The Impact of Dependence on Queueing Systems

Ismail Civelek
Carnegie Mellon University

Bahar Biller
Carnegie Mellon University, billerb@andrew.cmu.edu

Alan Scheller-Wolf
Carnegie Mellon University, awolf@andrew.cmu.edu

Follow this and additional works at: http://repository.cmu.edu/tepper
Part of the Management Sciences and Quantitative Methods Commons
The Impact of Dependence on Queueing Systems

Ismail Civelek,* Bahar Biller†, Alan Scheller-Wolf‡

Tepper School of Business
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213, USA

August 2009

Abstract

The effect of dependence is mostly ignored in queueing systems due to the difficulty in its analysis; the existing studies about dependent queueing systems have restrictive assumptions, such as considering only the M/M/1 queue with order of dependence lag-1 in interarrival and service times. In this paper, we conduct a simulation study to observe the impact of dependence in a single server queue, by using different probability distributions and higher orders of dependence. For generating multivariate input for our simulation, we use Vector-Auto-Regressive-to-Anything (VARTA) method that has never been used in queueing systems. We investigate nonmonotonic behavior of the performance of a single server queue with autocorrelated service times, which is a known observation in the literature but it hasn’t been explained by using VARTA as a novel approach to generate multivariate input.

*email: icivelek@andrew.cmu.edu
†email: billerb@andrew.cmu.edu
‡email: awolf@andrew.cmu.edu
1 Introduction

Most queueing models assume that job interarrival times, service demands, failure times, and repair requirements are independent and identically distributed, each being modeled as a renewal process (see, for example, Kelley [24]). These assumptions lead to models that are very easy to simulate, and are analytically tractable under suitable restrictions. Unfortunately, these models are often poor representations of real-life systems where correlations do, in fact, exist.

Nowadays, it is well known that dependent time-series input processes occur naturally in many service, telecommunication, and manufacturing systems. For example, Melamed et. al. [31] observe autocorrelation in sequences of compressed video frame bitrates, while Ware et. al. [43] 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. Recently, Biller and Nelson [6] demonstrate the existence of strong dependence in the pressure variable of a continuous-flow production line, while Biller [7] fit multivariate input models to the sales of a large vehicle manufacturer that are recorded at fixed time intervals and exhibit strong series dependence.

In this study, our goal is to perform a simulation study to observe effects of dependence on the queueing performance measures like the mean waiting time in the queue and mean slowdown of the jobs. In the related literature, the autocorrelation structure is considered only in the simplest case, lag-1 autocorrelation and M/M/1 system. Our research will expand this simple autocorrelated system into different distributions of interarrival and service times, and lag-p autocorrelation for \( p > 1 \). For this reason, we use Vector-Auto-Regressive-to-Anything (VARTA) method, which was developed by Biller and Nelson [5], for generating multivariate input to simulation and Johnson translation system for wide variety of distributional shapes. Therefore, our main contribution is to combine VARTA as a multivariate input generation technique with queueing theory in the context of effect of dependence.

Our objective in this paper is to profile the sensitivity of queueing models to the assumption of independence and quantify the performance decay in a broad range of systems as a result of ignoring dependencies despite their existence. A related study is performed by Livny et. al. [29] who examine the impact of autocorrelated exponentially distributed interarrival times and autocorrelated service demands on the performance of an infinite capacity single-stage, single-server
system without breakdowns (i.e., M/M/1) under a variety of traffic loads. The authors find that models that do not take autocorrelation into account can predict overly optimistic line lengths and waiting times. However, we still do not know how complex input model dependencies of different strengths and patterns might affect the performance of systems that may include multiple servers, finite buffers, products, unreliable stations of queues with arbitrary marginal distributions. The expected contribution of this study is to fill this gap, investigating the impact of dependent inputs on the performance of a variety of queuing systems, and understanding how the operating principles (i.e., factory physics) — that are very well understood under the assumption of independent inputs — change with dependencies in queueing systems. Considering the advent of increasingly complex systems, spawned by rapidly evolving technologies such as telecommunication and manufacturing, where dependencies in model inputs are both common and significant, we believe providing insight into the impact of dependencies on queuing system performance is a valuable contribution to both the academic community and practitioners.

Since analytical models can only handle the analysis of quite restricted dependence structures as well as queuing systems, we use discrete-event stochastic simulation to perform this study. The major challenge in our research is to find plausible models for representing dependencies in stochastic simulation. When the objective is to model, for example, the processing times of a workpiece across several workcenters in a serial production line, there does not exist many alternatives to the independence assumption. Recently, Biller and Nelson [5] have developed a comprehensive model for representing and generating multivariate time-series input processes with marginal distributions from the Johnson translation system and arbitrary autocorrelation structures represented in product-moment correlations. We use their framework, which allows arbitrary marginal distributions as well as dependence structures represented in rank correlations. Finally, we will present statistical summaries of queue performance, explore a variety of values for factors such as the strength of dependence, pattern of dependence, server utilization, and marginal distributions, and conduct a full-factorial experiment to permit the study of factor interactions.

The rest of the paper is organized as follows: the motivation behind this research is discussed in Section 2. Section 3 gives a comprehensive related literature survey and Section 4 introduces Vector-Auto-Regressive-to-Anything (VARTA) method, then Section 5 presents the experimental design of the systems, then Section 6 includes the simulation results and key findings. Section 7
analyzes the nonmonotonic behavior of the performance of a single server queue with negatively autocorrelated service times. Finally, the conclusion of our study is presented in Section 8, and the related appendices are followed by the section of conclusions.

2 Motivation

In manufacturing, service and telecommunication industries, the iid assumption is mostly used for the arrival, service and other related processes. However, those studies become very poor representations of the real system as autocorrelation and cross-correlation appear in almost all input processes of a real system. In Figure 1, we provide the examples of applications in which different forms of dependence, i.e., autocorrelation, correlation and cross-correlation occur. In this figure, while correlation corresponds the dependence between different processes of the same job in the same time horizon, while cross-correlation is the correlation between different processes of different jobs. For instance, there might exist a correlation between a customer’s interarrival and service times; and there might be a cross-correlation between two different customers’ interarrival and service times. The first attack to the problem of dependence in queueing systems is made in the call center case. The arrival stream to the call centers show burstiness and significant autocorrelation, and the structure of both arrival and service processes are time-dependent [9].

In other telecommunication areas, the dependence is very important in data and voice transfers, such as in ISDN (Integrated Service Digital Network) and ATM (Asynchronous Transfer Mode) technologies [16]. And also, WWW (World Wide Web) environment, the arrival of the internet users to the web sites and the think time exhibit strong autocorrelation and burstiness [14].

In manufacturing systems, the dependence is also ignored in both production environment and planning issues; but significant autocorrelations and cross-correlations have been observed in job arrival, machining time, machine failure and down times, vendor lead time, and material handling time. Altıok [2] observes that the down and failure times are positively correlated, which means that shorter down times follow shorter time to failure and longer down times tend to follow longer failure times. This positive correlation between time to failure and down time is very common in pharmaceutical manufacturing processes; such as, mixing, blending and tablet coating. The nozzles over the tablets are frequently replaced, and these down times are much shorter than the other usual machine failures that occur less and have longer down times [13]. In other manufacturing
applications, the dependence is significant in the pressure variable of a continuous-flow production line [6], correlation among different parts' processes in parallel production line using ATO(Assemble-to-Order) manufacturing[46], and it also affects the performance of sequential and ordering systems in JIT(Just-in-Time) manufacturing systems [42].

In service environments, the dependence becomes very important in modeling customer demand for airline tickets in revenue management area, where demand is price sensitive and changes over time [17]; so, demand shows burstiness and autocorrelation. In fact, service centers, such as, transportation stations, cinema and theaters behave like call centers in telecommunication industry; hence, the arrival stream of customers is bursty and time-independent.

<table>
<thead>
<tr>
<th>Application Area</th>
<th>Autocorrelation</th>
<th>Correlation</th>
<th>Cross-correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manufacturing</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Job interarrival time</td>
<td>✓</td>
<td>*</td>
<td>✓</td>
</tr>
<tr>
<td>Machining time</td>
<td>✓</td>
<td>*</td>
<td>✓</td>
</tr>
<tr>
<td>Machine failure time</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Machine down time</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Vendor lead time</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Material handling time</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td><strong>Service</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customer demand</td>
<td>✓</td>
<td>*</td>
<td>✓</td>
</tr>
<tr>
<td>Service time</td>
<td>✓</td>
<td>*</td>
<td>✓</td>
</tr>
<tr>
<td>Service centers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entertainment(i.e. cinema and theater)</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Transportation stations</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Telecommunication</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Call centers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Call interarrival time</td>
<td>✓</td>
<td>*</td>
<td>✓</td>
</tr>
<tr>
<td>Call service time</td>
<td>✓</td>
<td>*</td>
<td>✓</td>
</tr>
<tr>
<td>Data, video and voice transfer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISDN(Integrated Services Digital Network)</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>ATM(Asynchronous Transfer Mode)</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Internet(World Wide Web)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User interarrival time</td>
<td>✓</td>
<td>*</td>
<td>✓</td>
</tr>
<tr>
<td>User think time</td>
<td>✓</td>
<td>*</td>
<td>✓</td>
</tr>
</tbody>
</table>

Figure 1: Dependencies in manufacturing, service and telecommunication industries.

3 Related Literature Review

In the queueing literature, the dependence structure is mostly ignored, despite the fact that, there exists well known dependencies between interarrival and service times in computer, telecommunication, manufacturing and service systems as discussed in the previous section. The studies about
the impact of dependence on queueing systems can be separated into two main categories based on the approach: analytical and simulation. In analytical approach to the problem, the Markov renewal processes are mostly used; on the other hand, TES (Transfer-Expand-Sample) method [31] and Minification technique [27] are primarily used in the simulation studies. The literature review part is divided into three subparts: Analytical, simulation, and this paper.

The studies which use analytical or simulation methodology are presented in the Sections 3.1 and 3.2, respectively. Finally, we compare this paper with the existing literature regarding the methodology, assumptions about the system, and we discuss the other distinguishing features of our study in Section 3.3.

### 3.1 Analytical

In the analytical approach, the studies are first motivated by the telecommunication systems and call centers, where the correlated arrivals in data and voice transfer have significant impacts on the performance [16]. In those analytical studies, the dependence is considered in either short-range or long-range, and only lag-1 autocorrelation in interarrival and service times is studied in short-range. In other short-range studies, correlation between interarrival and service times are also analyzed. The most common result in those analytical papers is dramatic performance reduction from positive autocorrelation in interarrival times. In addition to positive autocorrelation in interarrival times, positive correlation between the interarrival and service times increase the approach to the steady-state value of the mean waiting time under restrictive assumptions [18] and correlation in interarrival times have very little effect on the departure process under heavy traffic conditions. Moreover, correlation in both short-range and long-range dependent cases, in service times have a great impact on the departure process regardless of the load of the system. Additionally, if the variability is reduced in the arrival stream by increasing the autocorrelation in interarrival times and keeping the other parameters the same, then the performance of the queue increases.

Runnenburg [36] studies the effect of dependence on the average waiting time using an integer-valued Markovian arrival process in an exponential renewal server. After this initial attempt to the problem, the distribution of the waiting time is investigated in Runnenburg [37] and it is characterized by the first autocorrelation. Hadidi [18] investigates the effect of positive correlation between
interarrival and service times in M/M/1 system. He shows that the waiting time distribution is hyperexponential, and analyzes the sensitivity of the convergence of the waiting time density to the correlation coefficient. He concludes that the mean waiting time goes faster to its limiting value when the correlation coefficient increases. In addition to Hadidi’s [18] results, Langaris [25] studies a correlated queue with infinitely many servers, where interarrival and service times are correlated. The paper shows that if the correlation coefficient increases, then the proportion of time during which the system is empty decreases; hence, the probability where there are n customers in the system goes to zero faster.

In the light of Markov renewal arrival stream in analytical studies, Szekli et. al. [41] analyze the behavior of MR/GI/1 queues with positively correlated arrivals. The main focus is the effect of dependency in arrival process on queueing performance and they study three main factors influencing the effect of correlation: (i) differences in the mean interarrival times, (ii) intensity in the Markov renewal process and (iii) variability in the point processes underlying the Markov renewal process. The paper emphasizes the impact of variability with the dependency and it shows that reducing the variability in the interarrival times by increasing the autocorrelation in arrival process increases the performance of the system. In addition to Markov renewal arrival process approach, Shioda [38] studies the problem of the departure process of the MAP/SM/1 queue and shows that the departure process is long-range dependent even when the interarrival times are short-range dependent, if the service time process is long-range dependent. The paper concludes that correlation in both short-range or long-range dependent cases, in service times have great impacts on the departure process regardless of the load of the system; however, the correlation in interarrival times have very little effect in heavily loaded systems. Furthermore, Patuwo et. al. [34] analyze the queueing performance in MR/M/1, which is M/M/1 with Markov renewal arrival stream, queue with an exponential renewal server. They show that the reduction in the performance of the system is the result of the correlation. Recently, in another paper for Markov arrival process approach, Boucherie et. al. [8] study the waiting time distribution in M/M/∞, M/G/∞ and G/M/∞ resequencing queue with dependent interarrival and service times. They use jumps of a semi-Markov process for generating dependent interarrival and service times based on the jumps of this process. Instead of dealing with the performance changes due to the dependence, their main concern is getting an approximation to their model by discretization of interarrival and service times.
The impact of dependencies is sometimes considered analytically, under heavy traffic conditions. Jacobs [23] investigates the effect of dependence between positively autocorrelated and correlated interarrival and service times on the performance measures of a single server queueing system under heavy traffic. The interarrival and service times are generated by a mixed exponential moving average method (EARMA). In addition to this heavy traffic study, Fendick et. al. [16] introduce the Index of Dispersion for Intervals (IDI) which is a three dimensional measure and captures the variability of interarrival and service times caused by individual autocorrelation structures and correlation between these processes. The paper considers M/M/1, M/G/1 and G/G/1 queues with lag-1 autocorrelation in multiple classes of traffic. The large reduction in queueing performance induced by increasing correlations appear again in this study. In an another similar study to Fendick et. al. [16] for modeling the voice and data traffic, Heffes et. al. [19] introduce a model where the autocorrelations are generated by a Markov modulated Poisson process.

As a recent study, Xu [46] makes a structural analysis of a queueing system with multiclassess of correlated arrivals and blocking. This study is motivated from assemble-to-order production systems, in which various components are manufactured or assembled at separate places and the ordering of these components involves the correlation structure. The paper uses a simple queueing model of Poisson arrivals and exponential service times with s parallel servers and considers correlation only in arrival process. In this study, she concludes that more positively correlated arrivals improve the worst performance, which has the longest queue among s parallel servers, by reducing the diversity among the servers. In addition to this study, Li and Xu [28] analyze the dependence structure and bounds for several basic parallel queueing systems with correlated arrival processes. In their parallel servers setup, they consider it as dependent single-server queues. Even though this correlated structure is not analytically tractable, the paper introduces various upper and lower bounds for the statistics of the joint performance measures. Moreover, they study both spatial dependence, which is the dependence among different queues, and the temporal dependence, which is the dependence over different time points.

In other recent analytical studies, Hwang et. al. [21] work on the exact analysis of a queue with 2-state D-MMBA (Discrete time Markov Modulated Batch Arrival) with autoregressive inputs and show that the effect of correlation on queueing performance is less significant when batch size distribution of the arrival process gets smaller. Furthermore, Adan et. al. [1] study the Markov
dependent interarrival and service times in MAP/G/1 queue. The performance measures are analyzed by specific transform methods and numerical examples are given to observe the impact of autocorrelation and cross-correlation between interarrival and service times.

### 3.2 Simulation

Simulation methodology is used to understand the behavior of the queueing systems with dependence due to the difficulty in understanding the impact of dependence analytically. For generating short-range dependent input to the experiments, there are two most used methods: TES (Transfer-Expand-Sample) method [31] and Minification technique [27]. However, for long-range dependent processes, the basic model is defined by nonincreasing sequence of nonnegative numbers and a Gamma process for exponential marginals [35]. The key results in simulation studies are: Autocorrelation in interarrival times have greater impact on the performance than the autocorrelation in service times and average waiting time is monotonically increasing in autocorrelated interarrivals as a function of autocorrelation in service times at a specific utilization level. In addition to most common results, the impact of positively autocorrelated arrivals increases monotonically as a function utilization and average waiting time is monotonically increasing for positively, and non-monotonically increasing in negatively autocorrelated service times as a function of autocorrelation in interarrival times at a specific utilization level. Furthermore, the pattern of the performance measures and the magnitude of the impact of dependence by changing the correlations in corresponding processes, depends on input modeling structure and the load of the system. The tail of the waiting time distribution is much fatter for the long-range dependent arrival input than the independent arrivals.

Livny et. al. [29] is the most comprehensive simulation study in the literature and they analyze the impact of lag-1 autocorrelation generated by either TES (Transfer-Expand-Sample) method [31] or Minification technique [27] on M/M/1 queue. They conclude that introducing autocorrelated arrivals have greater impact than the autocorrelation in service times, and the pattern of the performance measures depends on the input modeling and the load of the system. In both TES and Minification techniques, they observe monotonic increase of the average waiting time in autocorrelation in interarrival times as a function of autocorrelation in service times at a certain utilization level; however, the structure is not monotonic in autocorrelated service times as a function
of autocorrelated arrivals.

Considering the simulation studies using long-range dependency, Resnick et. al. [35] study the effect of long range dependence of autocorrelation in G/M/1 queue at where the dependence exists only in interarrival times. This pioneering paper in the long-range dependency, modifies the interarrival process by imposing dependence in the long-run horizon and investigates the impacts on the queueing performance. They observe that the tail of the waiting time distribution is much fatter for the long-range dependent arrival input than the independent arrivals, and the positively autocorrelated arrivals decrease the performance of the queue. In another long-range dependency study, Dahl et. al. [15] introduce the impact of long-memory arrivals in queueing performance. They analyze the degree of the long-memory, utilization of the server, number of jobs in the system and the interactions among those performance measures by using a factorial simulation experiment. They extend the work of Resnick et. al. [35] by comparing long-range dependent and iid arrivals explicitly, and isolating the stochasticity in arrivals by using deterministic service times; then, they observe strong sensitivities for the means and maxima of the queue statistics.

Regarding the practical relevance of the impact of dependence in manufacturing systems, Altıok [2] analyzes the case for modeling correlation in production systems. He states that the time to failure, vendor lead time, machine down time and pull systems in Kanban as possible examples of the dependent queueing structure in manufacturing environment. Moreover, he performs a simulation study with the assumption of stationary stochastic processes for interarrival and service times in lag-1 autocorrelation by using TES (Transfer-Expand-Sample) method [31]. He observes that waiting time metrics are highly affected by both autocorrelation and the load magnitudes; however, the inventory related metrics are less sensitive to autocorrelation under light traffic, and more sensitive under heavy traffic. In addition to this relevance in manufacturing systems, Takahashi and Nakamura [42] analyzes the impact of autocorrelation in JIT production systems. Similarly, they use TES method [31] to generate lag-1 autocorrelation. They observe that together with the coefficients of variation of production and transportation time, the marginal distribution of demand, and the autocorrelated demand have significant effects on both concurrent and sequential ordering systems.

In other studies using simulation and algorithmic procedures, Nelson and Taaffe study the specific $Ph_t/Ph_t/\infty$ queueing system in the focus of time-dependent behavior. In Nelson and
Taaffe [33], they analyze the single server case; and they extend their study to multiple server case in Nelson and Taaffe [32]. They develop a numerical method to evaluate time-dependent mean, variance and higher order moments of the number of jobs in a $Ph_t/Ph_t/\infty$ queueing system by providing finite sets of differential equations. These equations are integrated numerically by their procedure in order to calculate the time dependent behavior of the queueing performance. For algorithmic purposes, Iravani et. al. [22] recently develop a decomposition algorithm for parallel queues where the arrivals are correlated and service times are bulk. They also extend their algorithm to large systems as an approximation of the performance measures and perform numerical examples to test the accuracy of the decomposition algorithm.

3.3 Our study

Our work fits to the simulation studies of the problem and we use a comprehensive input modeling framework, VARTA, at the first time in dependent queue literature that represents dependence both in time sequence and among the components of the input process of interest. We work on single server queue with using a flexible system of distributions (namely Johnson family) as opposed to assuming exponentially distributed inputs. Unlike other studies in the literature, we work together with both autocorrelation in interarrival and service times, and correlation between interarrival and service times of a customer. By using a novel input model, we introduce cross-correlations between interarrival and service times of different customers as a feedback mechanism of the manager. In addition to novelty of our study, we perform simulations to allow lag-p autocorrelation for $p>1$ and to observe impact of pattern of dependence for lag-2 autocorrelation.

The main results are primarily checking the other simulation studies’ findings by using VARTA input modeling, and other observations in extended cases. Moreover, we explain the nonmonotonic behavior of the performance of a single-server queue with negatively autocorrelated service times. We observe that average waiting time is nonmonotonically increasing in autocorrelated interarrivals as a function of autocorrelation in service times at 80% utilization level, but monotonically increasing for other utilizations. Average waiting time is nonmonotonically increasing in autocorrelated service times as a function of autocorrelation in interarrival times at a specific utilization level. We also observe that the impact of positively autocorrelated arrivals increases monotonically as a function utilization. In addition to our results, we show that positive correlation between interarrival
and service times in the same customer increases the performance of the queue and the impact of correlation increases monotonically as a function of the utilization of the system.

4 Vector-Auto-Regressive-to-Anything (VARTA) method

In this section, we give the descriptions for product-moment and rank correlations, and VARTA process. VARTA is a comprehensive input model for representing and generating multivariate time-series input processes with marginal distributions from the Johnson translation system and arbitrary autocorrelation structures represented in product-moment correlations. We use their framework, which allows arbitrary marginal distributions as well as dependence structures represented in rank correlations. We refer the reader to Appendix 1 for more details about product-moment and rank correlations, and Appendix 2 for how VARTA works.

5 Design of Experiments

In this section, we discuss how to select the distributional properties of the input processes we will experiment with as well as the performance metrics we will use to measure the goodness of the resulting fits. We additionally construct the experimental setup and discuss the selection of the simulation run length, warm-up period and the number of replications that’ll ensure a prespecified level of error.

5.1 Factor Selection

In this paper, we choose the experimental factors from the distributional properties of the input processes of interest. More specifically, we consider the coefficient of variance, the coefficient of skewness, and the coefficient of kurtosis (denoted by $\rho$, $\sqrt{\beta_1}$ and $\beta_2$, respectively) as representative of dependence and pattern of dependence for the underlying dependence structure. We provide a brief discussion of each factor in the remainder of the section.

Our objective is to first experiment with exponential marginal distributions and then choose marginals relaxing the exponential assumptions, i.e., $\rho = 1$, $\sqrt{\beta_1} = 2$ and $\beta_2 = 9$, at a time. Therefore, we choose to work with the Johnson translation system that provides a unique representation for finite values of $\rho$, $\sqrt{\beta_1}$ and $\beta_2$ corresponding to a legitimate probability distribution. A cdf of
any Johnson-type random variable $X$ is specified through

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

where $\gamma$ and $\delta$ are shape parameters, $\xi$ is a location parameter, $\lambda$ is a scale parameter, and $f(\cdot)$ is one of the following transformations:

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

We refer the reader to Figure 2 for the partition of the two-dimensional plot of $\beta_1$ and $\beta_2$ into regions in which of a different Johnson family is used to match the third and fourth moments. Figures 3 and 4 are for the example shapes of the marginals we will choose from the Johnson translation system for modeling queueing inputs.

We characterize the dependence structure of an input process through its order of dependence $p$ and its autocorrelations defined up to order $p$, i.e., $\rho_x(h)$ for $h = 1, 2, \ldots, p$. We let $p$ take values between 0 and 5, where $p = 0$ corresponds to an independent process. For any given value of $p$ and sequence of autocorrelations, there exists a unique sequence of autoregressive coefficients denoted by $\alpha_h$, $h = 1, 2, \ldots, p$ and used for obtaining autocorrelations at higher orders using $\rho_x(h) = \sum_{h=1}^{p} \alpha_h \rho_x(h-1)$. Since our focus is on stationary processes, we require $\alpha_h$, $h = 1, 2, \ldots, p$ to take values in a way that the underlying base process is stationary. In other words, the reverse characteristic polynomial has no roots in and on the complex unit circle, i.e., $1 - \sum_{h=1}^{p} \alpha_h z^p \neq 0$ for $|z| \leq 1$ (Wei [45]).

In order to characterize the autocorrelation structure jointly specified by the order of dependence $p$ and the autocorrelations $\rho_x(h)$, $h = 1, 2, \ldots, p$ or the autoregressive coefficients $\alpha_h$, $h = 1, 2, \ldots, p$, with a single parameter, we define a measure that we will call strength of dependence, denote by $\eta$, and write as follows:

$$\eta = 1 + 2 \lim_{n \to \infty} \sum_{h=1}^{n-1} \left( 1 - \frac{h}{n} \right) |\rho_X(h)|.$$
Figure 2: The two-dimensional region of the square of skewness $\beta_1$ and kurtosis $\beta_2$ any legitimate random variable can have and its partition among the Johnson families.

Figure 3: (a) $S_L$ distributions with $\xi = 0$, $\lambda = 1$, and $\gamma = 0$. (b) $S_U$ distributions.
Notice that $\eta$ measures the rate of decay of dependencies as data points get farther apart from each other, and can be thought of as the number of dependent observations equivalent to one independent observation.

Another factor we consider in the design of our experiments is the pattern of dependence. For the order of dependence $p$, there are $2^p$ patterns to consider, e.g., $p = 1$, we take $0 \leq \rho_X(1) \leq 1$ as one pattern and $-1 \leq \rho_X(1) \leq 0$ as another pattern. When $p = 2$, we consider 4 different patterns, denoted by $(+,+), (-,+), (+,-)$ and $(-,-)$. $(+,+)$ implies that $\rho_X(1) > 0$ and $\rho_X(2) > 0$, while $(+,−)$ implies $\rho_X(1) > 0$ and $\rho_X(2) < 0$. We refer the reader to Figures 5 for comparison of the patterns when $p = 1$ and $|\rho_X(1)| = 0.50$ and to Figure 6 for the pattern when $p = 2$ and $|\rho_X(1)| = |\rho_X(2)| = 0.45$. The input processes of Figure 5 have the same values for $\eta$ that is equal
Figure 6: Autocorrelations of second-order autoregressive processes with different patterns.
to 3.062, despite the alternating pattern in the decay of autocorrelations when $\rho_X(1) = -0.50$. We make a similar observation for patterns $(+, +)$ and $(-, +)$ as well as patterns $(+, -)$ and $(-, -)$ of Figure 6. However, the strength of dependence in the processes with patterns $(+, -)$ and $(-, -)$ ($\eta = 18.647$) is much stronger than the dependence in the processes with patterns $(+, +)$ and $(-, +)$ ($\eta = 5.143$).

5.2 Determination of simulation control parameters

Due to the object oriented nature of the C++ programming language, we develop our simulations in this environment. Since the interest is on the steady-state analysis of the queueing systems, main challenge associated with the estimation of the performance measures via simulation is to overcome the problem of initial transient. We choose the initial conditions (i.e., system state when simulation starts) so as to have an empty system with idle servers and then apply the replication/deletion approach together with Welch’s graphical method. The implementation of this approach to the M/M/1 models with 25%, 50%, 66.7% and 80% system utilizations has suggested the selection of 4000 hours of warm-up period and 30,000 hours as the replication length. Our choice of warm-up period and replication length has been observed to be dependent. Thus, we perform the output data collected after 4000 hours in the next 26,000 hours. We refer the reader to chapter 4 for the estimation of the distributional properties of the performance measures of interest and Section 9 for a detailed description of the replication/deletion method in Law and Kelton [26].

6 Implementation

In this section, we will describe our integration of VARTA generation software and queueing model, then we will give the current simulation results.

6.1 The VARTA Multivariate Input Generation Software and Queueing Simulation Model in C++

The VARTA multivariate input generation software is developed in C++ environment by Biller and Nelson [5]. In our experiments, the autocorrelated data for processes like interarrival and service times, are generated by this software. We refer the reader to the authors for further technical details about the code.
We integrate VARTA multivariate input generation code with exponential marginals and rank correlations for the initial setup, M/M/1 queueing system simulation code. We use the C code for M/M/1 systems from Law and Kelton [26], but we modify it with our experimental purposes. For instance, we define the customers or jobs as they are coming to the server as a class that allows us to generate data when needed. Therefore, we avoid the definition of huge arrays, which we have experienced memory allocation problems at the early stage of simulations; then, we add quantile estimation and slowdown statistics to the original code.

6.2 M/M/1 with lag-1 autocorrelation and no correlation

We have simulation results from four different utilization levels: 25%, 50%, 66.6% and 80%. In the tables and figures, we see product-moment correlation values, but we use the corresponding rank correlation values in our experiments, as shown in Figure 7.

<table>
<thead>
<tr>
<th>Product-moment</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.55</td>
<td>-0.800473248</td>
</tr>
<tr>
<td>-0.40</td>
<td>-0.538238917</td>
</tr>
<tr>
<td>-0.25</td>
<td>-0.316403112</td>
</tr>
<tr>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.276178944</td>
</tr>
<tr>
<td>0.50</td>
<td>0.528701948</td>
</tr>
<tr>
<td>0.75</td>
<td>0.767559688</td>
</tr>
<tr>
<td>0.85</td>
<td>0.860952374</td>
</tr>
</tbody>
</table>

Figure 7: Corresponding rank correlation values for product-moment correlation.

In each experiment with respect to the utilization level, the lag-1 autocorrelation values for interarrival and service times are selected from the set \{-0.55, -0.40, -0.25, 0.00, 0.25, 0.50, 0.75, 0.85\}. We refer the reader to Livny et. al. [29] for detailed discussion about these values.

We use the results where there is no autocorrelation in both arrival and service processes for the mean waiting time as a benchmark for the experiment results. We include the benchmark result for 0.00 - 0.00 cell of the tables in a bold text format. In order to see the impact of dependence, we include the percentage discrepancy of the dependent simulation results from the benchmark case. The notation in tables for lag-1 autocorrelation in interarrival and service times are $\rho_a(1)$ and $\rho_S(1)$, respectively (See Figure 8 for 80% utilization results).

In the case where autocorrelation exists only in arrival process, we observe that the impact
of positive autocorrelation in interarrival time increases monotonically as a function of utilization, and the average waiting time increases monotonically as a function of the autocorrelation value at a certain utilization level. However, in the impact of negatively autocorrelated arrivals, the average waiting time does not increase monotonically as a function of the autocorrelation value at a certain load. In the Figure 8, the average waiting time decreases when we increase the autocorrelation value from $\rho_a(1) = -0.55$ to $\rho_a(1) = -0.40$ at this 80% utilization level, then it increases in the value of autocorrelation. Considering all the utilization levels, this nonmonotonic behavior only exists at 80% utilization level.

Regarding the autocorrelation in service times, the impact of dependence is very interesting. Although the impact of positively autocorrelated service times are monotonically increasing as a function of utilization, it’s not monotonic for the negative autocorrelation case. The average waiting time exhibits nonmonotonic structure for all utilization levels, at where it decreases from $\rho_s(1) = -0.55$ to $\rho_s(1) = -0.40$. We analyze this nonmonotonic behavior of the performance of the queue with negative autocorrelated service times in the next section of the paper.

### 6.3 The impact of correlation between interarrival and service times

In this experimental setup, there is only correlation between interarrival and service times of the same customer; this situation arises from our queueing simulation code, in which the jobs are defined as a class in C++ and assigned an interarrival and service time by required dependence. In this simulation, we also verify the results of analytical papers, in which positively correlated interarrival and service times are expected to decrease the average waiting time (See Figure 9 in interarrival and service plus only correlation plot). With this simulation, we get managerial insights
about the value of positively correlated interarrival and service times within a customer to increase the performance of the system. Thus, a positive feedback system between the arrival and service processes decreases the mean waiting time.

The impact of correlation is higher in high utilizations of the system, which is clearly understood in the literature. The intuition behind this result is that customers tend to wait more in the higher loads, so the positive/negative correlation between interarrival and service times affect the system performance more. For instance, it is beneficial to know that the service time of the upcoming job will be high if the interarrival time of that job is high. Thus, knowing this kind of feedback can be used as a signaling system in allocation of the resources.

6.4 The impact of cross-correlation between interarrival and service times

In this simulation, we are trying to understand the impact of the cross-correlation between the interarrival and service times of two different customers. The cross-correlation means the correlation between interarrival and service times through time; in other words lag-i correlation where \( i \geq 1 \). This situation is a proxy for feedback systems in a telecommunication, manufacturing or service system, where the manager can affect the service of a customer by observing the interarrival time of the previous customer. The experiment is performed, at where there are autocorrelations in both
interarrival and service times, as well as the correlation, and all the values are 0.40. The previous section analyzes the impact of traditional correlation, which is lag-0.

According to the results shown in figure 10, positive cross-correlation decreases the average waiting time. This result gives a managerial insight for the benefit of a feedback system or observing the arrival stream over time. In high cross-correlation level, 0.95, the average waiting time is getting even better as a function of utilization, which is not consistent with the case of the impact of correlation. This situation shows the significant value of having a cross-correlation concept between interarrival and service times of different jobs in the queueing system. Moreover, introduction of the cross-correlation in the literature of dependence modeling in the queueing systems is new because of our input modeling novelty, which is VARTA.

6.5 The effect of dependence pattern in M/M/1 with lag-2 autocorrelation

In this experiment, we analyze the impact of dependence pattern on the performance of the queueing system. The interarrival and service times’ lag-1 and lag-2 autocorrelations are defined. We pick the rank correlation of 0.3. In the tables P means lag-1 autocorrelation of 0.3 and N means lag-1 autocorrelation of -0.3. P and N designs are used another benchmark in addition to iid case. For lag-2 simulations, PP means lag-1 is 0.3 and lag-2 is 0.3; PN means lag-1 is 0.3 and lag-2 is -0.3.
NP and NN are defined in similar way with PN and PP. The percentage values in the tables are defined similarly before as the relative performance of the corresponding design to the iid case.

<table>
<thead>
<tr>
<th>Interarrival Load</th>
<th>iid</th>
<th>PP</th>
<th>PN</th>
<th>NP</th>
<th>NN</th>
<th>N</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.33267</td>
<td>50.30%</td>
<td>9.29%</td>
<td>-12.48%</td>
<td>-21.76%</td>
<td>-16.75%</td>
<td>30.14%</td>
</tr>
<tr>
<td>0.50</td>
<td>1.00520</td>
<td>68.54%</td>
<td>1.47%</td>
<td>-9.54%</td>
<td>-29.46%</td>
<td>-19.30%</td>
<td>35.08%</td>
</tr>
<tr>
<td>0.75</td>
<td>4.01590</td>
<td>82.48%</td>
<td>-4.15%</td>
<td>-5.37%</td>
<td>-29.34%</td>
<td>-18.72%</td>
<td>37.17%</td>
</tr>
</tbody>
</table>

Figure 11: Lag-2 autocorrelated interarrival times and iid service times.

In all utilization levels, if we compare the lag-1 and lag-2 experiments, the mean waiting times of different dependence patterns of autocorrelated interarrival times:

\[ PP > P > PN > NP > N > NN \]

This result is intuitive because if we introduce 0.3 in lag-2, then we increase the positive autocorrelated interarrival since the decay of lag-1 in lag-2 is less than 0.3. In other words, the strength of dependence is greater in PP than P. That is why, \( PP > P \). Similar argument holds other cases as well.

<table>
<thead>
<tr>
<th>Service Load</th>
<th>iid</th>
<th>PP</th>
<th>PN</th>
<th>NP</th>
<th>NN</th>
<th>N</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.33267</td>
<td>17.60%</td>
<td>4.88%</td>
<td>-4.53%</td>
<td>-9.28%</td>
<td>-7.09%</td>
<td>11.79%</td>
</tr>
<tr>
<td>0.50</td>
<td>1.00520</td>
<td>41.39%</td>
<td>1.74%</td>
<td>-4.59%</td>
<td>-16.86%</td>
<td>-11.36%</td>
<td>21.88%</td>
</tr>
<tr>
<td>0.75</td>
<td>4.01590</td>
<td>71.34%</td>
<td>-3.11%</td>
<td>-1.86%</td>
<td>-23.44%</td>
<td>-14.71%</td>
<td>31.83%</td>
</tr>
</tbody>
</table>

Figure 12: Lag-2 autocorrelated service times and iid interarrival times.

When we analyze the impact of dependence pattern in autocorrelated service times and iid interarrival times, the order of the performance of the designs are the same with one exception between PN and NP at 80% utilization level at where the relative performances are almost the same. Therefore,

\[ PP > P > PN > NP > N > NN \]

does not always hold for the queueing systems with autocorrelated service times.

If we compare the absolute values of the relative performances, the impact of the dependence is always greater in the autocorrelated interarrival times except PN at 50% utilization level, in which the relative performance is 1.47% and 1.74% in interarrival and service times cases, respectively.
The exception might be caused by the experimental design. This result is not surprising since it is well known in the literature that the impact of autocorrelation is greater in the interarrival times than the service times.

6.6 The impact of nonexponential marginals - G/M/1

In order to work with nonexponential marginals, we first experiment with exponential marginal distributions previously and then choose marginals relaxing the exponential assumptions now in this section, i.e., $\sqrt{\beta_1} = 2$ and $\beta_2 = 9$, at a time. Therefore, we choose to work with the Johnson translation system that provides a unique representation for finite values of $\sqrt{\beta_1}$ and $\beta_2$ corresponding to a legitimate probability distribution.

Considering nonexponential marginals, we perform G/M/1 queue simulations with relaxing $\beta_1 = 4$ and $\beta_2 = 9$ to deviate on $\beta_1$ and $\beta_2$ plane from exponential distribution. There is no clear pattern of the performance of the queue based on either $\beta_1$ or $\beta_2$. However, the average waiting time of the queue increases as the coefficient of variation of the interarrival distribution increases, which in fact is very intuitive. Clearly, higher positive autocorrelation or higher variation in interarrival times decreases the performance of the system.

7 Why isn’t the impact of negatively autocorrelated service times monotonic?

When service times are positively autocorrelated, then there are clusters of short and long processing times. Due to the clustering of the long processing times, the system observes more accumulation
of jobs/customers in the queue; therefore, it causes longer waiting times in the queue. When the service times are negatively autocorrelated, a long service time is followed by a short service time. Since there would be no clustering of long service times as in the case of positive autocorrelation, there would be enough time for the queue to empty. Thus, we would expect decreasing mean waiting times as functions of negative service times. This is indeed the case for negative service-time autocorrelations with relatively small strength. However, as the strength of the negative service autocorrelation increases, a very short service time is followed by a very long service time, during which incoming customers/jobs start accumulating in the queue. This explains the increasing mean waiting time with negative service time autocorrelation with high strength.

When we compare the waiting time sample path of -0.90 autocorrelated service time case and -0.50 case in figure 12, we observe that there are high peaks in -0.90 case, which might be the cause of the nonmonotonic behavior. Thus, the average waiting time in -0.90 case, 0.31076, is higher than -0.50 case, 0.30089. Moreover, the impact of positive autocorrelation is higher than the impact of negative autocorrelation, which shows the number of customers waiting in queue when the customer leaves the system.

We observe the increase in mean waiting time at -0.9 for 80% and 50%, -0.99 for 25%. We pick 25% utilization level, iid interarrival time and -0.9,-0.5, 0, 0.5 and 0.9 autocorrelated service times. We observe that the mean waiting time increases from 0.30089 to 0.31076 as the autocorrelation decreases from -0.5 to -0.9. In order to analyze this nonmonotonic behavior, we investigate
Figure 15: (a) Histogram of the frequencies of waiting times of (lag-1=-0.9 - lag-1=-0.5) autocorrelated service time cases, (b) Moving averages of -0.9, -0.5 and iid autocorrelated service times

the sample paths of the simulation actual generated data for the interarrival and service times. Moreover, we look at the performance measures for the waiting time and number in queue for the customers in the queue.

In figure 13, first plot is the histogram of the frequencies of the waiting time in sample path with lag-1=-0.9 autocorrelated service times case minus the waiting in sample path with lag-1=-0.5 autocorrelated service time case. The number of customers/jobs with zero waiting time in -0.5 cases is significantly larger than -0.9 case. In other words, large amount of zero-waiting time incidents outweigh other ones and the performance of the queue is better in -0.5 case than -0.9 case. Therefore, the tail of the waiting time distributions in -0.5 autocorrelated service time case is close to left more than -0.9 case. The nonmonotonic result, in which average waiting time of -0.5 case higher than -0.9 case, might be explained by this far left-tail of the waiting time distribution in -0.5 case. In addition to the histogram, there is a moving average of waiting time plot in Figure 13. Even if the moving-average of waiting time of -0.9 case approaches to be smaller than -0.5 case, moving average of the waiting time in -0.5 case is significantly shorter for early stages of the sample path.

8 Conclusion

In our simulation study, we observe that positively autocorrelated arrivals always increase the average waiting time and the impact of positive autocorrelation in interarrival time increases the
average waiting time monotonically as a function of the value of autocorrelation at a certain utilization level. However, the impact of negative autocorrelation in interarrival time only increases nonmonotonically as a function of the autocorrelation value for 80% utilization level, but it exhibits monotonic increase in other utilization levels.

As for the impact of dependent service times on the single server queue, the impact of negative autocorrelation in service time increases nonmonotonically as a function of the autocorrelation value for all utilization levels. We explain this nonmonotonic behavior by the tail behavior of the waiting time distribution. The mass shift in the waiting time distribution to zero causes the nonmonotonic behavior of negatively autocorrelated service times.

In addition to autocorrelated interarrival or service times, positive correlation between interarrival and service times of the same customer increases the performance of the system. Moreover, positive cross-correlation between interarrival and service times of different customers increases the performance of the system significantly. Furthermore, we perform simulations with nonexponential marginals by relaxing $\beta_1 = 4$ and $\beta_2 = 9$. We show the performance of the queue decreases as the coefficient of variation of the interarrival distribution increases, which in fact is very intuitive.

Our main contribution in this study is to combine VARTA as a multivariate input generation technique with queueing theory in the context of effect of dependence. By using a novel approach to generate multi-variate input, we introduce cross-correlation, which is the correlation between interarrival and service times through time, to dependent queue literature. This situation is a proxy for feedback systems in a telecommunication, manufacturing or service system, in which the manager can affect the service of a customer by observing the interarrival time of the previous customer.

Appendix

Appendix 1 - Product-Moment and Rank Correlations

The product-moment correlation matrix for a random vector $X = (X_1, \ldots, X_k)$ is the correlation matrix $\Sigma_X = (\Sigma_X : 1 \leq i, j \leq k)$ where

$$\Sigma_X(i,j) = \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)\text{Var}(X_j)}}$$
However, the rank correlation is in the following form:

$$\Sigma_X(i,j) = \frac{\text{cov}(F_i(X_i), F_j(X_j))}{\sqrt{\text{Var}(F_i(X_i))\text{Var}(F_j(X_j))}}$$

where $F_i$ and $F_j$ are the distribution functions of $X_i$ and $X_j$, respectively.

We prefer rank correlation to product-moment correlation for several reasons:

- Product-moment correlation is a measure of linear dependence, while rank correlation can capture nonlinear dependence relationships that may exist between real-world input processes.
- Rank correlation is invariant with respect to strictly increasing transformations of the random variables involved.
- Rank correlation is always defined, even if the random variables involved have infinite variance.

**Appendix 2 - VARTA Process**

In this simulation study, we use the VARTA input modeling framework for multivariate time-series processes with continuous marginal distributions that uses a highly flexible model to capture the important features present in the dependent systems. We achieve flexibility by combining Gaussian vector autoregressive processes and the Johnson family of distributions to characterize the process dependence and marginal distributions, respectively. The VARTA model we use for representing a stationary $k$-variate time-series input process $\{X_t; t = 0, 1, 2, \ldots\}$ has the following properties:

1. Each component time series $\{X_{i,t}; t = 0, 1, 2, \ldots\}$ has a Johnson-type marginal distribution that can be defined by $F_{X_i}$. In other words, $X_{i,t} \sim F_{X_i}$ for $t = 0, 1, 2, \ldots$ and $i = 1, 2, \ldots, k$.

2. The dependence structure is specified via Pearson product-moment correlations $\rho_X(i, j, h) = \text{Corr} [X_{i,t}, X_{j,t-h}]$, for $h = 0, 1, \ldots, p$ and $i, j = 1, 2, \ldots, k$. Equivalently, the lag-$h$ correlation matrices are defined by $\Sigma_X(h) = \text{Corr} [X_t, X_{t-h}] = [\rho_X(i, j, h)]_{(k \times k)}$, for $h = 0, 1, \ldots, p$, where $\rho_X(i, i, 0) = 1$.

We obtain the $i^{th}$ time series via the transformation $X_{i,t} = F_{X_i}^{-1}[\Phi(Z_{i,t})]$, which ensures that $X_{i,t}$ has distribution $F_{X_i}$ by well-known properties of the inverse cumulative distribution function. Therefore, the central problem is to select the autocorrelation structure, $\Sigma_Z(h)$, $h = 0, 1, \ldots, p$, for
the base process that gives the desired autocorrelation structure, \(\Sigma_X(h), h = 0, 1, \ldots, p\), for the input process.

We choose the base process \(Z_t\) as a stationary, standard Gaussian vector autoregressive process of order \(p\) with the representation

\[
Z_t = \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \cdots + \alpha_p Z_{t-p} + u_t, \ t = 0, \pm 1, \pm 2, \ldots,
\]

where \(Z_t = (Z_{1,t}, Z_{2,t}, \ldots, Z_{k,t})'\) is a \((k \times 1)\) random vector of the observations recorded at time \(t\) and the \(\alpha_i, i = 1, 2, \ldots, p\), are fixed \((k \times k)\) autoregressive coefficient matrices. Finally, \(u_t = (u_{1,t}, u_{2,t}, \ldots, u_{k,t})'\) is a \(k\)-dimensional white noise vector representing the part of \(Z_t\) that is not linearly dependent on past observations; it has \((k \times k)\) covariance matrix \(\Sigma_u\) such that

\[
E[u_t] = 0_{(k \times 1)} \text{ and } E[u_t u_{t-h}'] = \begin{cases} \\
\Sigma_u & \text{if } h = 0, \\
0_{(k \times k)} & \text{otherwise}.
\end{cases}
\]

We assume that the covariance matrix \(\Sigma_u\) is positive definite. Using the C++ programming language, we generate multivariate time series from this \(k\)-variate Gaussian vector autoregressive process of any required length, say \(T\). We now explain how to do this using standard theory:

- First, we obtain the starting values, \(z_{-p+1}, z_{-p+2}, \ldots, z_0\), using the autocorrelation structure, \(\Sigma_Z(h), h = 0, 1, \ldots, p\), and the implied system parameters, \(\alpha_1, \ldots, \alpha_p\) and \(\Sigma_u\). We also obtain a series of Gaussian white noise vectors, \(u_1, u_2, \ldots, u_T\). Then we generate the time series \(z_1, z_2, \ldots, z_T\) recursively as \(z_t = \alpha_1 z_{t-1} + \cdots + \alpha_p z_{t-p} + u_t\) for \(t = 1, 2, \ldots, T\).

- To generate \(z_{-p+1}, z_{-p+2}, \ldots, z_0\) as realizations of \(Z_{-p+1}, Z_{-p+2}, \ldots, Z_0\) whose joint distribution is given by a nonsingular \(kp\)-dimensional multivariate normal distribution, we choose a \((kp \times kp)\) matrix \(Q\) such that \(QQ' = \Sigma_Z\). Then we obtain the starting-value vector as \((z_0, z_{-1}, \cdots, z_{-p+1})' = Q (v_1, \cdots, v_{kp})'\), where the \(v_i\)'s are independent standard normal random variates. In this way, we ensure that the process starts stationary.

- To obtain the series of independent Gaussian white noise vectors, \(u_1, u_2, \ldots, u_T\), we first choose \(k\) independent univariate standard normal variates \(v_1, v_2, \ldots, v_k\), and then multiply by a \((k \times k)\) matrix \(P\) for which \(PP' = \Sigma_u\); that is, \(u_t = P (v_1, v_2, \ldots, v_k)'\). We repeat this
process $T$ times.

We let $\rho_Z(i, j, h)$ be the $(i, j)^{th}$ element of the lag-$h$ correlation matrix, $\Sigma_Z(h)$, and let $\rho_X(i, j, h)$ be the $(i, j)^{th}$ element of $\Sigma_X(h)$. The correlation matrix of the base process $Z_t$ directly determines the correlation matrix of the input process $X_t$, because

$$\rho_X(i, j, h) = \text{Corr}[X_{i,t}, X_{j,t-h}] = \text{Corr}\left[F^{-1}_{X_i}[\Phi(Z_{i,t})], F^{-1}_{X_j}[\Phi(Z_{j,t-h})]\right]$$

for all $i, j = 1, 2, \ldots, k$ and $h = 0, 1, 2, \ldots, p$, excluding the case $i = j$ when $h = 0$. Further, only $\text{E}[X_{i,t}X_{j,t-h}]$ depends on $\Sigma_Z$, since

$$\text{Corr}[X_{i,t}, X_{j,t-h}] = \frac{\text{E}[X_{i,t}X_{j,t-h}] - \text{E}[X_{i,t}]\text{E}[X_{j,t-h}]}{\sqrt{\text{Var}[X_{i,t}]\text{Var}[X_{j,t-h}]}}$$

and $\text{E}[X_{i,t}]$, $\text{E}[X_{j,t-h}]$, $\text{Var}[X_{i,t}]$, $\text{Var}[X_{j,t-h}]$ are fixed by $F_{X_i}$ and $F_{X_j}$ (i.e., $\mu_i = \text{E}[X_{i,t}]$, $\mu_j = \text{E}[X_{j,t-h}]$, $\sigma_i^2 = \text{Var}[X_{i,t}]$ and $\sigma_j^2 = \text{Var}[X_{j,t-h}]$ are properties of $F_{X_i}$ and $F_{X_j}$). Since $(Z_{i,t}, Z_{j,t-h})'$ has a nonsingular standard bivariate normal distribution with correlation $\rho_Z(i, j, h)$, we have

$$\text{E}[X_{i,t}X_{j,t-h}] = \text{E}\left[F^{-1}_{X_i}[\Phi(Z_{i,t})]F^{-1}_{X_j}[\Phi(Z_{j,t-h})]\right]$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F^{-1}_{X_i}[\Phi(z_{i,t})]F^{-1}_{X_j}[\Phi(z_{j,t-h})] \vartheta_{\rho_Z(i,j,h)}(z_{i,t}, z_{j,t-h}) dz_{i,t} dz_{j,t-h},$$

(2)

where $\vartheta_{\rho_Z(i,j,h)}$ is the standard bivariate normal probability density function with correlation $\rho_Z(i, j, h)$.

This development is valid for any marginal distributions $F_{X_i}$ and $F_{X_j}$ for which the expectation (2) exists. However, since $Z_{i,t}$ and $Z_{j,t-h}$ are standard normal random variables with a nonsingular bivariate distribution, the joint distribution of $X_{i,t}$ and $X_{j,t-h}$ is well-defined and the expectation (2) always exists in the case of Johnson marginals. Further, the Johnson translation system is a particularly good choice because

$$X_{i,t} = F^{-1}_{X_i}[\Phi(Z_{i,t})] = \xi_i + \lambda_i f^{-1}_{i} [(Z_{i,t} - \gamma_i)/\delta_i]$$

$$X_{j,t-h} = F^{-1}_{X_j}[\Phi(Z_{j,t-h})] = \xi_j + \lambda_j f^{-1}_{j} [(Z_{j,t-h} - \gamma_j)/\delta_j],$$

(3)

avoiding the need to evaluate $\Phi(Z)$, but still defining a bivariate Johnson distribution.
From (2) we see that the correlation between $X_{i,t}$ and $X_{j,t-h}$ is a function only of the correlation between $Z_{i,t}$ and $Z_{j,t-h}$, which appears in the expression for $\vartheta_{\rho_Z(i,j,h)}$. We denote the implied correlation Corr[$X_{i,t}, X_{j,t-h}$] by the function $c_{ijh}[\rho_Z(i,j,h)]$ defined as

$$
c_{ijh}[\rho_Z(i,j,h)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{X_i}^{-1}[\Phi(z_{i,t})]F_{X_j}^{-1}[\Phi(z_{j,t-h})] \vartheta_{\rho_Z(i,j,h)}(z_{i,t}, z_{j,t-h}) dz_{i,t} dz_{j,t-h} - \mu_i \mu_j \sigma_i \sigma_j.
$$

Thus, the problem of determining $\Sigma_Z(h), h = 0, 1, \ldots, p$, that gives the desired input correlation matrices $\Sigma_X(h), h = 0, 1, \ldots, p$, reduces to $pk^2 + k(k - 1)/2$ individual matching problems in which we try to find the value $\rho_Z(i,j,h)$ that makes $c_{ijh}[\rho_Z(i,j,h)] = \rho_X(i,j,h)$. When rank-type correlations are used, it is possible to find the $\rho_Z(i,j,h)$ values analytically:

$$
c_{ijh}[\rho_Z(i,j,h)] = \frac{6}{\pi} \sin^{-1} \left( \frac{\rho_Z(i,j,h)}{2} \right).
$$

We incorporate this multivariate data generation schema into the object-oriented software we developed for the simulation of queueing systems.

References


