thus, $G(r) \neq r$ holds for a set of positive probability at any other parameter setting $\boldsymbol{\psi} \neq \boldsymbol{\psi}^\dagger$, resulting in a positive value for (19). Thus, at $\boldsymbol{\psi} \neq \boldsymbol{\psi}^\dagger$,

$$\lim_{n \rightarrow \infty} \frac{1}{n-p} \sum_{t=p+1}^n \left(R_t - G(R_t)\right)^2 > 0.$$

Although we take $\boldsymbol{\psi} \neq \boldsymbol{\psi}^\dagger$, it still holds that, with probability one,

$$\sup_{r \in (0,1)} \left|G(r) - \hat{G}_n(r)\right| \text{ goes to } 0 \text{ as } n \rightarrow \infty.$$

The result follows from the extension of the Glivenko-Cantelli theorem to dependent processes that are metrically transitive (Doob 1953, Theorem 2.1) and therefore satisfy the Strong Law of Large Numbers. Thus, (20) and (21) approach 0 as $n \rightarrow \infty$, resulting in a positive lower bound and proving (15).

Finally, we need to show that $\boldsymbol{\psi}^\dagger$ always includes λ^* and ξ^* , and that if $\alpha_h = \alpha_h^*$, $h = 1, 2, \dots, p$, then $\boldsymbol{\psi}^\dagger = \boldsymbol{\psi}^*$ uniquely. That $\boldsymbol{\psi}^\dagger$ always includes λ^* and ξ^* follows immediately from the proof of Theorem 3 below (it is required for V_t to be marginally normally distributed). And if $\alpha_h = \alpha_h^*$, $h = 1, 2, \dots, p$, then V_t will not be $N(0, 1)$ unless all of the Johnson parameters are correct, from the uniqueness of the Johnson representation of the marginal distribution.

Thus, in either case, if the sample size n is sufficiently large so that (7) is close to the asymptotic limit, then $\hat{\boldsymbol{\psi}}_n$, which minimizes the former, will almost surely be close to $\boldsymbol{\psi}^\dagger$. ■

Theorem 3 *Let X_1, X_2, \dots, X_n be identically distributed random variables with a joint ARTA distribution characterized by the parameter vector $\boldsymbol{\psi}^*$ and assume that the type of the Johnson transformation f and the order of dependence p are known. Then $R_t(\boldsymbol{\psi})$ are i.i.d. uniform random variables on the unit interval $(0, 1)$ if and only if $\boldsymbol{\psi} = \boldsymbol{\psi}^*$.*

Proof. That $\boldsymbol{\psi} = \boldsymbol{\psi}^*$ implies $R_t(\boldsymbol{\psi})$ are i.i.d. $U(0, 1)$ follows immediately from the definition of $V_t(\boldsymbol{\psi})$ and the probability-integral transformation.

Suppose now that $R_t(\boldsymbol{\psi})$ are i.i.d. $U(0, 1)$. Then $V_t(\boldsymbol{\psi})$ must be an i.i.d. standard normal random variable. But since V_t is a linear combination of identically distributed terms of the form

$$W_j = \gamma + \delta f \left[\frac{X_j - \xi}{\lambda} \right], j = t - p, t - p + 1, \dots, t,$$

then the W_j , $j = t - p, t - p + 1, \dots, t$, must be identically distributed normal random variables with mean 0. However, W_j will only have the correct skewness and kurtosis for a normal distribution if $\lambda = \lambda^*$ and $\xi = \xi^*$ by the uniqueness of the Johnson transformation. Combining this with the fact that

$$X_j = \xi^* + \lambda^* f^{-1} \left[\frac{Z_j^* - \gamma^*}{\delta^*} \right],$$

where $Z_j^* = \sum_{h=1}^p \alpha_h^* Z_{j-h} + Y_j^*$ and Y_j^* are i.i.d. $N(0, g^2(p, \boldsymbol{\alpha}^*))$, gives

$$W_j = \gamma + \delta \left(\frac{Z_j^* - \gamma^*}{\delta^*} \right).$$

Notice that W_j is a linear transformation of Z_j^* , implying that it is also a Gaussian AR(p) process with the same autocorrelation structure as Z_j^* . By the uniqueness of the AR(p) representation, $\alpha_h = \alpha_h^*$, $h = 1, 2, \dots, p$ is required for

$$V_t(\boldsymbol{\psi}) = \frac{W_t - \sum_{h=1}^p \alpha_h W_{t-h}}{g(p, \boldsymbol{\alpha})}$$

to be independent. Given this fact, then we must have $\gamma = \gamma^*$ and $\delta = \delta^*$ for V_t to have mean 0 and variance 1. ■

ACKNOWLEDGMENT

This research was partially supported by National Science Foundation Grant numbers DMI-9821011 and DMI-9900164 and Sigma Xi Scientific Research Society Grant number 142.

REFERENCES

- Anderson, T. W. 1971. *The Statistical Analysis of Time Series*. New York: John Wiley and Sons.
- Anderson, T. W. 1993. Goodness of fit tests for spectral distributions. *The Annals of Statistics*, 21, 830–847.
- Bar-Shalom, Y. 1971. On the asymptotic properties of the maximum likelihood estimate obtained from dependent observations. *Journal of the Royal Statistical Society Series B*, 33, 72–77.
- Basawa, I. V., P. D. Feigin and C. C. Heyde. 1976. Asymptotic properties of maximum likelihood estimators for stochastic processes. *Sankhya - Series A*, 38, 259–270.
- Basawa, I. V. and B. L. S. Prakasa Rao. 1980. *Statistical Inference for Stochastic Processes*. London: Academic Press.
- Bazaraa, M. S., H. D. Sherali and C. M. Shetty. 1993. *Nonlinear Programming: Theory and Algorithms*. New York: John Wiley and Sons.

- Bhat, B. R. 1974. On the method of maximum likelihood for dependent observations. *Journal of the Royal Statistical Society Series B*, 36, 48–53.
- Biller, B. 2002. A comprehensive input-modeling framework and software for stochastic discrete-event simulation experiments. Doctoral Dissertation, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.
- Biller, B. and B. L. Nelson. 2002. Modeling and generating multivariate time-series input processes using a vector autoregressive technique. *ACM TOMACS*, forthcoming.
- Block, H. W., N. A. Langberg and D. S. Stoffer. 1990. Time series models for non-Gaussian processes. In *Topics in Statistical Dependence*, ed. W. H. Block, A. R. Sampson and T. H. Savits, pp. 69–83. Hayward, California: Institute of Mathematical Statistics.
- Cario, M. C. and B. L. Nelson. 1996. Autoregressive to anything: Time-series input processes for simulation. *Operations Research Letters*, 19, 51–58.
- Cario, M. C. and B. L. Nelson. 1998. Numerical methods for fitting and simulating autoregressive-to-anything processes. *INFORMS Journal on Computing*, 10, 72–81.
- Chatfield, C. 1999. *The Analysis of Time Series: An Introduction*. New York: Chapman and Hall.
- Chen, H. 2001. Initialization for NORTA: Generation of random vectors with specified marginals and correlations. *INFORMS Journal on Computing*, 13, 312–331.
- Crowder, M. J. 1976. Maximum likelihood estimation for dependent observations. *Journal of the Royal Statistical Society Series B*, 38, 45–53.
- Doob, J. L. 1953. *Stochastic Processes*. New York: John Wiley and Sons.
- Fishman, G. S. 1973. *Concepts and Methods in Discrete-Event Digital Simulation*. New York: John Wiley and Sons.
- 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*, eds. S. Andradottir, K. J. Healy, D. H. Withers and B. L. Nelson, pp. 296–302. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.
- Hartley, H. O. and J. W. K. Rao. 1967. Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54, 93–108.
- Heijmans, R. D. H. and J. R. Magnus. 1986. On the first-order efficiency and asymptotic normality of maximum likelihood estimators obtained from dependent observations. *Statistica*

- Neerlandica*, 40, 169–187.
- Hill, I. D., R. Hill and R. L. Holder. 1976. Fitting Johnson curves by moments. *Applied Statistics*, 25, 180–189.
- Kendall, M. G. and A. Stuart. 1979. *The Advanced Theory of Statistics*. New York: Macmillan.
- Kotz, S., N. Balakrishnan and N. L. Johnson. 2000. *Continuous Multivariate Distributions. Volume 1: Models and Applications*. New York: John Wiley.
- 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.
- Johnson, M. E. 1987. *Multivariate Statistical Simulation*. New York: John Wiley.
- Johnson, N. L. 1949. Systems of frequency curves generated by methods of translation. *Biometrika*, 36, 149–176.
- Lee, H. L., K. C. So, and C. S. Tang. 2000. The value of information sharing in a two-level supply chain. *Management Science*, 46, 626–643.
- Lehmann, E. L. 1998. *Elements of Large-Sample Theory*. New York: Springer-Verlag.
- Lewis, P. A. W., E. McKenzie and D. K. Hugus. 1989. Gamma processes. *Commun. Statist. - Stochastic Models*, 5, 1–30.
- Li, S. T. and J. L. Hammond. 1975. Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. *IEEE Transactions on Systems, Man, and Cybernetics*, 5, 557–561.
- Livny, M., B. Melamed and A. K. Tsolis. 1993. The impact of autocorrelation on queueing systems. *Management Science*, 39, 322–339.
- Mallows, C. L. 1967. Linear processes are nearly Gaussian. *Journal of Applied Probability*, 4, 313–329.
- Marquardt, D. W. 1963. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the SIAM*, 11, 431–441.
- 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*, eds. J. J. Swain, D. Goldsman, R. C. Crain and J. R. Wilson, pp. 135–144. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.
- Nelsen, R. B. 1998. *An Introduction to Copulas*. New York: Springer-Verlag.

- Olsson, D. M. 1974. A sequential simplex program for solving minimization problems. *Journal of Quality Technology*, 6, 53–57.
- Sarma, Y. R. 1986. Asymptotic properties of maximum likelihood estimators from dependent observations. *Statistics and Probability Letters*, 4, 309–311.
- Seber, G. A. F. 1977. *Linear Regression Analysis*. New York: John Wiley.
- Silvey, S. D. 1961. A note on the maximum-likelihood in the case of dependent random variables. *Journal of the Royal Statistical Society Series B*, 23, 444–452.
- Simmons, D. M. 1975. *Nonlinear Programming for Operations Research*. New Jersey: Prentice Hall.
- Song, W. T., L. Hsiao and Y. Chen 1996. Generating pseudorandom time series with specified marginal distributions. *European Journal of Operational Research*, 93, 1–12
- 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.
- Sweeting, T. J. 1980. Uniform asymptotic normality of the maximum likelihood estimator. *Annals of Statistics*, 8, 1375–1381.
- Wald, A. 1949. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20, 595–601.
- Ware, P. P., T. W. Page and B. L. Nelson. 1998. Automatic modeling of file system workloads using two-level arrival processes. *ACM TOMACS*, 8, 305–330.
- Wei, W. W. S. 1990. *Time Series Analysis: Univariate and Multivariate Methods*. New York: Addison Wesley.
- Weiss, L. 1971. Asymptotic properties of maximum likelihood estimators in some nonstandard cases. *Journal of the American Statistical Association*, 66, 345–350.
- Weiss, L. 1973. Asymptotic properties of maximum likelihood estimators in some nonstandard cases, II. *Journal of the American Statistical Association*, 68, 428–430.