Essays on Operations Management

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Operations Management and Manufacturing

by

İsmail Civelek

Tepper School of Business

Carnegie Mellon University

April 2010

Dissertation Committee:

Alan Scheller-Wolf (Chair)

Bahar Biller

Mustafa Akan

Kinshuk Jerath

Dissertation Abstract

This dissertation focuses on the perishable inventory theory with health-care applications, the multi-variate input modeling for stochastic simulations, and temporal and bivariate dependence modeling for interarrival and service times of the queueing systems. This dissertation contributes to the perishable inventory theory by introducing the critical level policy for the first time with application from the blood platelet inventory management. Moreover, this dissertation introduces the Vector-Auto-Regressive-to-Anything (VARTA) method as an advanced simulation input modeling to analyze the impact of temporal and bivariate dependent interarrival and service times in the queueing systems. A synopsis of the three chapters of the dissertation follows.

Chapter 1: "Blood Platelet Inventory Management with Protection Levels and Substitution"

We consider a discrete-time inventory system for a perishable product that has distinct demand streams for product of different ages; an example of such a system is blood platelets. In addition to inventory holding, outdating and shortage costs, our model includes substitution costs when a demand for a certain-aged item is satisfied by a different-aged item. Our objective is to minimize the expected cost over an infinite time horizon. We introduce the critical level policy to the perishable inventory literature, protecting the newest items against excessive Downward substitution, borrowing an approach from the spare parts literature. This reserves these newest items for future demand for procedures needing younger items (i.e. fresher blood platelets). We model the problem as a Markov Decision Process (MDP) and evaluate the costs of a common heuristic replenishment policy (with and without a protection level) against extant "near optimal" policies in the literature. We show that the protection level policy may outperform other policies, particularly if supplies are capacitated.

Chapter 2: "Failure Probability of VARTA in Higher Dimensions"

Vector-Autoregressive-To-Anything (VARTA) is a highly flexible model for driving largescale stochastic simulations by generating samples of stationary multivariate time series with arbitrary marginal distributions. The construction of this model relies on a stable vector autoregressive process with a positive definite autocorrelation matrix. We show that there exists multivariate time-series input processes for which the conditions of stability and positive definiteness are violated. We investigate the likelihood of this event with increasing number of component time-series processes and order of dependence by extending the onion method, which is used for sampling positive definite correlation matrices for random vectors, to sample positive definite autocorrelation matrices for multivariate time series. We find that the failure probability of VARTA reaches one with increasing number of component time series and order of dependence, but at a rate very much dependent on the rate of decay in temporal dependencies. We conclude with a discussion on an approximation of VARTA that might enable the simulation practitioner to avoid the failure of VARTA in high-dimensional settings.

Chapter 3: "The Impact of Dependence on Single-Server Queueing Systems"

In this study, we use advanced simulation input modeling to study the impact of bivariate and temporal dependencies among interarrival and service times on the performance of a single-server queue. The distinguishing feature of our study from those in the literature is to consider a wide variety of distributional shapes for the probability density functions of the interarrival and service times, and the patterns that arise in the temporal dependencies of the interarrival and service times. We generate dependent interarrival and service times via using the Vector-Auto-Regressive-to-Anything method, which has never before been used in queueing systems. We investigate the impact of dependent interarrival and service times on the average waiting time of M/M/1, M/G/1 and G/M/1 systems. We show that high variance and positive skewed nonexponential distributions decrease the performance of the single-server system. We also compare impact of temporal dependencies in interarrival and service times for M/M/k systems ($k \ge 2$) with the M/M/1 system, and conclude that the effect of dependence decreases in multi-server systems. Our main contribution is to combine this advanced input modeling method with queueing theory for investigating the impacts of dependent interarrival and service times on the average waiting time.

Acknowledgements

First, I would like to thank my dissertation committee, Alan Scheller-Wolf, Bahar Biller, Mustafa Akan and Kinshuk Jerath for working with me during my studies at Tepper. I'm very grateful for their feedback and support during my time at Carnegie Mellon.

I would like to express my gratitude to my advisor Alan Scheller-Wolf for his continuous support and mentorship during my studies. I am very thankful to him for always believing in me.

I thank all my colleagues at Tepper with whom I enjoyed taking classes together and discussing everything from research ideas to sports. I would like to acknowledge Ozgun Ekici, Zumrut Imamoglu, John Turner, Chester Xiang, Guoming Lai, Vineet Kumar, Erkut Sonmez, Viswanath Nagarajan, Borga Deniz, Sinan Sarpca and Nihat Altintas. I have great friendships and memories at Carnegie Mellon with all of you.

I also wish to thank our Ph.D. coordinator Lawrence Rapp for his patience and support. He made my life easier during my time at Carnegie Mellon.

I would like thank my mentor in high school and dear friend, Haluk Yardim. As my math teacher and mentor, he inspired me to go to Bilkent University and pursue a Ph.D. in the US. Also, I would like to thank my friends, Alper Ayhan and Suleyman Demirel, for always being there.

I would like to thank my parents, Himmet and Zinet, my future mother-in-law, Sheila, and my sister, Yasemin, for their support and love. I also thank my closest friends, Alptekin Cetin, Arda Balkanay and Can Ozlu, who are like brothers to me, for their endless friendship and support.

Finally, I would like to thank my beautiful fiancee, Laura Auer. I am thankful to Carnegie Mellon because I found the co-pilot of my life, Laura, here. I am the luckiest man alive for having her in my life. Without her love and patience, this dissertation would never be completed.

All errors are of my own.

Contents

1	Blo	od Platelet Inventory Management with Protection Levels and Substi-	
	tuti	on	1
	1.1	Introduction	1
	1.2	The model	5
		1.2.1 No protection level (c=0) $\ldots \ldots	7
		1.2.2 Positive protection level $(c>0)$	12
	1.3	Computational Complexity	14
	1.4	Numerical Results	15
	1.5	Sensitivity analysis on the cost parameters	19
	1.6	Capacity on order level	23
	1.7	Conclusion	25
2	Fail	ure Probability of VARTA in Higher Dimensions	27
	2.1	Introduction	27
	2.2	ARTA/VARTA Transformations and Reasons of Their Failure	30
	2.3	Sampling Positive Definite Autocorrelation Matrices	34
		2.3.1 Univariate Time-Series Setting	35
		2.3.2 Multivariate Time-Series Setting	37
	2.4	Analysis	39
		2.4.1 Univariate Time-Series Setting	40
		2.4.2 Multivariate Time-Series Setting	42
	2.5	Conclusion	45
3	$Th\epsilon$	e Impact of Dependence on Single-Server Queueing Systems	47
	3.1	Introduction	47
	3.2	Motivation	50

3.3	Litera	ture Review	51
3.4	VART	A for Modeling Interarrival and Service Times	54
3.5	Imple	mentation	56
	3.5.1	First-Order Autocorrelated, Exponentially Distributed Interarrival and	
		Service Times	58
	3.5.2	Second-Order Autocorrelated, Exponentially Distributed Interarrival	
		Times or Service Times	61
	3.5.3	First-Order Autocorrelated, Exponentially Distributed Service Times	
		and Lognormal Interarrival Times	63
	3.5.4	First-Order Autocorrelated, Exponentially Distributed Interarrival Time	\mathbf{s}
		and Lognormal Service Times	67
	3.5.5	Bivariate Dependence Between Exponentially Distributed Interarrival	
		and Service Times	71
	3.5.6	First-Order Autocorrelated, Bivariate Dependent and Exponentially	
		Distributed Interarrival and Service Times	72
	3.5.7	First-Order Autocorrelated Exponentially Distributed Interarrival and	
		Service Times for Multi-Servers	73
3.6	Concl	usion	75
Bibliog	graphy		82

List of Figures

Different demand models	18
Shortage (p_3, p_2, p_1) and outdating (m) cost parameters $\ldots \ldots \ldots \ldots$	20
Substitution costs: α_D^{NM} , α_D^{NO} , α_D^{MO} and α_U^{MN}	21
Substitution $(\alpha_U^{ON}, \alpha_U^{OM})$ and holding costs (h_3, h_2)	22
When to protect if order level, S , is limited. \ldots	24
Illustration of the second-order ARTA infeasibility.	30
The two-dimensional region of the square of skewness β_1 and kurtosis β_2 any	
legitimate random variable can have and its partition among the Johnson	
families	57
Waiting times of a single sample path for lag-one=-0.9 and lag-one=-0.5 au-	
to correlated service times	60
Histogram of the frequencies of waiting times of (lag-one=-0.9 - lag-one=-0.5)	
autocorrelated service time cases	61
Lognormal distributions (a) $\delta = 1$ with mean 1.65, (b) $\delta = 2$ with mean 1.13	63
	Different demand models

List of Tables

1.1	Values of cost variables used in comparison results	15
1.2	Comparison with cost data from Haijema et al. $[37]$	16
1.3	Comparison with modification of shortage costs $\ldots \ldots \ldots \ldots \ldots \ldots$	17
2.1	Probability of the ARTA(p) infeasibility for $p = 1, 2,, 20$	40
2.2	Behavior of $ARTA(p)$, $p = 1, 2,, 20$ when failure occurs	41
2.3	Probability of the $\operatorname{ARTA}(p)$ infeasibility with decreasing temporal dependencies.	42
2.4	Probability of the VARTA _k (1) infeasibility as a function of k	42
2.5	Probability of the $VARTA_k(1)$ infeasibility with decreasing temporal depen-	
	dencies.	43
2.6	Probability of the VARTA ₂ (p) infeasibility as a function of p	43
2.7	Probability of the $VARTA_2(p)$ infeasibility with decreasing temporal depen-	
	dencies.	44
2.8	Mean failure probability of $VARTA_k(p)$ with decreasing temporal dependencies.	45
3.1	First-Order Autocorrelated, Exponentially Distributed Interarrival and Ser-	
	vice Times, 25% utilization	59
3.2	Second-order autocorrelated, exponentially distributed interarrival times, and	
	independent and identically distributed service times $\ldots \ldots \ldots \ldots \ldots$	62
3.3	Second-order autocorrelated, exponentially distributed service times, and in-	
	dependent and identically distributed interarrival times $\ldots \ldots \ldots \ldots$	62
3.4	Temporal Dependence Decay	62
3.5	First-Order Autocorrelated, Exponentially Distributed Service and Interar-	
	rival Times, 50% utilization \ldots	64
3.6	First-Order Autocorrelated, Exponentially Distributed Service Times and Log-	
	normal Interarrival Times ($\delta = 1$), 50% utilization	64

3.7	First-Order Autocorrelated, Exponentially Distributed Service and Interar-	
	rival Times, 50% utilization $\ldots \ldots	66
3.8	First-Order Autocorrelated, Exponentially Distributed Service Times and Log-	
	normal Interarrival Times ($\delta = 2$), 50% utilization	66
3.9	First-Order Autocorrelated, Exponentially Distributed Interarrival and Ser-	
	vice times, 50% utilization \ldots	68
3.10	First-Order Autocorrelated, Exponentially Distributed Interarrival Times and	
	Lognormal Service Times ($\delta = 1$), 50% utilization	68
3.11	First-Order Autocorrelated, Exponentially Distributed Service and Interar-	
	rival Times, 50% utilization $\ldots \ldots	69
3.12	First-Order Autocorrelated, Exponentially Distributed Interarrival Times and	
	Lognormal Service Times ($\delta = 2$), 50% utilization	69
3.13	Bivariate Dependence Between Exponentially Distributed Interarrival and	
	Service Times	70
3.14	First-Order Autocorrelated, Bivariate Dependent and Exponentially Distributed	
	Interarrival and Service Times	72
3.15	First-Order Autocorrelated, Exponentially Distributed Interarrival and Ser-	
	vice Times for $M/M/2$, 40% utilization	73
3.16	First-Order Autocorrelated, Exponentially Distributed Interarrival and Ser-	
	vice Times for $M/M/3$, 26.67% utilization	74
3.17	First-Order Autocorrelated, Exponentially Distributed Interarrival and Ser-	
	vice Times for $M/M/2$, 80% utilization $\ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots$	75
3.18	First-Order Autocorrelated, Exponentially Distributed Interarrival and Ser-	
	vice Times for $M/M/3$, 80% utilization $\ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots$	75
3.19	First-Order Autocorrelated, Exponentially Distributed Interarrival and Ser-	
	vice Times, 50% utilization	78
3.20	First-Order Autocorrelated, Exponentially Distributed Interarrival and Ser-	
	vice Times, 80% utilization	78
3.21	First-Order Autocorrelated, Exponentially Distributed Interarrival and Ser-	
	vice Times, 99% utilization	78
3.22	First-Order Autocorrelated, Exponentially Distributed Service and Interar-	
	rival Times, 80% utilization $\ldots \ldots	79

3.23	First-Order Autocorrelated, Exponentially Distributed Service Times and Log-	
	normal Interarrival Times ($\delta = 1$), 80% utilization	79
3.24	First-Order Autocorrelated, Exponentially Distributed Service and Interar-	
	rival Times, 80% utilization $\ldots \ldots	79
3.25	First-Order Autocorrelated, Exponentially Distributed Service Times and Log-	
	normal Interarrival Times ($\delta = 2$), 80% utilization	80
3.26	First-Order Autocorrelated, Exponentially Distributed Service and Interar-	
	rival Times, 80% utilization $\ldots \ldots	80
3.27	First-Order Autocorrelated, Exponentially Distributed Interarrival Times and	
	Lognormal Service Times ($\delta = 1$), 80% utilization	80
3.28	First-Order Autocorrelated, Exponentially Distributed Service and Interar-	
	rival Times, 80% utilization $\ldots \ldots	81
3.29	First-Order Autocorrelated, Exponentially Distributed Interarrival Times and	
	Lognormal Service Times ($\delta = 2$), 80% utilization	81

Chapter 1

Blood Platelet Inventory Management with Protection Levels and Substitution

We consider a discrete-time inventory system for a perishable product that has distinct demand streams for product of different ages; an example of such a system is blood platelets. In addition to inventory holding, outdating and shortage costs, our model includes substitution costs when a demand for a certain-aged item is satisfied by a different-aged item. Our objective is to minimize the expected cost over an infinite time horizon. We introduce the critical level policy to the perishable inventory literature, protecting the newest items against excessive Downward substitution, borrowing an approach from the spare parts literature. This reserves these newest items for future demand for procedures needing younger items (i.e. fresher blood platelets). We model the problem as a Markov Decision Process (MDP) and evaluate the costs of a common heuristic replenishment policy (with and without a protection level) against extant "near optimal" policies in the literature. We show that the protection level policy may outperform other policies, particularly if supplies are capacitated. ¹

1.1. Introduction

One week after the 9/11 terrorist attacks, Altman [4] reminded the American public about the perishability of blood. He stated, "Few people realize that blood is perishable and cannot be stored indefinitely. Blood centers function more as pipelines than banks, and there is a steady need for donors." Altman's emphasis on the need of more blood donors stems from

 $^{^1\}mathrm{Co}\text{-authors:}$ Itir Karaesmen and Alan Scheller-Wolf

the perishability of blood, and also points to the importance of effective utilization of blood resources. Of course, blood is not needed only in times of crisis; the American Red Cross states that "every 2 seconds an American needs blood," which shows that blood inventory management decisions could be life-saving, every day.

Not only is the utilization of blood very important, but in addition Wilson et al. [90] show that handling blood products has recently become increasingly expensive. For instance, he cites statistics from the Canadian Institute for Health Information showing that the total expenditures of the Canadian Blood Services increased 51% in 2001-02 compared to an increase in health care costs of 25% on average. Furthermore, the 2005 Blood Collection and Utilization report states 8.4% of surveyed hospitals in the US reported that elective surgery was postponed on one or more days in 2004 due to blood inventory shortages. In addition, more than 1.5 million components of blood platelets are transfused each year in the US (Sullivan et al. [80]), while at the same time 17% of platelet units collected in the US were outdated in 2004. Recently Landro [51] reports a decrease in overall blood collection in 27% of the US blood centers because of the swine-flu pandemic, and she emphasizes the blood centers' plans to allocate blood to the sickest patients. All of these statistics show that it is crucial that society should find ways to reduce costs and improve utilization within blood supply chains, so as to make the best use of limited blood resources.

Blood inventory management has been extensively studied in the OM literature. We refer readers to the survey papers of Nahmias [65], Prastacos [73], Pierskalla [71] and Karaesmen et al. [45]. Blood products are prototypical examples of age-based perishable products for which consumers have specific preferences, differentiating by product age. However, papers with age-differentiated demand in multi-product, multi-period settings are limited in the literature. An exception is Haijema et al. ([37], [36]), who analyze the perishable inventory problem of blood platelets, which are the most expensive and the most perishable blood product, having only four to six days of shelf life. In their study, demand for "young" platelets come from oncology and hematology, while demand for "any" aged blood comes from traumatology and general surgery. They use a combined Markov Decision Process (MDP) and simulation approach to find near-optimal heuristics in their setting. Recently, Kopach et al. [48] construct a red blood cell inventory management system with two demand rates (urgent/ non-urgent). They use a queueing model with simulation to compare different control techniques using data from the Canadian Blood Services.

We consider an age differentiated product with three periods of lifetime, under an heuris-

tic inventory policy, NIS (see Karaesmen et al. [45]). This policy orders a constant amount of New items in each period; this is a very common inventory replenishment practice in grocery stores and blood banks, and has also been shown to be effective in the literature (Deniz et al. [27]). Our model includes Upward and Downward substitution, in which Old blood platelets could be given to New blood platelet demand in Upward substitution, and vice versa in Downward substitution. Haijema et al. ([37], [36]) refer to Upward substitution as mismatch; according to the industry example in Haijema et al. ([37], [36]) from a Dutch blood bank, Upward substitution is very common in practice. Our model includes (possibly negative) costs for substitution, as well as (positive) costs if blood is outdated or if demand is left unsatisfied. Moreover, as blood banks and hospitals have limited space and blood must be refrigerated, we include an inventory cost.

Historically blood platelet transfusion was shown to reduce death from bleeding in patients with acute leukaemia in the 1950s (Hersh et al. [39]). Since then, transfusion of platelets has grown to be a significant part of treatment of conditions such as cancer, organ transplant, haematopoietic stem cell transplantation, marrow failure, AIDS, hepatitis, cardiovascular surgery and traumatology [79]. In addition to the treatments mentioned in Haijema et al. [37], in which oncology patients request the freshest platelets and traumatology patients have no preference on the age of the blood platelet, Fontaine et al. [30] reports that platelet demand from organ transplant operations is realized a couple of hours in advance of the operation and it is treated as an emergency receiving the freshest blood platelets. Moreover, a recent review of blood platelets and liver transplantation stated that "platelets are critically involved in liver injury and in liver generation via serotonin-mediated mechanisms" [70]. Therefore, our model with three age-differentiated demand streams can be seen as a model for practice in which transplant patients need new-aged platelets (3 periods of shelf time), cancer and hematology patients need medium aged platelets (2 periods of shelf time) (or fresher), and traumatology patients have no preference on the age of blood platelets. Note that in the current study of Fontaine et al. [30] in the Stanford Blood Center, the hospital blood bank doesn't release blood platelets for two days after donation due to testing. Therefore, the practical shelf life of blood platelets are often three days, and our modeling framework that considers blood platelets with three periods of lifetime fits the practice in such a blood bank.

With respect to the transfusion of blood platelets, there is no cross-matching of blood types unless the blood platelets contain a significant amount of red blood cells [79]. Currently

in transfusion medicine, plateletphresis, which is an automated method to separate platelets from other whole blood components, allows the collection of blood platelets, while returning plasma and red blood cells to the donor (The unit cost of producing blood platelet by plateletphresis is 538.72 on average [2]). Regarding transfers from other blood banks in case of inventory shortages, Brodheim et al. [15] considers a cycle stock model for red blood with life time of m periods; then, Prastacos [73] develops an allocation policy of red blood cells in a regional blood bank with n locations. His objective is to minimize expected average shortages and expected average outdates in this region. Fontaine et al. [30] reports that there is no trade of blood platelets; however, hospitals are able to get the freshest platelets from other blood centers or hospitals in the time shortage. Thus, modeling a single product, blood platelets at a single location, without considering different blood types in our study agrees with transfusion practice.

One issue confronted in blood banks is the need to maintain stocks of fresh inventory. To protect such inventory, we introduce the critical level policy to the perishable inventory literature, borrowing it from the spare parts literature. This policy protects the newest items against excessive downward substitution, possibly leaving some demand for the oldest blood unsatisfied in order to reserve the freshest blood for future demand requesting younger blood platelets (like transplants or oncology). Our main goal is to evaluate the effectiveness of using the NIS heuristic along with a critical level, as compared to the "near optimal" but more complex policies in the literature.

As mentioned above, this protection level policy is related to the critical level inventory policies for single product and multiple demand systems in spare parts inventory management (Veinott [85], Topkis [84], Deshpande et al. [28], Arslan et al. [6], Dekker et al. [26], Kranenburg and van Houtum [49], Zhao et al. [92]). In this research stream, there are multiple demand streams that are prioritized, a single non-perishable product of a single age in inventory, and the goal is to set a reorder level along with a critical number; when the inventory falls below this critical number the low priority demand will not be served. For the first time, we introduce a critical level type inventory policy to multi-product, multiperiod perishable inventory systems. Considering the transfusion practice given above and our modeling framework comprised of three age-differentiated groups of blood platelets, protecting New-aged blood platelets against excessive demand of Old-aged platelets could help, because protecting New-aged platelets for a period (day) against excessive demand from the traumatology department may allow serving more cancer patients in the future.

In addition to the efficient utilization of blood platelet inventories, Moroff [63] from the American Red Cross explains another crucial issue in the transfusion of platelets. He points out the substantial increase in usage of blood platelets over the last 15-20 years in the US, because of the enhanced supportive care required by cancer patients and the use of stem cell transplants. He also states that the total demand from patients receiving blood platelets is increasing, because of the aging population and "aggressive medical practices." The demand for blood platelets is increasing in Europe, too. Condon [24] reports an over 50% increase in demand between 2001 and 2006 in Ireland. Another similar demand increase for platelets recently caused critical shortage for cancer patients in Scotland in 2009 (Moss 2009). Such shortages may also come about due to factors such as epidemics and disasters, that can reduce the overall platelet supply dramatically. For instance, Landro [51] reports that 27% of the US blood centers faced reductions in overall blood collections due to the swine-flu pandemic. Therefore, hospital blood bank managers increasingly face capacities on platelet shipments from blood centers. In this study, we show that our protection policy may be particularly helpful when there is a capacity on platelet supply.

In the next section, we introduce the details of the modeling framework of our blood platelet inventory management problem with substitution, protection levels and capacities.

1.2. The model

Considering the transfusion practice in US blood banks, the actual shelf life of platelets is about three days [30]. Therefore, in our model we have three age-differentiated stocks of blood platelets: New, Medium and Old. At the beginning of every period, a fixed number of the newest blood platelets are ordered from a capacitated supply with no lead-time. This assumption is reasonable since blood platelet donors typically donate platelets regularly according to a schedule [30]. Unused blood from the previous period ages: New platelets become Medium-aged, Medium-aged become Old platelets, and Old platelets become outdated.

In the critical level protection model, we protect the newest blood platelet against excessive demand of Old items. In other words, there is no limit on substitution from New-aged item inventory to Medium-aged item demand or Medium-aged inventory to Old-aged item demand. (There is no need to protect *n*-aged items against excessive n+1-aged item demand since *n*-aged item will age to n+1 in the next period and we are in a discrete-time framework with all demand filled at the end of a period.) The demand process is discrete and nonnegative for all periods. We assume that the demand process is iid for each demand stream denoted by D_3 , D_2 and D_1 , as New, Medium and Old respectively (the subscript denotes the number of periods of lifetime remaining). Demand processes that are not iid still fit our modeling framework, but they would make the transition probabilities and one-period cost expressions more complicated.

The shortage, inventory holding and outdating costs are denoted as p_i , h_i for i = 1, 2, 3, and m respectively. In addition to considering outdating costs, we include shortage and substitution costs. As an example, the demand to New-aged blood platelets, which is primarily needed for organ transplants, are realized a couple hours before the operation [30]. If such blood is unavailable and the transplant is postponed, we capture this in shortage costs. If such donors are unavailable, but instead older blood platelets are available for use, our modeling framework captures this situation in substitution costs. We denote downward substitution costs as α_D^{AB} when substituting A to B for $AB = \{NM, NO, MO\}$; and we use α_U^{EF} for the Upward substitution when substituting E to F for $EF = \{MN, ON, OM\}$. Finally, we use S to denote the (constant) New blood platelet inventory level before demands are realized. Recall that we assume infinite supply of platelets (S may be constrained and we analyze this capacitated case in Section 1.6) and zero lead time.

We can thus represent our model's states as (i, j), where *i* and *j* denote the inventory levels of Medium and Old blood platelets before demand is realized: Our inventory control problem is a discrete Markov Chain (MC) with $(S + 1)^2$ states. Since our discrete MC is positive recurrent, we can easily find the limiting probabilities, π_{ij} for all i, j = 0, 1, ..., S. Then, we can represent the expected cost as

$$\sum_{i=0}^{S} \sum_{j=0}^{S} \pi_{ij} C_{ij} \tag{1.1}$$

where π_{ij} is the limiting probability and C_{ij} is the one-period cost of state (i, j).

We give examples of the cost expressions below. Note that these costs are based on substitution assumptions which specify the substitution priorities. In the expressions we use the following assumptions:

• Excessive demand of New items gets priority for substitution over excessive Medium or Old item demand.

- Excessive Medium-item demand gets priority for substitution over excessive Old item demand.
- Inventory of a specific age satisfies its own demand to the extend possible before being used for substitution.
- We substitute items of the "closest" available age.

For instance, if we are short New items, we substitute from Medium first and then from Old item inventory, if both have excessive inventory. However, if we are short Medium items, we substitute from Old items first, only after satisfying excessive New item demand from excess Old item inventory. Similarly, if we are short both New and Old items, we satisfy New item demand first, and then Old. Other substitution rules of course are possible. These will change our specific cost expressions, but not our general solution framework.

1.2.1. No protection level (c=0)

Our model without protection level extends the NIS policy studied in Deniz et al. [27] from two periods of life time to three periods. In addition to the limiting probabilities, we need to calculate expected one-period costs, C_{ij} , for each state (i, j). The expected one-period cost, C_{ij} , is the summation of expected one-period shortage cost, expected one-period inventory cost, expected one-period outdating cost and expected one-period substitution cost.

Limiting probabilities

Recall that state (i, j) corresponds to inventory level of *i* for Medium-aged platelets and *j* for Old-aged platelets after demand is realized. The limiting probabilities, π_{ij} , are formulated for fixed order level, *S*, in four different cases of the state space: (0,0), (0,k), (k,0) and (i,j)for $1 \le k \le S$ and $1 \le i, j \le S$. First, π_{00} is

$$\pi_{00} = \sum_{i=0}^{S} \sum_{j=0}^{S} \pi_{ij} \quad [\Pr(D_3 + D_2 + D_1 \ge S + i + j) + \Pr(D_3 \ge S, D_3 + D_2 \ge S + i, D_3 + D_2 + D_1 < S + i + j)].$$

There are two possible ways of transferring from state (i, j) to state (0, 0) depending on the total demand. The first term shows the case when the total demand exceeds the total available inventory. In addition, the second term is for the case when the total demand is less than the total available inventory. Since our perishable product, blood platelet, ages for each period and outdates after Old-age (3 days), $D_3 \ge S$ and $D_3 + D_2 \ge S + i$ finish all New-aged and Medium-aged inventory. Therefore, state (i, j) goes to state (0, 0) even if there are left-over Old-aged platelets after demand is realized, because these left-over Old platelets outdate.

For $1 \leq k \leq S$, the limiting probability, π_{0k} , is

$$\pi_{0k} = \sum_{i=k}^{S} \sum_{j=0}^{S} \quad \pi_{ij} \quad [\Pr\left(D_3 \ge S, D_3 + D_2 = S + i - k, D_1 \le j\right) \\ + \quad \Pr\left(D_3 \ge S, D_3 + D_2 + D_1 = S + i - j - k, D_1 > j\right)].$$

There are two possible different transitions from (k, j) to (0, k) for $1 \le k \le S$ and $0 \le j \le S$ depending on whether there is substitution for Old platelets. Note that there is no transition from state (i, j) to (0, k) for $0 \le i < k$ and $1 \le k \le S$, because transition to state (0, k)requires at least k amount of Medium-aged platelet inventory before demand is realized.

For $1 \leq k \leq S$, the limiting probability, π_{k0} , is

$$\pi_{k0} = \sum_{i=0}^{S} \sum_{j=0}^{S} \pi_{ij} \quad [\Pr\left(D_3 = S - k, D_2 + D_1 \le i + j, D_2 \ge i\right) \\ + \quad \Pr\left(D_3 = S - k, D_2 + D_1 = i + j, D_2 < i\right) \\ + \quad \Pr\left(D_3 + D_2 + D_1 = S + i - j - k, D_3 < S - k\right)].$$

There are three different possible transitions to state (k, 0) from state (i, j) for $1 \le k \le S$ and $0 \le i, j \le S$. Since we need to go state (k, 0), there should be exactly k amount of New-aged platelet left-over inventory after demand is realized. If $D_3 < S - k$, then extra New platelet inventory exceeding k is used for downward substitution to excessive older platelet demands, and the total demand should be equal to S + i - j - k.

Finally, the rest of the limiting probabilities, π_{ij} for $1 \leq i, j \leq S$, are calculated by

$$\pi_{ij} = \sum_{k=j}^{S} \sum_{l=0}^{S} \pi_{kl} \quad [\Pr\left(D_3 = S - i, D_2 = k - j, D_1 \le l\right) \\ + \quad \Pr\left(D_3 = S - i, D_1 + D_2 = k + l - j, D_1 > l\right)].$$

Note that there is no transition from state (k, l) to (i, j) for $0 \le k < j$ and $0 \le l \le S$, because transition to state (i, j) requires at least j amount of Medium-aged platelet inventory before demand is realized. Hence, two possible transitions to state (i, j) from (k, l) exist depending on the value of Old platelet demand, D_1 .

Shortage costs

There are three unit shortage costs: p_3 , p_2 and p_1 . In state (i, j), the expected one-period shortage cost for New platelets is

$$p_{3}E\left[\left(\left(D_{3}-S\right)^{+}-\left(i-D_{2}\right)^{+}-\left(j-D_{1}\right)^{+}\right)^{+}\right].$$
(1.2)

Note that implicit in this expression is the assumption that each stock of inventory serves its demand first, and any left-over inventory is first used to satisfy demand for New items.

The expected one-period shortage cost for Medium-aged platelets is

$$p_2 E\left[\left((D_2 - i)^+ - (S - D_3)^+ - \left((j - D_1)^+ - (D_3 - S)^+\right)^+\right)^+\right].$$
 (1.3)

Similarly implicit in this expression is the assumption that each stock of inventory serves its demand first, and any left-over inventory is first used to satisfy demand for New items, then excessive Medium item demand. Hence, $((j - D_1)^+ - (D_3 - S)^+)^+)$ states that excessive Medium platelet demand is satisfied from left-over Old platelet inventory only after these left-over Old platelets are used for excessive New platelet demand.

Finally, the expected one-period shortage cost for Old platelets is

$$p_1 E\left[\left(\left(D_1 - j\right)^+ - \left(S + i - D_2 - D_3\right)^+\right)^+\right].$$
(1.4)

Inventory holding costs

As there are three ages of blood platelets and outdating of left-over Old platelets, there are only two different inventory costs: New-aged and Medium-aged. Left-over New platelets are refrigerated and age to Medium in the next period; thus in state (i, j), the expected one-period inventory holding cost for New platelets is

$$h_3 E\left[\left((S-D_3)^+ - (D_1+D_2-i-j)^+\right)^+\right].$$
(1.5)

Note that this expression implicitly assumes Medium and Old inventories are used in substitution for each other before New items are substituted. Similarly we can find the expected one-period inventory holding cost for Medium platelets:

$$h_2 E\left[\left((i-D_2)^+ - (D_3-S)^+ - (D_1-j)^+\right)^+\right].$$
(1.6)

Outdating cost

There is no inventory cost for Old platelets, because left-over blood platelets after three periods (5-6 days from donation) become medical waste in transfusion practice. Therefore, the hospital blood bank manager incurs an outdating/waste cost for these left-over Old platelets. The expected one-period outdating cost for state (i, j) is

$$mE\left[\left((j-D_1)^+ - (D_2-i)^+ - \left((D_3-S)^+ - (i-D_2)^+\right)^+\right)^+\right].$$
 (1.7)

Again this expression incorporates our substitution assumptions that excessive demand for New platelets is first satisfied from Medium item inventory, then Old platelet left-over inventory.

Downward substitution costs

This (possibly negative) mismatching cost is incurred when traumatology patients (Old platelet demand) are treated with fresher platelets (New or Medium). Similarly treating oncology patients with the freshest (New) blood platelet incurs a downward substitution cost. Thus, there are three different downward substitution costs in our modeling framework: New to Medium, New to Old and Medium to Old. One may wonder why the blood bank manager would incur any (positive) cost in satisfying demand needing Old platelets with fresher platelets. The supply of blood platelet inventory is often constrained, so the order size of the freshest platelets is limited. Therefore, there is an opportunity cost in using platelets and the hospital blood bank manager should incur positive downward substitution costs.

Explicitly, the cost of downward substituting New items to satisfy excessive demand of Medium items is

$$\alpha_D^{NM} E\left[\min\left\{ \left(S - D_3\right)^+, \left(\left(D_2 - i\right)^+ - \left(j - D_1\right)^+\right)^+ \right\} \right].$$
(1.8)

In this expression, $((D_2 - i)^+ - (j - D_1)^+)^+$ states that New to Medium downward substitution happens if there is still extra Medium platelet demand after Old to Medium Upward substitution, because of our assumption on substitution priorities.

For the downward substitution from New to Old, the expected one-period cost is

$$\alpha_D^{NO}E\left[\min\left\{\left((S-D_3)^+ - (D_2-i)^+\right)^+, \left((D_1-j)^+ - (i-D_2)^+\right)^+\right\}\right].$$
(1.9)

In any demand realization and inventory state, excessive demand for Old platelets is first satisfied from left-over Medium platelet inventory (Medium to Old downward substitution), then from left-over New platelet inventory (New to Old downward substitution). Hence, $((D_1 - j)^+ - (i - D_2)^+)^+$ represents the amount of excessive demand for Old platelets need-ing left-over New platelets. However, extra demand for Medium platelets are satisfied first from this left-over New platelet inventory (New to Medium downward substitution); then, excessive demand for Old platelets could use New items if there are such available left-overs.

Considering the last downward substitution, the expected one-period cost of Medium to Old in state (i, j) is

$$\alpha_D^{MO} E\left[\min\left\{\left((i-D_2)^+ - (D_3 - S)\right)^+, (D_1 - j)^+\right\}\right].$$
(1.10)

Note that excessive demand for New platelets has priority on left-over Medium platelet inventory over extra Old platelet demand. Hence, Medium to New Upward substitution occurs before Medium to Old downward substitution.

Upward substitution costs

The mismatching cost used in Haijema et al. [37] corresponds to Upward substitution cost in our model. Since patients needing fresh blood platelets are treated with older platelets, Haijema et al. [37] reports that patients often suffer from this mismatching. In our model, there are three different Upward substitution costs: Medium to New, Old to New and Old to Medium. The expected one-period Medium to New Upward substitution cost is

$$\alpha_U^{MN} E\left[\min\left\{ (D_3 - S)^+, (i - D_2)^+ \right\} \right].$$
(1.11)

Note that excessive New platelet demand is satisfied from left-over Medium platelet inventory regardless of Old platelet demand because of the substitution priority.

As for Old to New Upward substitution, the expected one-period cost is

$$\alpha_U^{ON} E\left[\min\left\{\left((D_3 - S)^+ - (i - D_2)^+\right)^+, (j - D_1)^+\right\}\right].$$
(1.12)

Similar to Medium to New, excessive New platelet demand has priority over left-over Medium platelet inventory.

Finally, the expected one-period Old to Medium Upward substitution cost is

$$\alpha_U^{OM} E\left[\min\left\{ (D_2 - i)^+, \left((j - D_1)^+ - (D_3 - S)^+ \right)^+ \right\} \right].$$
(1.13)

Note that excessive Medium platelet demand is satisfied from left-over Old platelet inventory after this left-over inventory is used for excessive New platelet demand because of the substitution priorities.

1.2.2. Positive protection level (c>0)

In our heuristic for the protection level model, we set a fixed level, c; limiting New platelet substitution to Old platelet demand: We only permit New to Old substitution when the New inventory level, after satisfying New and excessive Medium demand, is greater than c. If we are short Medium platelets, there is no protection against satisfying this excessive demand by New item inventory, since left-over New platelets will age to Medium in the next period. The limiting probabilities, π_{ij} , and costs, C_{ij} , change slightly with the addition of the protection level, c. In the C_{ij} expressions, only the shortage cost of Old items, the inventory cost of New items and the downward substitution cost of New to Old will change, because protecting unsold New-item inventory against excessive demand of Old-item affects substitutions involving these quantities.

Limiting probabilities

Compared to the limiting probabilities of the model without protection level, only π_{i0} 's for $0 \leq i \leq c$ change, because in other states the protection against excessive Old platelet demand has no effect. Firstly, the limiting probability of state (0, 0), π_{00} , is

$$\pi_{00} = \sum_{i=0}^{S} \sum_{j=0}^{S} \pi_{ij} \quad [\Pr\left(D_3 \ge S, D_2 + D_1 \ge i+j\right) + \Pr\left(D_3 < S, D_3 + D_2 \ge S+i\right) \\ + \Pr\left(D_3 \ge S, D_2 \ge i, D_2 + D_1 < i+j\right)].$$

Note that the transition from state (i, j) to (0, 0) simply depends on the total demand of New and Old platelets when $D_3 \ge S$. Since there is a positive protection level in the model, transition to state (0, 0) is affected when $D_3 \le S$. Therefore, $D_3 + D_2 \ge S + i$ ensures that there is no left-over New platelet inventory even with the protection level.

Considering the transition to state (k, 0) for $1 \le k < c$, there is no downward substitution from left-over New platelet inventory to excessive Old platelet demand; because left-over inventory is already lower than c after D_3 is realized and excessive Medium platelet demand is satisfied from any left-over New-item inventory. Then, π_{k0} for $1 \le k < c$, it is

$$\pi_{k0} = \sum_{i=0}^{S} \sum_{j=0}^{S} \pi_{ij} \quad [\Pr(D_3 = S - k, D_2 < i, D_2 + D_1 = i + j)]$$

+
$$\Pr(D_3 = S - k, D_2 + D_1 \le i + j, D_2 \ge i)$$

+ $\Pr(D_3 + D_2 + D_1 = S + i - j - k, D_3 < S - k, D_2 > i)$
+ $\Pr(D_2 > i, D_1 > j, D_3 + D_2 = S + i - k)].$

In the transition probability, the first term represents the case when there is only downward substitution from left-over Medium platelet inventory to excessive Old platelet demand and a total $i-D_2$ amount of left-over Medium platelet inventory is substituted. In contrast to the first term, the left-over Old platelet inventory is used by excessive Medium platelet demand in the second term of the transition probability. As for the third term, the left-over New platelet inventory beyond k is used by excessive Medium platelet demand, because there is no protection against New to Medium downward substitution. Finally, the fourth term is almost the same as the third term except for excessive Old platelet demand; however there is no downward substitution from New to Old since k < c.

The transition to state (c, 0) is very similar to the transition to state (k, 0) for $1 \le k < c$. Then, π_{c0} is

$$\pi_{c0} = \sum_{i=0}^{S} \sum_{j=0}^{S} \pi_{ij} \quad [\Pr\left(D_3 = S - c, D_2 < i, D_2 + D_1 = i + j\right) \\ + \quad \Pr\left(D_3 = S - c, D_2 + D_1 \le i + j, D_2 \ge i\right) \\ + \quad \Pr\left(D_3 + D_2 + D_1 = S + i - j - c, D_3 < S - c, D_2 > i\right) \\ + \quad \Pr\left(D_2 > i, D_1 > j, D_3 + D_2 = S + i - c\right) \\ + \quad \Pr\left(D_3 + D_2 + D_1 \ge S + i - j - c, D_3 < S - c, D_2 \le i\right)].$$

Note that the first four terms of the transition probability from state (i, j) to (c, 0) are the same as the previous case, π_{k0} , when k = c. The fifth term of the transition probability represents the case when c amount of left-over New platelet inventory is protected against excessive Old platelet demand.

Costs

Recall that only shortage cost for Old platelets, inventory holding cost for New platelets and New to Old downward substitution cost change for the model without protection level. The expected one-period shortage cost for Old platelets becomes

$$p_1 E\left[\left((D_1 - j)^+ - \left((i - D_2)^+ - (D_3 - S)^+\right)^+ - \left((S - D_3)^+ - (D_2 - i)^+ - c\right)^+\right)^+\right].$$
(1.14)

Note that $((i - D_2)^+ - (D_3 - S)^+)^+$ represents how many left-over New platelet units are used for excessive Medium platelet demand because of the substitution priority and no protection on New to Medium downward substitution. Then, $((S - D_3)^+ - (D_2 - i)^+ - c)^+$ ensures that a restricted amount of left-over New platelet inventory is substituted to satisfy excessive Old item demand up to the protection level, c.

Considering positive protection levels in our model, the hospital blood bank manager stochastically carries more left-over New platelets. Hence, the expected one-period inventory holding cost of New platelets increases to

$$h_3 E\left[\left((S-D_3)^+ - (D_2-i)^+ - \gamma\right)^+\right],$$
 (1.15)

where $\gamma = \min\left\{\left((S - D_3)^+ - (D_2 - i)^+ - c\right)^+, \left((D_1 - j)^+ - (i - D_2)^+\right)^+\right\}$. Similar to the shortage cost for New platelets, $\left((S - D_3)^+ - (D_2 - i)^+ - c\right)^+$ enables the manager to protect some left-over New platelet inventory against excessive Old platelet demand and carry them as Medium platelet to the next period.

Finally, the New to Old downward substitution cost becomes

$$\alpha_D^{NO} E\left[\min\left\{\left((S-D_3)^+ - \left((D_1-j)^+ - (i-D_2)^+\right)^+\right)^+, c\right\}\right].$$
(1.16)

Again the protection level, c, in the expression just ensures that excessive Old platelet demand won't be satisfied after the threshold value, c, of left-over New platelet inventory.

1.3. Computational Complexity

Recall that our modeling framework aims to improve the decision making process of the hospital blood bank manager, whose objective is to minimize the expected cost. In her decision process, she has to decide a fixed order level, S, and the protection level, c, so as to minimize:

$$\sum_{i=0}^{S} \sum_{j=0}^{S} \pi_{ij} C_{ij}$$

For every, S and c, we need to compute π_{ij} and C_{ij} for $(S + 1)^2$ states. To calculate both π_{ij} and C_{ij} , we first need to calculate the transition probabilities. Denote by M_3 the maximum demand the hospital can realize for New-aged blood platelets in a period, and M_2 and M_1 ; these values are for Medium-aged and Old-aged items, respectively. For

Table 1.1: Values of cost variables used in comparison results

Variable	Value
Shortage	750
m	150
h_3	1
h_2	1
Mismatching	200

simplification, we can assume $M_1 = M_2 = M_3 = M_d$. For a given S and c, we can calculate π_{ij} 's in a couple of seconds. However, significant computational effort is needed to compute the C_{ij} 's. For instance, for a fixed S and c the complexity is on the order of $O(M_d^3)$ for each (i, j) to compute a C_{ij} , because we must calculate the expectations over D_1 , D_2 and D_3 for each i, j.

Considering an exhaustive search on the order and protection levels, we need to compute $(S + 1)^2$ number of C_{ij} 's for every S and c since no structural results are apparent. Since there are total of $(S_{max} - 1) S_{max}/2$ different (S, c) pairs, we need to calculate costs, C_{ij} , that many times. Notice that, $S_{max} = 3M_d$ in the worst case scenario. Therefore, the total complexity for calculating C_{ij} 's is in order of $O(1.5M_d^4(3M_d - 1))$. In a realistic scenario assuming $M_d = 100$, we need to calculate around 4.5×10^{10} different C_{ij} 's (the overall complexity is 3×10^{13} in Haijema et al. [37]). In a small size example, $M_d = 5$, the total time to calculate expected cost is over two hours on a dual-core Intel Pentium computer. In light of this complexity, we use simulation to compare our heuristic with the existing near-optimal heuristics in the literature [37]. Thus, we run a sample path for fixed S and c over 1 million periods to calculate costs directly.

1.4. Numerical Results

In this section, we compare our protection level policy with existing policies presented in Haijema et al. [37] using real cost data from a Dutch blood bank. Then, we analyze the robustness of our results to different demand models, and summarize the corresponding managerial insights. In Section 1.5, we perform sensitivity analysis about our twelve cost parameters; then we analyze the capacitated order level case in Section 1.6.

Table 1.1 summarizes the unit cost values given in Haijema et al. [37]. Recall that we modify their model to fit our framework: We assume that there are three age differentiated demand streams instead of two. We initially adopt their costs. Their "mismatching cost"

I I I I I I I I I I I I I I I I I I I			J .	
	NIS	NISprot	1D	2D
S^*	12	12	17	(17, 14)
c^*	0	0	0	0
$cost^*$	486.75	486.75	506.60	482.65
Shortage	37.53	37.53	20.77	16.32
Holding	7.65	7.65	5.93	6.41
Outdating	307.99	307.99	164.60	194.77
Downward substitution	102.62	102.62	185.46	149.35
Upward substitution	30.96	30.96	129.83	115.80
Total substitution	133.58	133.58	315.30	265.15

Table 1.2: Comparison with cost data from Haijema et al. [37]

represents both Upward and downward substitution costs in our model such that

$$\alpha_D^{NM}=\alpha_D^{NO}=\alpha_D^{MO}=\alpha_U^{MN}=\alpha_U^{ON}=\alpha_U^{OM}=200.$$

In addition to this, all shortage costs are the same for all ages:

$$p_3 = p_2 = p_1 = 750$$

As for the demand process, we choose a Poisson process for all demand streams with mean 7, 2 and 1 for New, Medium and Old platelet demands, respectively. Haijema et al. [37] report 70% of demand the blood bank in their study is for "young" and 30% is for "any" blood platelets. Since our modeling framework separates "any" demand into Medium and Old-aged platelets, we initially choose 20% for Medium and 10% for Old-aged platelets. Later we will change our demand stream to get insights about the impact of demand structure on the protection level policy.

Table 1.2 shows the costs for four different policies: NIS, NIS with positive protection level and the 1D and 2D policies from Haijema et al. [37]. 1D policy consider one orderup-to level for total inventory and 2D policy takes both total inventory and freshest platelet inventory into account. However, we only focus on order-up-to level for freshest platelet inventory since we extend NIS policy with protection levels. Deniz et al. [27] shows that NIS policy is efficient and widely used in transfusion practice and grocery stores. The first three rows represent the optimal order level, S^* , the optimal protection level, c^* , and the minimum cost of a sample path of one million periods. In this table, (17, 14) represents the total order level and order level of New-items for the 2D heuristic policy; the cost values are in terms of million. According to the simulation results, NIS with positive protection level

_			L L	<u> </u>
	NIS	NIS prot	1D	2D
S^*	12	12	17	(17, 13)
c^*	0	2	0	0
$cost^*$	478.84	451.47	504.71	460.56
Shortage	29.29	8.63	18.87	9.69
Holding	7.65	10.91	5.93	6.66
Outdating	307.99	358.07	164.29	112.43
Downward substitution	103.03	46.43	185.60	279.84
Upward substitution	30.88	27.43	130.02	51.93
Total substitution	133.91	73.86	315.63	331.77

Table 1.3: Comparison with modification of shortage costs

is not beneficial; it performs worse than 2D and NIS without protection level policies. This result is intuitive. Because there is no priority difference between the three age-differentiated demand streams ($p_3 = p_2 = p_1 = 750$), there is no benefit to reserving New platelets for future excessive Medium platelet demand. The major cost differences between the NIS and 1D policies are outdating, upward substitution and total substitution costs. The NIS policy has almost twice the outdating cost as 1D and 2D. However, 1D and 2D have significantly higher substitution costs than NIS. In addition, NIS has more holding cost than 1D and 2D policies. These results all arise from the same behavior: Since both 1D and 2D policies take the total inventory level into account, unlike NIS, they are more adaptive to higher inventory levels than NIS and tend to hold less inventory. Therefore, we observe more substitution costs and less outdating and inventory holding costs in 1D and 2D than NIS. Note that 2D has higher downward substitution than 1D because of the extra order-up-to level for the freshest platelets. The trade off between 1D and 2D is paying more on downward substitution cost and saving on outdating cost, because 2D will order to bring new platelets up to 14 even if this causes the total inventory exceeds 17.

In order to effectively apply a "critical level" type policy in our modeling framework, we need different priorities across different demand streams. Furthermore, as the demand from organ transplants and oncology patients are typically more important than traumatology patients because of the emergency or risk, it is reasonable to assume $p_3 > p_2 > p_1$. Thus in a second simulation run, we choose $p_3 = 1000$, $p_2 = 750$ and $p_1 = 300$; these results are presented at Table 1.3. In this case NIS with a protection level of two outperforms all other policies when the hospital blood bank manager has different shortage costs for different aged blood platelets. The protection level policy has the highest holding and outdating costs because of carrying protected left-over New platelet inventory. However, the protection level policy has lower substitution and shortage costs. Incurring lower substitution cost is intuitive since reserving freshest platelets limits the substitution. However, lower shortage cost in a protection level policy is not obvious since rejecting excessive Old platelet demands incurs a shortage cost. In fact, NIS with protection level of two has more shortage cost for Old items than NIS without protection level. On the other hand, the protection level decreases shortage costs for both New and Medium items because it reserves inventory. Since there is more social cost of losing demand from organ transplants and cancer patients, the protection level policy is clearly shown in Table 1.3 as a decrease in shortage and substitution costs. Even if the protection policy increases outdating and holding costs, the benefit from substitution and shortage costs is larger for this case.



Figure 1.1: Different demand models

We now analyze the robustness of our results to the demand structure. For instance, some hospitals in the US serve primarily oncology patients (i.e. cancer treatment centers) and these hospitals' blood bank managers might face more demand from oncology patients. Figure 1.1 shows cost comparisons of the four different policies in four different demand settings with two different shortage cost settings. "Mostly New" corresponds to demand for 60% New, 10% Medium, 30% Old; "Balanced" corresponds to demand for 30% New, 40% Medium, 30% Old; "Mostly Medium" corresponds to demand for 10% New, 60% Medium, 30% Old; "Mostly Old" corresponds to demand for 10% New, 30% Medium, 60% Old. The first plot corresponds the case when $p_3 = p_2 = p_1 = 750$. Intuitively there is no incentive to protect if age-differentiated demand stream is not prioritized, hence the protection level is zero. Except for the "Mostly Medium" case, NIS outperforms both 1D and 2D policies in all cases. For this case, since most demand is for Medium platelets, and 2D is a more adaptive policy than NIS, 2D benefits more on the outdating cost than NIS benefits from the substitution cost in "Mostly Medium" case.

In the second plot of Figure 1.1, we use $p_3 = 1000$, $p_2 = 750$ and $p_1 = 300$ to put more social cost on losing organ transplants and cancer patients. In all demand structures, NIS with positive protection level outperforms other policies because of prioritizing demand from high-risk patients. The cost gap between NIS with positive protection level and other policies increases as the proportion of demand for fresher platelets increases. This result is intuitive since the penalty cost of losing high-risk patients is high.

In our experiments in this section, we show that a protection level policy may be beneficial for a blood bank manager if the shortage cost of losing demand from high-risk patients is high; a protection level policy can decrease the substitution and holding costs significantly but increases the outdating and holding costs. However, the manager has no incentive to protect if there is no prioritization of high-risk patients over elective surgeries or traumatology patients. In this case, 1D and 2D may be better than the NIS policy because they are more adaptive than NIS. In the next section, we analyze the sensitivity of our results on the cost parameters used in the model and provide managerial insights.

1.5. Sensitivity analysis on the cost parameters

In the previous section we showed that a protection level policy may be beneficial for the hospital blood manager if she has different priorities for blood platelet demand from different departments in the hospital: Hence, ordered shortage costs, $p_3 > p_2 > p_1$, allows a protection level policy to possibly be superior. To explore the relative importance of other parameters, we perform a sensitivity analysis for the minimum cost of our four different policies on each of the 12 cost parameters: We calculate the minimum cost for each policy and vary the cost parameter of interest. In our base analysis, we use the original cost data shown in Table

1.1. The demand process is the same as the original data from the Dutch blood bank [37]: Poisson with mean 7 for New, 2 for Medium and 1 for Old. We performed similar sensitivity analyzes with different demand streams, but the conclusions were unchanged.



Figure 1.2: Shortage (p_3, p_2, p_1) and outdating (m) cost parameters

Three plots of Figure 1.2 show our numerical results conveying changes with respect to the shortage cost parameters. The plots for p_3 and p_2 have similar impact on the optimal policy: There is no protection level as p_3 and p_2 get closer to zero because of the decrease in value of protecting. But, for higher values of p_3 and p_2 the manager may choose to reserve the freshest platelets if the social cost of losing these patients is high enough; in these cases NIS with positive protection level policy outperforms other policies. However, there is no incentive to protect for high values of p_1 . Therefore, NIS with positive protection level outperforms other policies with high margin if p_1 gets closer to zero, because low p_1 takes away one of the downside of the protection level policy. The last plot in Figure 2 is for mand the result is similar to p_1 . Since the outdating cost is the other downside of a protection level policy, NIS with a positive protection level policy outperforms other policies as m gets closer to zero. But unlike the result in p_1 , 1D and 2D outperform NIS for large values of m. This result is intuitive, because both 1D and 2D are more adaptive policies than NIS, which allow them to better control outdating cost.



Figure 1.3: Substitution costs: α_D^{NM} , α_D^{NO} , α_D^{MO} and α_U^{MN}

Three plots of Figure 1.3 show our numerical results analyzing the Downward substitution parameters. The plots for all Downward substitution parameters have similar impact on the optimal policy: Since substitution cost is significant in both 1D and 2D, NIS outperforms both these policies in high values of these cost parameters. Conversely both 1D and 2D are better than NIS as these Downward substitution parameters get closer to zero, because incurring these substitutions become cheaper. This result is intuitive because both 1D and 2D are more adaptive than NIS and Downward substitution occurs more in order to reduce outdating. For high values of these cost parameters, the protection level policy outperforms the other policies the most in the case of α_D^{NO} , because saving substitution cost from New to Old substitutions is the advantage of the protection level policy and higher α_D^{NO} results increase the cost gap between our protection level policy and other policies. As for the first Upward substitution cost parameter, α_U^{MN} , plot, the protection level policy is better than the other policies with low values of this parameter, because there are stochastically more Medium item inventory in protection level policy and the manager may incur more substitution from Medium to New (This is a potential second-order benefit of protection). However, NIS without protection level outperforms other policies as α_U^{MN} increases. The cost lines are almost flat after a certain value of α_U^{MN} because there is very little Upward substitution in the optimal policies after this point.



Figure 1.4: Substitution $(\alpha_U^{ON}, \alpha_U^{OM})$ and holding costs (h_3, h_2)

The analysis with respect to the remaining two Upward substitution and holding cost parameters are shown in Figure 1.4. NIS outperforms both 1D and 2D in high values of α_U^{ON} but the protection level policy slightly outperforms NIS without protection level in low values of α_U^{ON} . This result is intuitive because the protection policy carries more inventory than NIS without a protection level and thus substitution from Old to New occurs is more commonly in the protection level policy. Regarding the other Upward substitution cost parameter α_U^{OM} , the 2D policy outperforms other policies for low values α_U^{OM} . Note 1D and 2D incur more Upward substitution costs – See Table 1.3. As for the holding cost parameters, h_3 and h_2 , in Figure 1.4, the protection level policy outperforms other policies with very low h_3 and h_2 ; but 2D outperforms the rest as h_3 and h_2 increase. These results are intuitive since the protection level policy carries more inventory. However, the holding cost is not a big cost factor in blood platelet inventory management; hence low values of h_3 and h_2 are realistic assumptions.

Summarizing our sensitivity analysis on the cost parameters, the protection level policy performs better with low values of m, p_1 , α_U^{MN} , h_3 and h_2 ; and high values of p_3 , p_2 , α_D^{NM} and α_D^{NO} . The performance comparison of the protection level and NIS without protection level depends on the cost of carrying more inventory, outdating and substitution. Because the protection level policy carries more inventory, it outdates more than NIS without protection level policy. On the other hand, the NIS without protection level policy pays more shortage and substitution costs than the protection level policy. As for the performance comparison of the protection level policy and the 1D and 2D policies, because the 1D and 2D policies pay more substitution cost and less to outdating costs than the protection level policy, these two trade-offs primarily determine the performance difference of these policies.

1.6. Capacity on order level

In our previous analysis, there was no capacity on the order level of blood platelets, S; thus, the hospital blood bank manager could order as many blood platelets as she wanted from the regional blood center. However, blood platelet supply is often limited and it is exposed to many risks: Increasing demand from cancer patients, epidemics and natural disasters. Recently Landro [51] reports a decrease in overall blood collection in 27% of the US blood centers because of the swine-flu pandemic, and she emphasizes the blood centers' plans to allocate blood to the sickest patients due to this reduction. Therefore, we analyze the performance of the protection level policy when there is a limit on New item inventory.

Figure 1.5 shows when the protection level policy outperforms not protecting with respect to the tightness of capacity of supply and downward substitution cost. In this simulation study, we set the cost parameters: $p_3 = 1000$, $p_2 = 750$, $p_1 = 300$, m = 150, $h_3 = h_2 = 1$ and $\alpha_U^{MN} = \alpha_U^{ON} = \alpha_U^{OM} = 200$. In addition, we use the following demand stream: Poisson with mean 7 for New, Poisson with mean 2 for Medium and Poisson with mean 1 Old. (Our results are robust for different demand models.) The horizontal axis represents the downward substitution cost from 'Low' (0) to 'Medium' (375) and 'High' (750) and the vertical axis represents the capacity on order level, S: 'Loose' (20) to 'Medium' (10) and 'Tight" (1). In Figure 1.5, the blue shaded region, 'Protection', shows the area in which the protection level policy outperforms the other policies.



Figure 1.5: When to protect if order level, S, is limited.

Different perishable goods may fall into different region on Figure 1.5. For instance, bananas and milk at grocery stores may always lie with the No protection region, because there is no such downward substitution cost and often the supply of groceries isn't constrained. However, blood platelets with limited supply may lie on the Protection region, as the opportunity cost of using fresher platelets is high. Furthermore, even if blood platelets with a low downward substitution cost are usually in the No Protection region, the protection level policy may perform better during a supply shortage from the regional blood bank (e.g. when supply is reduced due to epidemics or disasters). In this case protecting the freshest left-over platelets for future patients needing fresher units is beneficial. This result is very intuitive: The hospital blood bank manager would choose to serve tomorrow's cancer patients over today's elective surgeries because of the high shortage cost for oncology. Recall that capacity on blood platelets is in fact a major issue currently facing blood platelet supply chains [51]. Therefore, our protection policy may be a helpful managerial decision tool for the hospital blood bank manager in times when she has capacity on order levels from regional blood banks.

1.7. Conclusion

We consider a discrete-time inventory system for blood platelets that has distinct demand streams for product of different ages. In addition to inventory holding, outdating and shortage costs, our modeling framework includes substitution costs when a demand for a certainaged item is satisfied by a different-aged item. Since the decision maker in our problem is the hospital blood bank manager, our objective is to minimize the expected cost for the hospital over an infinite time horizon. We introduce the critical level policy to the perishable inventory literature for the first time, protecting the newest items against excessive downward substitution. This reserves these newest items for future demand for procedures needing fresher items. We model the problem as MDP and evaluate the costs of a common heuristic replenishment policy, NIS, (with and without a protection level) against "near optimal" policies from the literature (1D and 2D).

We show that NIS with positive protection level may outperform the other heuristics when the hospital blood bank manager has different shortage costs for different aged blood platelets. Since the blood platelets have three days of actual shelf life in practice, protecting some of the freshest platelets for one period, i.e. one day or shift of eight hours, against excessive Old-item demand from traumatology patients stochastically increases Mediumaged blood platelet inventory in the next period. This protection level policy could improve the hospital blood bank's performance when there is a significant difference between the shortage costs of different aged blood platelets: Protecting some New-aged platelets against excessive demand from traumatology patients in the current period directly helps the blood bank manager to satisfy demand of Medium-aged platelets from oncology and hematology operations in the next period, and indirectly satisfy demand for New-aged transplantation patients via substitution in the next period as well.

Considering our different costs, we perform a sensitivity analysis on twelve cost parameters to investigate the policy yielding the minimum expected cost. We find that NIS with positive protection level policy outperforms the rest of the policies for certain parameter settings: low values of cost parameters for outdating, holding, shortage for Old and Upward substitution from Medium to New; and high values of cost parameters for shortage for New, shortage for Medium, Downward substitution from New to Medium and New to Old. Since the protection level policy carries more inventory, and thus outdates more than NIS without a protection level policy, the performance difference between NIS with and without a protection level depends mostly on the costs of outdating and substitution. Similarly, the protection level policy's performance against 1D and 2D depends on these cost parameters, especially on the substitution costs. Because both the 1D and 2D policies are more adaptive, they incur more substitution costs than the protection level policy while paying less outdating cost. In other words, the "price of adaptiveness" of the 1D and 2D determine the value of the protection level policy to the hospital blood bank manager.

We show that our protection level policy may be particularly beneficial when the supply of blood platelets is tight. In addition, the protection level policy may be beneficial even if the downward substitution cost is very low in tight supply capacities. This result may be very useful for hospital blood bank managers to efficiently utilize their blood platelet inventory in the face of recent decreases in overall blood collection across the US: Protecting the freshest blood platelets for future demand needing fresher items is especially beneficial for the hospital blood bank manager if there is a capacity on order levels from regional blood banks.

In our study, we bring a "critical level" type policy to the multi-age differentiated product and multi-period perishable inventory literature considering one of the most perishable products, blood platelets. Our protection level policy can possibly improve the hospital blood bank manager's decision process and a future collaboration with a US blood bank would be tremendous opportunity to test our work. In additional future work, we are working on reducing the search space of optimal order and protection levels, and extending our modeling framework in a continuous time and nonstationary order level framework. As a final note, our protection level policy could be interesting for inventory management of other perishable products or those subject to obsolescence, i.e. electronics and groceries.
Chapter 2

Failure Probability of VARTA in Higher Dimensions

Vector-Autoregressive-To-Anything (VARTA) is a highly flexible model for driving largescale stochastic simulations by generating samples of stationary multivariate time series with arbitrary marginal distributions. The construction of this model relies on a stable vector autoregressive process with a positive definite autocorrelation matrix. We show that there exists multivariate time-series input processes for which the conditions of stability and positive definiteness are violated. We investigate the likelihood of this event with increasing number of component time-series processes and order of dependence by extending the onion method, which is used for sampling positive definite correlation matrices for random vectors, to sample positive definite autocorrelation matrices for multivariate time series. We find that the failure probability of VARTA reaches one with increasing number of component time series and order of dependence, but at a rate very much dependent on the rate of decay in temporal dependencies. We conclude with a discussion on an approximation of VARTA that might enable the simulation practitioner to avoid the failure of VARTA in high-dimensional settings. ¹

2.1. Introduction

An important step in the design of stochastic simulation is input modeling, i.e., modeling the uncertainty in the input environment of the system being studied. Input modeling is often characterized as selecting appropriate univariate probability distributions to represent the primitive inputs of interest, and it would indeed be this simple if the relevant input processes

¹Co-author: Bahar Biller

could be represented as a sequence of independent random variables having identical distributions. When such univariate models do apply, there are a number of software packages that support automated input modeling; good reviews are available in Vincent [86] and Law and Kelton [53].

However, these simple models fail to capture the stochastic properties of the input processes that exhibit multivariate and/or temporal dependencies that occur naturally in many service, communications, and manufacturing systems (see [62] and [87] for example studies). A close look at the existing input-modeling literature reveals that much of the previous work on time-series input processes for stochastic simulation is based on linear, univariate time-series models such as the autoregressive moving average process. However, Mallows [60] shows that the linearity of these models imply normal marginal distributions and there are many physical situations in which the marginals of the time-series processes are non-normal. This has led a number of researchers to model time series with marginals from exponential, gamma, geometric, or general discrete distributions. However, these models allow only limited control of the dependence structure and a different model is required for each type of marginal distribution [11].

A way to overcome these limitations is to construct the desired process by a monotone transformation of a Gaussian linear process. For example, Cario and Nelson ([18], [20]) take this approach to develop models for representing and generating stationary univariate time-series processes with arbitrary marginal distributions. The central idea is to transform a Gaussian autoregressive process, which Cario and Nelson call the base process, into the desired univariate time-series input process that they presume as having an Autoregressive-To-Anything (ARTA) distribution. The authors manipulate the autocorrelations of the Gaussian base process in order to achieve the desired autocorrelations for the input process. A similar transformation-based model suggested in the simulation inputmodeling literature is the Normal-To-Anything (NORTA) process designed specifically for random vectors [19]. It is constructed by simply transforming a multivariate Gaussian base random vector into the desired process via the use of the inverse cumulative distribution function (cdf). The most recent addition to these transformation-based family of methods is the Vector-Autoregressive-To-Anything (VARTA) process of Biller and Nelson [12] that simply pulls together the theory behind the ARTA process and the NORTA process and extends it to the multivariate time-series inputs.

Although the ARTA/VARTA processes are regarded as two of the highly flexible uni-

variate/multivariate time-series models in the current simulation input-modeling literature, they might fall short in the simulation of some stochastic systems for the following reason: There exist sets of marginal distributions with feasible dependence structures that are not representable by the ARTA/VARTA transformations. Both Li and Hammond [56] and Lurie and Goldberg [58] give examples where this appears to be the case for the NORTA transformation and Ghosh and Henderson [33] prove the existence of a joint distribution that is not representable as the transformation of the corresponding multivariate normal random vector. Although these studies focus on random vectors, similar results can be generalized to time-series input processes and this is what we aim to do in this paper. Biller and Nelson [8] have shown that if the autocorrelation matrix of the base vector autoregressive process is positive definite, then the autocorrelation matrix of the input process is positive definite. In this paper we show that the reverse of this statement is not true. For example, in a second-order univariate time-series setting with a standard normal marginal distribution, the positive definiteness of the input autocorrelation matrix requires the base lag-one and lag-two autocorrelations $\rho_Z(1)$ and $\rho_Z(2)$ to satisfy the inequality given by

$$\rho_Z(2) > \sqrt{3} \sin\left\{\frac{12}{\pi} \left[\arcsin\left(\frac{\rho_Z(1)}{2}\right) \right]^2 \right\} - \cos\left\{\frac{12}{\pi} \left[\arcsin\left(\frac{\rho_Z(1)}{2}\right) \right]^2 \right\}.$$

This inequality is denoted by the dashed line in Figure 1. However, not all possible pairs of $\rho_Z(1)$ and $\rho_Z(2)$ satisfy $\rho_Z(2) > 2\rho_Z^2(1)-1$, which insures the positive definiteness of the base autocorrelation matrix. This second inequality is denoted by the solid line in Figure 1. Thus, it is possible for $\rho_Z(1)$ and $\rho_Z(2)$ to fail to form a positive definite base autocorrelation matrix despite the positive definiteness of the input autocorrelation matrix and the likelihood of this event, which is illustrated in Figure 2.1 by the area between the solid curve and the dashed curve, is 2.62%. Although the use of the three-dimensional Gaussian distribution as the base process provides a great deal of flexibility for representing uncertainty in this second-order univariate time-series setting, it comes at the expense of not working for some input processes with feasible dependence structures. In their past work, Ghosh and Henderson [32] and Koruwicka and Cooke [50] investigate the likelihood of the failure of the NORTA transformation as a function of the number of components of the multivariate process. Our objective in this paper is to extend their investigation to a k-dimensional p^{th} -order time-series setting. In other words, we aim to determine the likelihood of the VARTA infeasibility as a function of the number of component time series and the order of dependence.



Figure 2.1: Illustration of the second-order ARTA infeasibility.

The rest of the paper is organized as follows. We provide the description of the ARTA and VARTA transformations and the reasons of their failure in Section 2.2. We introduce the extension of the onion method of Ghosh and Henderson [32] for sampling positive definite autocorrelation matrices representative of the stochastic properties of VARTA in Section 2.3. We present the results of our numerical study in Section 2.4 and in Section 2.5, we conclude with the summary of the paper and the discussion of an approximation to VARTA the simulation practitioner might use to avoid the failure of VARTA in high-dimensional settings.

2.2. ARTA/VARTA Transformations and Reasons of Their Failure

When the problem of interest is to construct a stationary univariate time series $\{X_t; t = 1, 2, ...\}$ with given marginal distribution F and first p autocorrelations $\rho_X(h)$, h = 1, 2, ..., p, the basic approach is to construct a p^{th} -order ARTA process [18]. The p^{th} -order ARTA process (ARTA(p)) defines a time series with uniform marginals on [0, 1] via the transformation $U_t = \Phi(Z_t)$, where the base process $\{Z_t; t = 1, 2, ...\}$ is a stationary, standard, Gaussian autoregressive process of order p with the autocorrelation structure given by $\rho_Z(h)$, h = 1, 2, ..., p. The input time series $\{X_t; t = 1, 2, ...\}$ is obtained via the transformation

 $X_t = F^{-1}[\Phi(Z_t)]$, which ensures that X_t has distribution F by well-known properties of the inverse cdf. Therefore, the central problem is to select the autocorrelation structure, $\rho_Z(h)$, h = 1, 2, ..., p, for the base process Z_t that gives the desired autocorrelation structure, $\rho_X(h)$, h = 1, 2, ..., p, for the input process X_t . It is easily shown that the base autocorrelation $\rho_Z(h)$ depends only on the input autocorrelation $\rho_X(h)$. The determination of the dependence structure for the base process is thus equivalent to solving p different correlation-matching problems.

Now we switch our focus to the representation of a stationary k-variate p^{th} -order timeseries input process $\{\mathbf{X}_t; t = 0, 1, 2, \ldots\}$, where $\mathbf{X}_t = (X_{1,t}, X_{2,t}, \ldots, X_{k,t})'$, with the following properties: (1) Each component time series $\{X_{i,t}; t = 0, 1, 2, ...\}$ has an arbitrary continuous marginal distribution denoted by F_i , i.e., $X_{i,t} \sim F_{X_i}$ for $i = 1, 2, \ldots, k$ and $t = 0, 1, 2, \dots$ (2) The dependence structure is specified via Pearson product-moment (or rank) correlations $\rho_{\mathbf{X}}(i, j, h) = \operatorname{Corr} [X_{i,t}, X_{j,t-h}]$ (or $\rho_{\mathbf{X}}(i, j, h) = \operatorname{Corr} [F_i(X_{i,t}), F_j(X_{j,t-h})]$) for i, j = 1, 2, ..., k and h = 0, 1, ..., p. Equivalently, the lag-h autocorrelation matrices are defined by $\Sigma_X(h) = [\rho_{\mathbf{X}}(i,j,h)]_{(k\times k)}$, for $h = 0, 1, \dots, p$, where $\rho_{\mathbf{X}}(i,i,0) = 1$. Biller and Nelson (2003) extends the theory behind the ARTA(p) process to represent multivariate time-series process $\{X_{i,t}; i = 1, 2, \dots, k, t = 1, 2, \dots\}$ with the k-dimensional p^{th} -order VARTA process (VARTA_k(p)). To do that, the authors choose the base process \mathbf{Z}_t as the stationary, standard, Gaussian vector autoregressive process of order p with representation $\mathbf{Z}_t = \sum_{h=1}^p \boldsymbol{\alpha}_h \mathbf{Z}_{t-h} + \mathbf{u}_t$ (Lütkepohl 1993). The $\boldsymbol{\alpha}_h, h = 1, 2, \dots, p$, are fixed $k \times k$ auto regressive coefficient matrices and $\mathbf{u}_t = (u_{1,t}, u_{2,t}, \dots, u_{k,t})'$ is a k-dimensional white noise vector representing the part of \mathbf{Z}_t that is not linearly dependent on past observations. The structure of \mathbf{u}_t is assumed to be such that $\mathrm{E}[\mathbf{u}_t] = \mathbf{0}_{(k \times 1)}$ and $\mathrm{E}[\mathbf{u}_t \mathbf{u}_{t+h}'] = \mathbf{\Sigma}_u$ if h = 0and $E[\mathbf{u}_t \mathbf{u}'_{t+h}] = \mathbf{0}_{(k \times k)}$ otherwise. Choosing Σ_u as $\Sigma_Z(0) - \sum_{h=1}^p \alpha_h \Sigma'_Z(h)$, where $\Sigma_Z(h)$ is the lag-h base autocorrelation matrix, ensures that each component series of the base process \mathbf{Z}_t , i.e., $\{Z_{i,t}; t = 1, 2, ...\}$ for i = 1, 2, ..., k, is marginally standard normal. Finally, the i^{th} component input time series $\{X_{i,t}; t = 1, 2, ...\}$ is obtained via the transformation $X_{i,t} = F_i^{-1}[\Phi(Z_{i,t})].$

As in the construction of the ARTA(p) process with k = 1, the challenge associated with the construction of the VARTA_k(p) process is to match the autocorrelation structure of the Gaussian vector autoregressive base process, $\rho_{\mathbf{Z}}(i, j, h)$, i, j = 1, 2, ..., k, h = 01, 2, ..., p, to the desired autocorrelation structure of the input process, $\rho_{\mathbf{X}}(i, j, h)$, i, j = 1, 2, ..., k, $h = 01, 2, \ldots, p$. This correlation-matching problem corresponds to the solution of

$$\rho_{\mathbf{X}}(i,j,h) = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{Y_i}^{-1}[\Phi(z_i)]F_{Y_j}^{-1}[\Phi(z_j)]\vartheta_{\rho_{\mathbf{Z}}(i,j,h)}(z_i,z_j)dz_idz_j - \mu_i\mu_j}{\sigma_i\sigma_j}$$

for $\rho_{\mathbf{Z}}(i, j, h)$ for a prespecified value of $\rho_{\mathbf{X}}(i, j, h)$, where $\vartheta_{\rho}(\cdot)$ is the standard bivariate normal probability density function with correlation ρ , $\mu_i = \mathbb{E}[Y_{i,t}]$ and $\sigma_i^2 = \operatorname{Var}[Y_{i,t}]$ for i = 1, 2. Thus, the problem of adjusting the correlation structure of the base process decomposes into $pk^2 + k(k-1)/2$ correlation-matching problems. Solving the correlationmatching problems might be a difficult task when the Pearson product-moment correlations are used. Fortunately, the corresponding function has the properties that allows the implementation of an efficient numerical search procedure to find $\rho_{\mathbf{Z}}(i, j, h)$ within a predetermined precision. Good references for numerical search procedures exploiting these properties are Cario and Nelson [20], Chen [22], and Biller and Nelson [8]. On the other hand, when rank-type correlations are used, it is possible to find $\rho_{\mathbf{Z}}(i, j, h)$ analytically via $\rho_{\mathbf{X}}(i, j, h) = 6/\pi \sin^{-1}(\rho_{\mathbf{Z}}(i, j, h)/2)$. To isolate the problem of estimating the probability of the VARTA infeasibility, we assume the use of rank-type autocorrelations in the remainder of the paper.

Next, we provide the procedure that generates multivariate time-series data of length n with component marginal distributions F_i , i = 1, 2, ..., k and input autocorrelation matrices $\Sigma_{\mathbf{X}}(h) = [\rho_{\mathbf{X}}(i, j, h); i, j = 1, 2, ..., k]$ prespecified for h = 0, 1, 2, ..., p: (1) Solve the correlation-matching problem for the base autocorrelation matrix $\Sigma_{\mathbf{Z}}(h)$ that would match the prespecified input autocorrelation matrix $\Sigma_{\mathbf{X}}(h)$ for h = 0, 1, 2, ..., p. (2) Obtain base process parameters $\alpha_1, \alpha_2, ..., \alpha_p$ and Σ_u from $\Sigma_{\mathbf{Z}}(h), h = 0, 1, 2, ..., p$ by using the multivariate Yule-Walker equations (Lütkepohl 1993). More specifically, first compute $\boldsymbol{\alpha} = \Sigma \Sigma_{\mathbf{Z}}^{-1}$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_p)$ and $\boldsymbol{\Sigma} = (\Sigma_Z(1), \Sigma_Z(2), ..., \Sigma_Z(p))$ are $(k \times kp)$ -dimensional matrices and

$$\boldsymbol{\Sigma}_{\mathbf{Z}} = \begin{pmatrix} \boldsymbol{\Sigma}_{Z}(0) & \boldsymbol{\Sigma}_{Z}(1) & \dots & \boldsymbol{\Sigma}_{Z}(p-2) & \boldsymbol{\Sigma}_{Z}(p-1) \\ \boldsymbol{\Sigma}_{Z}'(1) & \boldsymbol{\Sigma}_{Z}(0) & \dots & \boldsymbol{\Sigma}_{Z}(p-3) & \boldsymbol{\Sigma}_{Z}(p-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{\Sigma}_{Z}'(p-1) & \boldsymbol{\Sigma}_{Z}'(p-2) & \dots & \boldsymbol{\Sigma}_{Z}'(1) & \boldsymbol{\Sigma}_{Z}(0) \end{pmatrix}_{k\nu \times kp}$$
(2.1)

provides the full characterization of the base autocorrelation structure. Then, compute covariance matrix $\Sigma_{\mathbf{Y}}$ as equivalent to $\Sigma_{Z}(0) - \sum_{h=1}^{p} \alpha_{h} \Sigma'_{Z}(h)$. (3) Obtain starting values for $\mathbf{z}_{-p+1}, \mathbf{z}_{-p+2}, \ldots, \mathbf{z}_{0}$ using base autocorrelation structure $\Sigma_{Z}(h), h = 0, 1, \ldots, \nu$ and base process parameters $\alpha_{1}, \ldots, \alpha_{p}$ and Σ_{u} . Additionally, generate a series of Gaussian white noise vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_T$. So that, we can generate time series $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T$ recursively from $\mathbf{z}_t = \boldsymbol{\alpha}_1 \mathbf{z}_{t-1} + \cdots + \boldsymbol{\alpha}_p \mathbf{z}_{t-p} + \mathbf{u}_t$ for $t = 1, 2, \ldots, T$: (i) To generate $\mathbf{z}_{-p+1}, \mathbf{z}_{-p+2}, \ldots, \mathbf{z}_0$ as realizations of $\mathbf{Z}_{-p+1}, \mathbf{Z}_{-p+2}, \ldots, \mathbf{Z}_0$, whose joint distribution is given by a nonsingular kpdimensional multivariate normal distribution, choose a $(kp \times kp)$ matrix \mathbf{Q} such that $\mathbf{Q}\mathbf{Q}' = \mathbf{\Sigma}_{\mathbf{Z}}$, and then obtain the starting-value vector via $(\mathbf{z}'_0, \mathbf{z}'_{-1}, \ldots, \mathbf{z}'_{-p+1})' = \mathbf{Q} (v_1, \ldots, v_{kp})'$, where the v_i 's are independent standard normal random variates. (ii) To obtain an independent Gaussian white noise vector, first choose k independent univariate standard normal variates v_1, v_2, \ldots, v_k , and then multiply them with a $(k \times k)$ matrix \mathbf{P} for which $\mathbf{P}\mathbf{P}' = \mathbf{\Sigma}_u$ holds. Repeat this procedure for a total of T times to generate $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_T$. (4) Finally, transform the generated base time-series $\{z_{i,t}; i = 1, 2, \ldots, k, t = 1, 2, \ldots, T\}$ with standard normal marginal distributions into the desired process with cdfs $F_i, i = 1, 2, \ldots, k$ by implementing the inverse transformation method: $X_{i,t} = F_i^{-1} [\Phi(Z_{i,t})]$ for $i = 1, 2, \ldots, k$ and $t = 1, 2, \ldots, T$.

The success in the implementation of this data-generation procedure is dependent on the positive definiteness of both the base autocorrelation matrix $\Sigma_{\mathbf{Z}}$ and the variance-covariance matrix $\Sigma_{\mathbf{u}}$ as well as the stability of the underlying vector autoregressive base process, i.e., the roots of the reverse characteristic polynomial, $|\mathbf{I}_{(k \times k)} - \boldsymbol{\alpha}_1 z - \boldsymbol{\alpha}_2 z^2 - \cdots - \boldsymbol{\alpha}_p z^p| = 0$, lie outside of the unit circle in the complex plane ($\mathbf{I}_{(k \times k)}$ is the $(k \times k)$ identity matrix). It is easily shown that a stationary vector autoregressive process has a positive definite autocorrelation matrix and thus, a non-positive-definite base autocorrelation matrix always results in an unstable base process. However, the failure of the base process to be unstable, despite the positive definiteness of the base autocorrelation matrix, might appear as a possible reason of failure in a time-series setting unlike in a random-vector setting. It is shown in Biller and Nelson [8] that if the underlying base process is a stationary time series, then the input time series is also stationary and the stationarity of the base process is insured by the construction of a stable base vector autoregressive process. This can be explained by the fact that any stationary process with a stable state-space representation can be represented in the form of a stationary vector autoregressive model. Thus, stability implies stationarity [59], i.e., a non-stationary process is unstable, but an unstable process is not necessarily nonstationary. A close look at the existing literature reveals that the difference between stability and stationarity has been somewhat ambiguous. In the case of the second-order ARTA process, the region of stability is a triangle bounded by $\alpha_2 - \alpha_1 = 1$, $\alpha_2 + \alpha_1 = 1$, and $-2 \leq \alpha_1 \leq 2$. As long as the resulting autoregressive coefficients α_1 and α_2 fall on the boundary or outside

of this region, the underlying base process is unstable and the base autocorrelation matrix is not positive definite. Thus, for the second-order ARTA process, stability and positive definiteness coincide and the positive definiteness of the base autocorrelation matrix appears to be sufficient for insuring the stationarity of the resulting time series. We investigate whether this result continues to hold beyond second-order univariate time series in Section 4.

2.3. Sampling Positive Definite Autocorrelation Matrices

Ghosh and Henderson [32] have introduced a method, which is called the onion method, to sample exactly and quickly from the uniform distribution on the set $\Omega_k = \{\Sigma_{\mathbf{X}}(0) :$ $\Sigma_{\mathbf{X}}(0) = \Sigma'_{\mathbf{X}}(0), \Sigma_{\mathbf{X}}(0) \succ 0, diag(\Sigma_{\mathbf{X}}(0)) = 1$ of $k \times k$ positive definite correlation matrices of the k-dimensional NORTA transformation when viewed as a subset of $\Re^{k(k-1)/2}$. More specifically, the procedure of Ghosh and Henderson [32] samples $\Sigma_{\mathbf{X}}(0)$ uniformly from the convex, closed, compact, and full-dimensional set Ω_k in a way that the density $f(\Sigma_{\mathbf{X}}(0)) \propto 1$ for any $\Sigma_{\mathbf{X}}(0) \in \Omega_k$, where f is a function of the k(k-1)/2 upper-diagonal elements of $\Sigma_{\mathbf{X}}(0)$. This method of uniform sampling of positive definite correlation matrices is iterative in that it starts with a one-dimensional matrix and then grows out the matrix to the dimension desired by successively adding an extra row and the corresponding mirrored column chosen from an appropriate distribution. Ghosh and Henderson [32] report that Marsaglia and Olkin [61] use a similar matrix-growing approach in their algorithm to sample correlation matrices with a given set of eigenvalues, but they apply it to transform diagonal elements of arbitrary positive definite matrices to 1 in order to form correlation matrices from them. Other noteworthy references include Ouellette [68] discussing the uses of the layering approach and Guttman [34] proposing a numerical method for computing inverses of large nonsingular matrices. When compared to these methods, the exact sampling method of Ghosh and Henderson [32] scales very well with dimension as the uniform sampling of a positive definite correlation matrix reduces to the problem of sampling from a univariate beta distribution and a jointnormal independent random vector. Ghosh and Henderson [32] also report that for a given sample size the results are more accurate in the sense that confidence-interval widths are smaller for this sampling method. Thus, we choose to extend the onion method of Ghosh and Henderson [32] to our multivariate time-series setting for sampling positive definite

autocorrelation matrices representative of the stochastic properties of VARTA. Additionally, we discuss how to insure a prespecified rate of decay in the temporal-dependence structure. We first focus on a univariate time-series setting in Section 3.1 and then on a multivariate time-series setting in Section 3.2.

2.3.1. Univariate Time-Series Setting

The focus of this section is on the sampling of a $(p+1) \times (p+1)$ input autocorrelation matrix Σ_X of the form

$$\boldsymbol{\Sigma}_{X} = \begin{bmatrix} 1 & \rho_{X}(1) & \rho_{X}(2) & \dots & \rho_{X}(p-1) & \rho_{X}(p) \\ \rho_{X}(1) & 1 & \rho_{X}(1) & \dots & \rho_{X}(p-2) & \rho_{X}(p-1) \\ \rho_{X}(2) & \rho_{X}(1) & 1 & \dots & \rho_{X}(p-3) & \rho_{X}(p-2) \\ \rho_{X}(3) & \rho_{X}(2) & \rho_{X}(1) & \dots & \rho_{X}(p-4) & \rho_{X}(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{X}(p-1) & \rho_{X}(p-2) & \rho_{X}(p-3) & \dots & 1 & \rho_{X}(1) \\ \rho_{X}(p) & \rho_{X}(p-1) & \rho_{X}(p-2) & \dots & \rho_{X}(1) & 1 \end{bmatrix}_{(p+1)\times(p+1)}$$

for the ARTA(p) process. The objective is to generate Σ_X in a way that $f(\Sigma_X) \propto 1$, $\forall \Sigma_X \in \Omega_{p+1} == \{\Sigma_X : \Sigma_X = \Sigma'_X, \Sigma_X \succ 0, diag(\Sigma_X) = 1\}$. To simplify the presentation of the sampling algorithm, we define notation Σ_{κ} for the upper-left $\kappa \times \kappa$ submatrix of the autocorrelation matrix Σ_X and $\Sigma_{\kappa-1}$ for the completed matrix at the $(\kappa - 1)^{\text{th}}$ step, i.e.,

$$\mathbf{\Sigma}_{\kappa} = \left[egin{array}{cc} \mathbf{\Sigma}_{\kappa-1} & \mathbf{q}_{\kappa} \ \mathbf{q}'_{\kappa} & 1 \end{array}
ight]_{\kappa imes \kappa}$$

where $\mathbf{q}_{\kappa} = (\rho_X(\kappa-1), \rho_X(\kappa-2), \dots, \rho_X(1))'$ is a $(\kappa-1) \times 1$ column vector. We additionally use notation f_{κ} for the marginal density of Σ_{κ} at the κ^{th} step of the matrix completion and write $f_{\kappa}(\Sigma_{\kappa})$ as $\propto [\det(\Sigma_{\kappa})]^{\frac{p+1-\kappa}{2}}, \forall \Sigma_{\kappa} \in \Omega_{\kappa}, 2 \leq \kappa \leq p+1$ (Ghosh and Henderson 2003). Now we are ready to present the closed-form expression for the conditional probability density function of the vector \mathbf{q}_{κ} given $\Sigma_{\kappa-1}$:

$$\varphi\left(\mathbf{q}_{\kappa}\right) \propto \left[1 - \mathbf{q}_{\kappa}^{\prime} \boldsymbol{\Sigma}_{\kappa-1}^{-1} \mathbf{q}_{\kappa}\right]^{\frac{p+1-\kappa}{2}},$$
$$\forall \mathbf{q}_{\kappa} = \left[\rho_{X}(\kappa-1), \mathbf{q}_{\kappa-1}\right]^{\prime} \in \Psi_{\kappa-1} = \left\{\mathbf{q} \in \Re^{\kappa-1} \mid \mathbf{q}^{t} \boldsymbol{\Sigma}_{\kappa-1}^{-1} \mathbf{q} \leq 1\right\}$$

Notice that the density φ represents the joint density of $\rho_X(\kappa - 1)$ and $\mathbf{q}_{\kappa-1} = (\rho_X(\kappa - 2), \rho_X(\kappa - 3), \dots, \rho_X(1))'$ that has already been sampled in the $(\kappa - 1)^{\text{th}}$ step of the matrix completion from the probability density function $\varphi(\mathbf{q}_{\kappa-1})$. More importantly, the problem of interest is to generate $\rho_X(\kappa - 1)$ at κ^{th} step of the autocorrelation matrix generation. If

 $\mathbf{q}_{\kappa-1}$ is fixed at \mathbf{q} and $\mathbf{\Sigma}_{\kappa-1}$ is fixed at \mathbf{M} , then we have the following expression for the conditional density function of $\rho_X(\kappa-1)$ given $\mathbf{q}_{\kappa-1}$:

$$\varphi\left(\rho_X(\kappa-1)\right) = \varphi\left(\rho_X(\kappa-1)|\mathbf{\Sigma}_{\kappa-1} = \mathbf{M}, \mathbf{q}_{\kappa-1} = \mathbf{q}\right)$$
$$\propto \left[1 - \left[\rho_X(\kappa-1) \quad \mathbf{q}'\right] \mathbf{M}^{-1} \left[\rho_X(\kappa-1) \quad \mathbf{q}'\right]\right]^{\frac{p+1-\kappa}{2}}$$
(2.2)

To simplify the derivation of the conditional distribution of $\rho_X(\kappa - 1)$, we rewrite matrix Σ_{k-1} and hence, **M** as

$$\mathbf{\Sigma}_{\kappa-1} = \left[egin{array}{cc} \mathbf{\Sigma}_{\kappa-2} & \mathbf{q}_{\kappa-1} \ \mathbf{q}_{\kappa-1}' & 1 \end{array}
ight] = \left[egin{array}{cc} 1 & \mathbf{p}_{\kappa-1} \ \mathbf{p}_{\kappa-1}' & \mathbf{\Sigma}_{\kappa-2} \end{array}
ight],$$

where $\mathbf{p}_{\kappa-1} = (\rho_X(1), \rho_X(2), \dots, \rho_X(\kappa-2))$ is a $(\kappa-2)$ -dimensional row vector. Thus, fixing $\mathbf{p}_{\kappa-1}$ at \mathbf{p} and $\boldsymbol{\Sigma}_{\kappa-2}$ at \mathbf{B} leads to

$$\mathbf{M} = \begin{bmatrix} 1 & \mathbf{p} \\ \mathbf{p}' & \mathbf{B} \end{bmatrix} \quad \text{with} \quad \mathbf{M}^{-1} = \begin{bmatrix} 1 & \mathbf{p} \\ \mathbf{p}' & \mathbf{B} \end{bmatrix}^{-1} = \begin{bmatrix} 1 + \mathbf{p}\mathbf{V}\mathbf{p}' & -\mathbf{p}\mathbf{V} \\ \mathbf{V}\mathbf{p}' & \mathbf{V} \end{bmatrix},$$

where $\mathbf{V} = (\mathbf{B} - \mathbf{p'p})^{-1}$. Inserting the expression for \mathbf{M}^{-1} into (2.2) results in the following expression for $\varphi(\rho_X(\kappa - 1))$:

$$\begin{split} \varphi\left(\rho_X(\kappa-1)\right) &\propto \left[1 - \left[\begin{array}{cc} \rho_X(\kappa-1) & \mathbf{q}' \end{array}\right] \left[\begin{array}{cc} 1 + \mathbf{p} \mathbf{V} \mathbf{p}' & -\mathbf{p} \mathbf{V} \\ \mathbf{V} \mathbf{p}' & \mathbf{V} \end{array}\right] \left[\begin{array}{cc} \rho_X(\kappa-1) \\ \mathbf{q} \end{array}\right] \right]^{\frac{p+1-\kappa}{2}} \\ &= \left[\rho_X(\kappa-1) \left(\mathbf{q}' \mathbf{V} \mathbf{p}' + \mathbf{p} \mathbf{V} \mathbf{q}\right) - \rho_X^2(\kappa-1) \left(1 + \mathbf{p} \mathbf{V} \mathbf{p}'\right) + 1 - \mathbf{q}' \mathbf{V} \mathbf{q} \right]^{\frac{p+1-\kappa}{2}} \\ &= \left[Y + Z \rho_X(\kappa-1) - W \rho_X^2(\kappa-1)\right]^{\frac{p+1-\kappa}{2}} \\ &\propto \left[1 - \left(\frac{2\rho_X(\kappa-1)W - Z}{\sqrt{4YW + Z^2}}\right)^2\right]^{\frac{p+1-\kappa}{2}}, \end{split}$$

where $W = 1 + \mathbf{p}\mathbf{V}\mathbf{p}'$, $Y = 1 - \mathbf{q}'\mathbf{V}\mathbf{q}$, and $Z = \mathbf{p}\mathbf{V}\mathbf{q} + \mathbf{q}'\mathbf{V}\mathbf{p}'$. Thus, we generate $\rho_X(\kappa - 1)$ at the κ^{th} step using $\varphi(\rho_X(\kappa - 1))d\rho_X(\kappa - 1) \propto [1 - x^2]^{(p+1-\kappa)/2}dx$, where $x^2 = (2\rho_X(\kappa - 1)W - Z)^2/(4YW + Z^2)$. As a result of changing the variables via $x^2 = y$, the density function simplifies to $\varphi(\rho_X(\kappa - 1))d\rho_X(\kappa - 1) \propto [1 - y]^{\alpha_2 - 1}y^{\alpha_1 - 1}dy$, where $\alpha_1 = 1/2$ and $\alpha_2 = (p + 3 - \kappa)/2$. Thus, the generation of a value for $\rho_X(\kappa - 1)$ at the κ^{th} step reduces to the sampling of y from a univariate beta distribution with parameters 1/2 and $(p + 3 - \kappa)/2$ and solving equality $y = (2\rho_X(\kappa - 1)W - Z)^2/(4YW + Z^2)$ for $\rho_X(\kappa - 1)$. This procedure can be further modified to account for the decay in the temporal-dependence structure at the κ^{th} step via $|\rho_X(\kappa-1)| \leq A_{\kappa} |\rho_X(\kappa-2)|$ for $2 \leq \kappa \leq p+1$, where $\{A_k; k=3,4,\ldots,p+1\}$ is a series of constants insuring different rates of decay in temporal dependencies. In this particular case, the sampling algorithm reduces to sampling from a bounded beta distribution with parameters $\alpha_1 = 1/2$ and $\alpha_2 = (p+3-\kappa)/2$ and bounds given by $(2A_{\kappa}W|\rho_X(\kappa-2)|-Z)^2/(4YW+Z^2)$ and $(2A_{\kappa}W|\rho_X(\kappa-2)|+Z)^2/(4YW+Z^2)$.

2.3.2. Multivariate Time-Series Setting

In this section we restrict our attention to the sampling of a $k(p+1) \times k(p+1)$ input autocorrelation matrix $\Sigma_{\mathbf{X}}$ of the form

$$\boldsymbol{\Sigma}_{\mathbf{X}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} (0) & \boldsymbol{\Sigma}_{\mathbf{X}} (1) & \dots & \boldsymbol{\Sigma}_{\mathbf{X}} (p) \\ \boldsymbol{\Sigma}'_{\mathbf{X}} (1) & \boldsymbol{\Sigma}_{\mathbf{X}} (0) & \dots & \boldsymbol{\Sigma}_{\mathbf{X}} (p-1) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}'_{\mathbf{X}} (p) & \boldsymbol{\Sigma}'_{\mathbf{X}} (p-1) & \dots & \boldsymbol{\Sigma}_{\mathbf{X}} (0) \end{bmatrix}_{k(p+1) \times k(p+1)}$$

where $\Sigma_{\mathbf{X}}(h) = [\rho_{\mathbf{X}}(i, j, h)]_{k \times k}$ is the $k \times k$ correlation matrix at lag-h for $h = 0, 1, \ldots, p$. We sample $\Sigma_{\mathbf{X}}$ in three consecutive steps: We first generate the $\kappa \times \kappa$ principal minor $\Sigma_{\kappa} = [\rho_{\mathbf{X}}(i, j, 0)]_{\kappa \times \kappa}$ of $\Sigma_{\mathbf{X}}(0)$ in stages $\kappa = 1, 2, \ldots, k$, then generate the k-dimensional column vector $\mathbf{q}_{k} = (\rho_{\mathbf{X}}(1, 1, 1), \rho_{\mathbf{X}}(2, 1, 1), \ldots, \rho_{\mathbf{X}}(k, 1, 1))'$ in stage k + 1, and finally generate the k-dimensional column vector $\mathbf{q}_{k} = (\rho_{\mathbf{X}}(1, j, h), \rho_{\mathbf{X}}(2, j, h), \ldots, \rho_{\mathbf{X}}(k, j, h))'$ in stage κ for $\kappa = k + 3, k + 4, \ldots, k(p + 1)$.

In the first k stages we simply implement the sampling algorithm of Ghosh and Henderson (2003) to generate the $\kappa \times \kappa$ principal minors, $\kappa = 1, 2, ..., k$, of $\Sigma_{\mathbf{X}}(0)$, which is simply a $k \times k$ correlation matrix. Since the conditional density function φ for the $(\kappa - 1)$ -dimensional column vector \mathbf{q}_{κ} of the $\kappa \times \kappa$ principal minor is given by

$$\varphi\left(\mathbf{q}_{\kappa}\right) \propto \left(1-\mathbf{q}_{\kappa}' \boldsymbol{\Sigma}_{\kappa-1}^{-1} \mathbf{q}_{\kappa}\right)^{\frac{k(p+1)-\kappa}{2}}, \forall \mathbf{q}_{\kappa} \in \Psi_{\kappa}$$

where

$$\mathbf{\Sigma}_{\kappa} = \left[egin{array}{cc} \mathbf{\Sigma}_{\kappa-1} & \mathbf{q}_{\kappa} \ \mathbf{q}'_{\kappa} & 1 \end{array}
ight]_{\kappa imes \kappa}$$

we sample \mathbf{q}_{κ} at the κ^{th} step of the algorithm by (i) first sampling the random variate y from a beta distribution with $\alpha_1 = (\kappa - 1)/2$ and $\alpha_2 = (k(p+1) - \kappa + 2)/2$, (ii) setting $r = \sqrt{y}$, (iii) sampling a $(\kappa - 1)$ -dimensional random vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_{\kappa-1})'$ whose components are standard normally distributed, and (iv) finally setting $\mathbf{q}_{\kappa} = (\rho_{\mathbf{X}}(1, \kappa, 0), \rho_{\mathbf{X}}(2, \kappa, 0), \dots, \rho_{\mathbf{X}}(\kappa - 1, \kappa, 0))' = \boldsymbol{\Sigma}_{\kappa-1}^{1/2} \boldsymbol{\omega}$, where $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_{\kappa-1})' = (r\theta_1, r\theta_2, \dots, r\theta_{k-1})'$. In the $(k+1)^{\text{th}}$ -stage of the algorithm, we first sample the random variate y from a beta distribution with $\alpha_1 = k/2$ and $\alpha_2 = (k(p+1) - k + 1)/2$, set $r = \sqrt{y}$, and then impose conditions $|\rho_X(i, 1, 1)| \leq A_{i1} |\rho_X(i, 1, 0)|$, i = 1, 2, ..., k to insure decreasing temporal dependencies. Thus, the joint distribution for \mathbf{q}_{k+1} is given by

$$\varphi\left(\mathbf{q}_{k+1}\right) \propto \left(1 - \mathbf{q}_{k+1}' \boldsymbol{\Sigma}_{k}^{-1}(0) \mathbf{q}_{k+1}\right)^{\frac{kp-1}{2}}, \forall \mathbf{q}_{k+1} \in \Psi_{k+1}.$$

We generate \mathbf{q}_{k+1} using the polar transformation [47] as a result of which we set \mathbf{q}_{k+1} = $\Sigma_{\mathbf{X}}^{1/2}(0) \boldsymbol{\omega}$, where $\boldsymbol{\omega} = (r\theta_1, r\theta_2, \dots, r\theta_k)'$ whose components have means of 0 and standard deviations of r and $\boldsymbol{\Sigma}_{\mathbf{X}}^{1/2}(0)$ is the Cholesky factorization of the lag-0 correlation matrix $\Sigma_{\mathbf{X}}(0)$. Thus, the generation of \mathbf{q}_{k+1} is equivalent to the generation of a zero-mean random vector whose components are identically normally distributed with standard deviations of rand whose correlation matrix is $\Sigma_{\mathbf{X}}(0)$ and therefore, every component of \mathbf{q}_{k+1} can be generated using the marginal-conditional characterization of the multivariate normal distribution. The use of Corollary 3.3.1 and Theorem 3.3.4 of [83] for this purpose further results in a series of truncated normal distributions. More specifically, the joint distribution of random variables $\rho_{\mathbf{X}}(i, 1, 1), i = 1, 2, \dots, \ell$ is captured by an ℓ -dimensional normal distribution with zero-mean vector and a variance-covariance matrix $r^{\ell} \Sigma_{\mathbf{X}}^{1/2}(0)[\ell,\ell]$, where $\Sigma_{\mathbf{X}}^{1/2}(0)[\ell,\ell]$ stands for the upper-left $\ell \times \ell$ submatrix of $\Sigma_{\mathbf{X}}^{1/2}(0)$. Thus, the marginal distribution of $\rho_{\mathbf{X}}(\ell, 1, 1)$ given $\rho_{\mathbf{X}}(\ell-1,1,1), \rho_{\mathbf{X}}(\ell-2,1,1), \ldots, \rho_{\mathbf{X}}(1,1,1)$ is a normal distribution with mean μ_{ℓ} and variance σ_{ℓ}^2 for $\ell = 2, 3, \ldots, k$ and μ_{ℓ} and σ_{ℓ}^2 are determined as follows: We partition the vector $(\rho_{\mathbf{X}}(\ell, 1, 1), \rho_{\mathbf{X}}(\ell - 1, 1, 1), \rho_{\mathbf{X}}(\ell - 2, 1, 1), \dots, \rho_{\mathbf{X}}(1, 1, 1))'$ into two subvectors given by $s_1 = \rho_{\mathbf{X}}(\ell, 1, 1)$ and $\mathbf{s}_2 = (\rho_{\mathbf{X}}(\ell - 1, 1, 1), \rho_{\mathbf{X}}(\ell - 2, 1, 1), \dots, \rho_{\mathbf{X}}(1, 1, 1))'$, organize the entries of $r^{\ell} \Sigma_{\mathbf{X}}^{1/2}(0)[\ell, \ell]$ by reversing the indices in a decreasing order in matrix $\bar{\Sigma}$, and finally partition $\bar{\Sigma}$ into submatrices $\bar{\Sigma}_{1,1}$, $\bar{\Sigma}_{1,2} = \bar{\Sigma}'_{2,1}$, and $\bar{\Sigma}_{2,2}$, where $\bar{\Sigma}_{1,1}$ is the upper-left 1×1 submatrix, $\Sigma_{2,2}$ is the lower-right $(\ell - 1) \times (\ell - 1)$ submatrix, and $\Sigma_{1,2}$ is the row submatrix composed of the first row of Σ associated with the last $\ell - 1$ columns. In this particular case $\mu_{\ell} = \bar{\Sigma}_{1,2}\bar{\Sigma}_{2,2}^{-1}\mathbf{s}_2$ and $\sigma_{\ell}^2 = \bar{\Sigma}_{1,1} - \bar{\Sigma}_{1,2}\bar{\Sigma}_{2,2}^{-1}\bar{\Sigma}_{2,1}$. Now we can sample a value for $\rho_{\mathbf{X}}(\ell, 1, 1)$ from a truncated normal distribution with mean μ_{ℓ} , variance σ_{ℓ}^2 , and truncation characterized by $|\rho_{\mathbf{X}}(\ell, 1, 1)| \leq A_{\ell,1} |\rho_{\mathbf{X}}(\ell, 1, 0)|.$

Now we switch our focus to the sampling of column vector \mathbf{q}_{κ} in stage κ for $\kappa = k + 2, k + 3, \ldots, k(p+1)$. We know that the joint distribution of vectors \mathbf{q}_k and $\mathbf{q}_{\kappa-k-1}$, where

 $\mathbf{q}_{\kappa} = (\mathbf{q}_k, \mathbf{q}_{\kappa-k-1})'$ corresponds to the upper column vector in stage κ , is given by

$$\varphi\left(\mathbf{q}_{k},\mathbf{q}_{\kappa-k-1}
ight)\propto\left[1-\left[\begin{array}{cc}\mathbf{q}_{k}^{\prime}&\mathbf{q}_{\kappa-k-1}^{\prime}\end{array}
ight]\mathbf{\Sigma}_{\kappa-1}^{-1}\left[\begin{array}{cc}\mathbf{q}_{k}\\\mathbf{q}_{\kappa-k-1}\end{array}
ight]
ight]^{rac{k(p+1)-\kappa}{2}}$$

As a result of fixing $\Sigma_{\kappa-1}^{-1}$ at **V**, which is further decomposed into $k \times k$ submatrix $\mathbf{V}_{1,1}$, $k \times (\kappa - k - 1)$ submatrix $\mathbf{V}_{1,2}$ (= $\mathbf{V}'_{2,1}$), and $(\kappa - k - 1) \times (\kappa - k - 1)$ submatrix $\mathbf{V}_{2,2}$, and fixing $\mathbf{q}_{\kappa-k-1}$ at **q** whose components have already been generated in the previous steps, we obtain the marginal distribution of \mathbf{q}_k conditional on $\mathbf{q}_{\kappa-k-1}$ as follows:

$$\begin{split} \varphi\left(\mathbf{q}_{k}\right) &\propto \left[1 - \left[\begin{array}{cc} \mathbf{q}_{k}^{\prime} & \mathbf{q}^{\prime} \end{array}\right] \left[\begin{array}{cc} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{array}\right] \left[\begin{array}{cc} \mathbf{q}_{k} \\ \mathbf{q} \end{array}\right] \right]^{\frac{k(p+1)-\kappa}{2}} \\ &\propto \left[1 - \mathbf{q}_{k}^{\prime} \mathbf{V}_{11} \mathbf{q}_{k} - \mathbf{q}^{\prime} \mathbf{V}_{21} \mathbf{q}_{k} - \mathbf{q}^{\prime} \mathbf{V}_{12} \mathbf{q}_{k} - \mathbf{q}^{\prime} \mathbf{V}_{22} \mathbf{q} \right]^{\frac{k(p+1)-\kappa}{2}} \\ &\propto \left[1 - \mathbf{q}_{k}^{\prime} \mathbf{V}_{11} \mathbf{q}_{k} - \left(\mathbf{q}^{\prime} \mathbf{V}_{21} + \mathbf{q}^{\prime} \mathbf{V}_{12}^{\prime}\right) \mathbf{q}_{k} - \mathbf{q}^{\prime} \mathbf{V}_{22} \mathbf{q} \right]^{\frac{k(p+1)-\kappa}{2}} \\ &\propto \left[\left(1 - \mathbf{q}_{k}^{\prime} \mathbf{V}_{22} \mathbf{q}_{k}\right) + \left(-\mathbf{q}^{\prime} \mathbf{V}_{21} - \mathbf{q}^{\prime} \mathbf{V}_{12}^{\prime}\right) \mathbf{q}_{k} - \mathbf{q}^{\prime} \mathbf{V}_{11} \mathbf{q} \right]^{\frac{k(p+1)-\kappa}{2}} \end{split}$$

Next we define scalar $B = 1 - \mathbf{q}'_k \mathbf{V}_{22} \mathbf{q}_k$, $(1 \times k)$ -dimensional row vector $\mathbf{C} = -\mathbf{q}' \left(\mathbf{V}_{2,1} + \mathbf{V}'_{1,2} \right)$, and $k \times k$ square matrix $\mathbf{D} = \mathbf{V}_{1,1}$, and summarize the sampling problem as follows:

$$\varphi(\mathbf{q}_{k}) \propto \left[B + \mathbf{C}\mathbf{q}_{k} - \mathbf{q}_{k}'\mathbf{D}\mathbf{q}_{k}\right]^{\frac{k(p+1)-\kappa}{2}}, \ k+2 \le \kappa \le k \ (p+1)$$
$$= \left[B\left(1 - \left[\mathbf{q}_{k}'\frac{\mathbf{D}}{B}\mathbf{q}_{k} - \frac{\mathbf{C}}{B}\mathbf{q}_{k}\right]\right)\right]^{\frac{k(p+1)-\kappa}{2}}$$
$$\propto \left[1 - \mathbf{q}_{k}'\widetilde{\mathbf{D}}\mathbf{q}_{k}\right]^{\frac{k(p+1)-\kappa}{2}}$$

where $\widetilde{\mathbf{D}}_{ii} = \frac{1}{B} \mathbf{D}_{i,i} - \frac{1}{B} \mathbf{C}_i$ for i = 1, 2, ..., k. Finally, we use the approach described for the $(k+1)^{\text{th}}$ -stage to insure prespecified rates of decay in temporal dependencies in stages k+2, k+3, ..., k(p+1). The difference appears due to the replacement of $\Sigma_{\mathbf{X}}^{1/2}(0)$ with $\widetilde{\mathbf{D}}^{-1/2}$, where $\widetilde{\mathbf{D}}_{ii} = \frac{1}{B} D_{ii} - \frac{1}{B} C_i$ for i = 1, 2, ..., k.

2.4. Analysis

In this section we generate positive definite autocorrelation matrices using the sampling algorithm of Section 2.3, choose the replication number as 15000, and estimate the likelihood of the following types of failure: (1) Given that $\Sigma_{\mathbf{X}}$ is positive definite, the base autocorrelation matrix $\Sigma_{\mathbf{Z}}$ is not positive definite. (2) Both $\Sigma_{\mathbf{X}}$ and $\Sigma_{\mathbf{Z}}$ are positive definite and the base process is stable, but the covariance matrix of the white noise $\Sigma_{\mathbf{u}}$ is not positive definite. (3) Given that $\Sigma_{\mathbf{X}}$, $\Sigma_{\mathbf{Z}}$, and $\Sigma_{\mathbf{u}}$ are positive definite, the base autoregressive process is unstable. (4) Both $\Sigma_{\mathbf{X}}$ and $\Sigma_{\mathbf{Z}}$ are positive definite, but neither the covariance matrix of the white noise vector, $\Sigma_{\mathbf{u}}$, is positive definite nor the base vector autoregressive process is stable. Thus, the total probability of the VARTA_k(p) infeasibility in representing arbitrary multivariate time series with positive definite autocorrelation matrices is the sum of the probabilities of these types of failure.

In Section 2.4.1 we restrict our attention to a univariate time-series setting and in Section 2.4.2 to a multivariate time-series setting. We estimate the probability of the VARTA infeasibility together with the corresponding 95% confidence interval as a function of the number of component time series, the order of dependence, and the temporal-dependence decay rate. We perform our experiments using MATLAB with Intel Pentium 1.6 GHz processor CPU speed.

2.4.1. Univariate Time-Series Setting

We compute the failure probability of the ARTA(p) process in representing univariate time series whose autocorrelation structure has the form presented in Section 2.3.1 without any constraints insuring any ordering relationships among its autocorrelations $\rho_{\mathbf{X}}(h)$, h = 1, 2, ..., p. We tabulate our findings in Table 2.1 for p = 1, 2, ..., 20. We observe that the only reason

Order of Dependence	Failure Probability	Order of Dependence	Failure Probability
1	(0.00%, 0.00%)	11	(42.56%, 44.50%)
2	(2.31%, 2.93%)	12	(48.85%, 50.81%)
3	(3.45%, 4.21%)	13	(53.74%, 55.70%)
4	(6.44%, 7.44%)	14	(61.53%, 63.43%)
5	(11.26%, 12.52%)	15	(64.93%, 66.79%)
6	(14.49%, 15.89%)	16	(70.30%, 72.08%)
7	(19.35%, 20.93%)	17	(74.89%, 76.57%)
8	(24.20%, 25.90%)	18	(78.63%, 80.21%)
9	(30.06%, 31.88%)	19	(81.93%, 83.41%)
10	(36.39%, 38.29%)	20	(85.60%, 86.94%)

Table 2.1: Probability of the ARTA(p) infeasibility for p = 1, 2, ..., 20.

for the ARTA infeasibility is the failure of Σ_Z to be positive definite for the given positive definite Σ_X . In other words, the positive definiteness of the base autocorrelation matrix is sufficient to insure the existence of a stable base autoregressive process. In the next section we find that this observation does not extend to the multivariate time-series settings. In Table 2.1 we additionally report the determinant of Σ_X ($|\Sigma_X|$), the minimum eigenvalue of Σ_X (eig(Σ_X)), and the sum of the distances between the eigenvalues of Σ_X and the value of zero (eig(all)). As consistent with the previous findings reported for the random-vector settings, we observe that the ARTA infeasibility occurs as the boundary of the positive definite input autocorrelation matrices is approached in the univariate time-series setting. This observation extends readily to the multivariate time-series setting and thus, to the VARTA infeasibility in the next section. Therefore, we restrict our focus to the computation of the VARTA infeasibility and its decomposition into different reasons of failure in Section 2.4.2.

р	$ \mathbf{\Sigma}_X $	$ ext{eig}(\mathbf{\Sigma}_X)$	eig(all)
2	(0.02236, 0.02653)	(0.01087, 0.01277)	(2.24, 2.28)
3	(0.01415, 0.01660)	(0.01079, 0.01227)	(2.74, 2.79)
4	(0.00821, 0.00959)	(0.01083, 0.01191)	(3.29, 3.35)
5	(0.00531, 0.00615)	(0.01184, 0.01276)	(3.65, 3.71)
6	(0.00295, 0.00343)	(0.01149, 0.01225)	(4.14, 4.20)
7	(0.00225, 0.00262)	(0.01212, 0.01281)	(4.41, 4.47)
8	(0.00119, 0.00139)	(0.01152, 0.01213)	(4.91, 4.98)
9	(0.00092, 0.00108)	(0.01174, 0.01229)	(5.16, 5.23)
10	(0.00051, 0.00060)	(0.01091, 0.01139)	(5.69, 5.76)
11	(0.00038, 0.00045)	(0.01095, 0.01142)	(5.85, 5.92)
12	(0.00021, 0.00026)	(0.00978, 0.01019)	(6.36, 6.44)
13	(0.00015, 0.00018)	(0.00964, 0.01004)	(6.55, 6.62)
14	(9.03E - 05, 10.9E - 05)	(0.00871, 0.00906)	(7.08, 7.16)
15	(7.08E - 05, 8.94E - 05)	(0.00836, 0.00870)	(7.27, 7.35)
16	(4.01E - 05, 5.30E - 05)	(0.00743, 0.00774)	(7.83, 7.93)
17	(3.08E - 05, 3.99E - 05)	(0.00689, 0.00718)	(8.09, 8.19)
18	(1.89E - 05, 2.50E - 05)	(0.00607, 0.00633)	(8.54, 8.65)
19	(1.19E - 05, 1.71E - 05)	(0.00559, 0.00585)	(8.86, 8.97)
20	(7.14E - 06, 1.06E - 05)	(0.00489, 0.00512)	(9.29, 9.41)

Table 2.2: Behavior of ARTA(p), p = 1, 2, ..., 20 when failure occurs.

Next we impose constraints insuring different rates of decay in temporal dependencies via the selection of A_{κ} , $\kappa = 1, 2, ..., p + 1$ as 1, 1/2, 1/4, and 1/8. The smaller the value of A_{κ} , $\kappa = 1, 2, ..., p+1$, the faster the rate of decay in temporal dependencies. We present our findings for p = 1, 2, ..., 10 in Table 2.4.1. A close look at Table 2.3 shows the significant impact of the decay rate on the ARTA infeasibility. The failure probability does not exceed 3% in any of our experimental settings reported for p = 1, 2, ..., 10. Thus, we conclude that

Order of Dependence	Te	mporal-Depend	dence Decay Ra	ate
р	1	1/2	1/4	1/8
1	(0.0%, 0.0%)	(0.0%, 0.0%)	(0.0%, 0.0%)	(0.0%, 0.0%)
2	(0.5%, 0.8%)	(0.3%, 0.5%)	(0.4%, 0.6%)	(0.2%, 0.4%)
3	(1.3%, 1.7%)	(1.1%, 1.5%)	(0.7%, 1.0%)	(0.6%, 0.8%)
4	(1.7%, 2.2%)	(1.2%, 1.5%)	(1.0%, 1.3%)	(0.9%, 1.3%)
5	(2.0%, 2.4%)	(1.2%, 1.6%)	(1.2%, 1.6%)	(1.1%, 1.5%)
6	(2.3%, 2.8%)	(1.5%, 1.9%)	(1.4%, 1.8%)	(1.3%, 1.6%)
7	(2.4%, 2.9%)	(1.6%, 2.0%)	(1.7%, 2.2%)	(1.5%, 1.9%)
8	(2.4%, 2.9%)	(1.5%, 1.9%)	(1.5%, 1.9%)	(1.5%, 1.9%)
9	(2.2%, 2.7%)	(1.8%, 2.3%)	(1.5%, 2.0%)	(1.6%, 2.0%)
10	(2.0%, 2.5%)	(1.6%, 2.1%)	(1.6%, 2.0%)	(1.4%, 1.8%)

Table 2.3: Probability of the ARTA(p) infeasibility with decreasing temporal dependencies.

the ARTA transformation is a highly flexible model with the ability of working for most time-series processes in univariate time-series settings.

2.4.2. Multivariate Time-Series Setting

As in the previous section, we first investigate the likelihood of the VARTA infeasibility without assuming any relationships among correlations $\rho_{\mathbf{X}}(i, j, h)$, i, j = 1, 2, ..., k, h = 0, 1, ..., p. We report our findings on the VARTA infeasibility in Table 2.4 for a first-order process whose number of components changes between 2 and 5.

Reasons of	Number of Components (k)						
Failure	2	3	4	5			
Reason 1	(20.2%, 21.5%)	(31.2%, 32.7%)	(42.3%, 43.8%)	(49.8%, 51.4%)			
Reason 2	(0.9%, 1.2%)	(0.6%, 0.9%)	(0.5%, 0.7%)	(0.0%, 0.0%)			
Reason 3	(6.4%, 7.2%)	(26.6%, 27.9%)	(25.0%, 26.4%)	(41.3%, 42.8%)			
Reason 4	(1.4%, 1.8%)	(1.9%, 2.4%)	(4.2%, 4.8%)	(6.9%, 7.7%)			
Total Failure							
Probability	(29.6%, 31.0%)	(61.4%, 62.9%)	(73.2%, 74.6%)	(99.9%, 100.0%)			

Table 2.4: Probability of the VARTA_k(1) infeasibility as a function of k.

Our focus on the first-order VARTA process is motivated by the fact that any VARTA_k(p) process can be represented as a VARTA_{kp}(1) process [8]. Recall that "Reason 1" in Table 2.4 denotes the failure of $\Sigma_{\mathbf{Z}}$ to be positive definite, "Reason 2" represents the failure of $\Sigma_{\mathbf{u}}$ to be positive definite, "Reason 3" corresponds to the failure of the roots of the reverse characteristic polynomial to satisfy the stability condition, and finally "Reason 4" corresponds

to the lack of a stable vector autoregressive process whose reverse characteristic polynomial satisfies the stability condition and the variance-covariance matrix of the white noise vector is positive definite. We find that the total failure probability reaches the value of one very quickly, i.e., we estimate the mean failure probability as 99.98% for the five-dimensional first-order VARTA₅(1) process and unlike in the univariate time-series setting, the failure of $\Sigma_{\mathbf{Z}}$ to be positive definite is not the only reason for the VARTA infeasibility. However, like in the univariate time-series setting, we find that decay rate in temporal dependencies has a significant (positive) impact as tabulated in Table 2.5. For $|\rho_{\mathbf{X}}(i, j, h)| \ge |\rho_{\mathbf{X}}(i, j, h + 1)|$, $h \ge 1$, the mean failure probability for the VARTA₅(1) process never reaches the value of one; instead, it converges to the probability of 41%. We observe that the failure probabil-

Table 2.5: Probability of the $\text{VARTA}_k(1)$ infeasibility with decreasing temporal dependencies.

# of Components	Temporal-Dependence Decay Rate					
k	1	1/2	1/4	1/8		
2	(3.0%, 3.5%)	(2.2%, 2.7%)	(1.6%, 2.0%)	(1.3%, 1.7%)		
3	(26.2%, 27.6%)	(18.8%, 20.1%)	(10.4%, 11.4%)	(7.5%, 8.4%)		
4	(34.0%, 35.5%)	(22.9%, 24.3%)	(11.5%, 12.6%)	(7.3%, 8.1%)		
5	(37.9%, 44.0%)	(23.6%, 24.9%)	(10.3%, 11.3%)	(5.3%, 6.0%)		

ity decreases even more significantly with increasing values of the decay rate in temporal dependencies.

Next we restrict our attention to a bivariate time-series setting and estimate the likelihood of the VARTA infeasibility as a function of the order of dependence p without any constraints associated with decaying temporal dependencies (see Table 2.6). We observe

	$\underline{\qquad}$									
p	Failure		Reasons of Failure							
	Probability	Reason 1	Reason 2	Reason 3	Reason 4					
1	(29.5%, 31.0%)	(20.2%, 21.5%)	(0.8%, 1.2%)	(6.3%, 7.1%)	(1.4%, 1.8%)					
2	(88.1%, 89.1%)	(8.2%, 9.1%)	(3.1%, 3.7%)	(64.2%, 65.7%)	(10.9%, 12.0%)					
3	(97.2%, 97.7%)	(13.8%, 14.9%)	(0.6%, 0.9%)	(54.1%, 55.7%)	(26.7%, 28.1%)					
4	(99.7%, 99.8%)	(17.0%, 18.3%)	(0.0%, 0.1%)	(54.9%, 56.5%)	(25.6%, 27.0%)					
5	(100.0%, 100.0%)	(18.5%, 19.8%)	(0.0%, 0.0%)	(50.6%, 52.2%)	(28.66%, 30.12%)					

Table 2.6: Probability of the VARTA₂(p) infeasibility as a function of p

that the failure probability increases very rapidly with the order of dependence. Wei [89] notes that the value of 3 for p appears to be a good approximation for the maximum order

of dependence in most practical applications. We find that the mean failure probability of the VARTA₂(3) infeasibility is 97.53% with the major reason of failure as the lack of a stable vector autoregressive process despite the positive definiteness of the base autocorrelation matrix $\Sigma_{\mathbf{Z}}$. Thus, unlike in the univariate time-series setting, the positive definiteness of $\Sigma_{\mathbf{Z}}$ is not sufficient for the existence of a stable vector autoregressive process, i.e., failure might occur even when the resulting base autocorrelation matrix is positive definite. On the other hand, the failure probabilities decrease significantly as we assume decaying temporal dependencies over time (see Table 2.7). Notice that the mean probability of the VARTA₂(p)

Order of Dependence	Temporal-Dependence Decay Rate					
р	1	1/2	1/4	1/8		
1	(3.00%, 3.57%)	(2.21%, 2.71%)	(1.61%, 2.03%)	(1.38%, 1.78%)		
2	(3.60%, 4.22%)	(2.54%, 3.07%)	(1.97%, 2.43%)	(1.95%, 2.42%)		
3	(4.05%, 4.70%)	(3.75%, 4.38%)	(2.62%, 3.15%)	(2.24%, 2.74%)		
4	(5.45%, 6.19%)	(4.29%, 4.96%)	(2.77%, 3.32%)	(2.22%, 2.71%)		
5	(5.04%, 5.77%)	(4.39%, 5.07%)	(2.87%, 3.43%)	(2.36%, 2.87%)		
6	(5.64%, 6.40%)	(4.35%, 5.03%)	(2.98%, 3.54%)	(2.33%, 2.83%)		
7	(6.03%, 6.82%)	(3.99%, 4.65%)	(2.84%, 3.40%)	(2.44%, 2.96%)		
8	(5.92%, 6.70%)	(3.99%, 4.64%)	(2.72%, 3.27%)	(2.34%, 2.85%)		
9	(5.77%, 6.54%)	(4.38%, 5.06%)	(2.71%, 3.26%)	(2.38%, 2.90%)		
10	(5.57%, 6.33%)	(4.10%, 4.76%)	(2.85%, 3.41%)	(2.33%, 2.84%)		

Table 2.7: Probability of the VARTA₂(p) infeasibility with decreasing temporal dependencies.

infeasibility is estimated as 6.43% for $|\rho_{\mathbf{X}}(i, j, h)| \ge |\rho_{\mathbf{X}}(i, j, h + 1)|$, $h \ge 1$. This particular value of the failure probability is around five times higher than the one computed for the ARTA(p) process in Section 2.4.1. However, the failure probability is still very low, indicating a strong support for the applicability of the VARTA₂(p) process in a time-series setting with a high order of dependence.

We conclude this section with the presentation of the mean probability of the VARTA_k(p) infeasibility for p = 1, 2, ..., 5 and k = 1, 2, ..., 5 when $|\rho_{\mathbf{X}}(i, j, h)| \ge |\rho_{\mathbf{X}}(i, j, h+1)|, h \ge 1$. Despite the fact that the VARTA process can get infeasible quickly with increasing dimensions, its performance is very much dependent on the rate of decay in temporal dependencies. As indicated by the results reported in Table 2.8, VARTA appears to be a highly flexible model with the ability of representing the stochastic properties of the p^{th} -order multivariate time-series processes with $p \le 3$ and strictly decreasing temporal dependencies.

Order,	Number of Components, k										
р			1 2 3		5						
1	(0.0%, 0.0%)	(3.0%, 3.5%)	(26.2%, 27.6%)	(34.0%, 35.5%)	(40.2%, 41.7%)						
2	(0.5%, 0.7%)	(3.6%, 4.2%)	(28.9%, 30.3%)	(41.7%, 43.3%)	(52.9%, 54.5%)						
3	(0.8%, 1.2%)	(4.0%, 4.7%)	(32.7%, 34.2%)	(47.5%, 49.1%)	(54.8%, 56.4%)						
4	(0.8%, 1.1%)	(4.3%, 5.0%)	(34.1%, 35.6%)	(48.8%, 50.4%)	(60.4%, 61.9%)						
5	(0.7%, 1.0%)	(5.6%, 6.3%)	(35.3%, 36.8%)	(52.4%, 54.0%)	(62.8%, 64.3%)						

Table 2.8: Mean failure probability of $VARTA_k(p)$ with decreasing temporal dependencies.

2.5. Conclusion

VARTA is a flexible multivariate input model used for generating stationary multivariate time series with prespecified marginal distributions and positive definite autocorrelation matrices. The question we ask in this paper is whether we can represent any stationary multivariate time series with arbitrary feasible dependence structures. We start our study with the extension of the exact sampling algorithm of Ghosh and Henderson [32] to our time-series setting for sampling positive definite autocorrelation matrices. Using the resulting sampling algorithm, we estimate the likelihood of the ARTA/VARTA infeasibility as a function of the order of dependence, the number of component time-series processes, and the rate of decay in temporal dependencies. Our findings suggest that ARTA is a highly flexible model with the ability of representing the stochastic properties of most univariate time-series processes. However, without assuming any pattern of decay in temporal dependencies, we find that the VARTA process fails very rapidly in representing arbitrary multivariate time series as a function of both the order of dependence and the number of components. We estimate the mean failure probability as 99.98% for the first-order five-dimensional VARTA₅(1) process and as 100.00% for the fifth-order two-dimensional VARTA₂(5) process. In this particular case, we recommend the simulation practitioner to consider the representation of the VARTA process via the copula-vine specification of Kurowicka and Cooke [50]. This representation requires the characterization of the multivariate temporal dependence structure as a mix of pairwise correlations and pairwise conditional correlations. We refer the reader to Biller [12] for details on this particular representation of VARTA. However, the probability of the VARTA infeasibility starts diminishing very quickly with increasing decay rates. When we assume that $|\rho_{\mathbf{X}}(i,j,h)| \ge |\rho_{\mathbf{X}}(i,j,h+1)|, i,j = 1, 2, ..., k, h = 1, 2, ..., p - 1$, the mean failure probabilities for the VARTA₅(1) and VARTA₂(5) processes decrease to 41.00% and

5.407%, respectively. Thus, we conclude that VARTA is a highly flexible, easily implementable method for driving large-scale discrete-event stochastic simulations whose inputs are represented by multivariate time series with temporal dependencies decaying over time.

Chapter 3

The Impact of Dependence on Single-Server Queueing Systems

In this study, we use advanced simulation input modeling to study the impact of bivariate and temporal dependencies among interarrival and service times on the performance of a single-server queue. The distinguishing feature of our study from those in the literature is to consider a wide variety of distributional shapes for the probability density functions of the interarrival and service times, and the patterns that arise in the temporal dependencies of the interarrival and service times. We generate dependent interarrival and service times via using the Vector-Auto-Regressive-to-Anything method, which has never before been used in queueing systems. We investigate the impact of dependent interarrival and service times on the average waiting time of M/M/1, M/G/1 and G/M/1 systems. We show that high variance and positive skewed nonexponential distributions decrease the performance of the single-server system. We also compare impact of temporal dependencies in interarrival and service times for M/M/k systems ($k \ge 2$) with the M/M/1 system, and conclude that the effect of dependence decreases in multi-server systems. Our main contribution is to combine this advanced input modeling method with queueing theory for investigating the impacts of dependent interarrival and service times on the average waiting time. ¹

3.1. Introduction

Most queueing models assume that job interarrival and service times, failure times, and repair requirements are independent and identically distributed, each being modeled as a renewal process (see, for example, Kelley [46]). These restricted assumptions lead to models

 $^{^1\}mathrm{Co}\text{-authors:}$ Bahar Biller and Alan Scheller-Wolf

that are very easy to simulate, and which are analytically tractable. Unfortunately, these models are often poor representations of real-life systems where correlations do, in fact, exist. It is well known that dependent time-series input processes occur naturally in many service, telecommunication, and manufacturing systems. For example, Melamed et. al. [62] observe autocorrelation in sequences of compressed video frame bitrates, while Ware et. al. [87] 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. Further examples of dependent systems are provided in Section 2.

The goal of this paper is to perform a simulation study to observe the effects of dependence on the average waiting time of the single-server queue. In the literature, autocorrelations are often considered in M/M/1 systems with lag-one autocorrelations in interarrival and service times. In this paper, we go beyond the use of such an input model for interarrival and service times; we use the Vector-Autoregressive-To-Anything (VARTA) input model capturing a wide variety of distributional shapes for the probability density functions of interarrival and service times, and lag-two autocorrelations. Specifically, VARTA input model is introduced by Biller and Nelson [8] to generate multivariate time series for discrete-event stochastic simulations. Additionally, we use the Johnson translation system [44] to represent the marginal distributions of interarrival times and service demands. This allows us to consider both unimodal and bimodal distributional shapes with any combination of first four moments (i.e., mean, variance, skewness, and kurtosis). Therefore, our main contribution is to combine VARTA as a multivariate input generation technique with queueing theory to generate insights into the effects of dependence on queueing performance.

A related study is performed by Livny et. al. [57] who examine the impact of autocorrelated, exponentially distributed interarrival and service times 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 ignoring the autocorrelation in the interarrival times and/or the autocorrelation in the service times can predict overly optimistic line lengths and waiting times. However, we still do not know how the bivariate and temporal dependencies of different strengths and patterns among interarrival and service times affect the performance of this infinite-capacity, single-stage, single-server system with arbitrary marginal distributions. We fill this gap by investigating the impact of dependent inputs on the queuing performance and understanding how the operating principles (i.e., factory physics) — that are very well understood under the assumption of independent inputs — change with bivariate and temporal dependencies. 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 within queuing systems, we use discrete-event stochastic simulation to perform this study. The first challenge is thus to find a plausible and yet parsimonious model for representing the dependencies in the stochastic simulation of the queueing system. We overcome this challenge by using the VARTA model representing and generating interarrival and service times with given marginal distributions from the Johnson translation system and positive definite autocorrelation matrices. We present statistical summaries of queue performance, exploring a variety of values for factors such as the pattern of dependence, server utilization, and wide variety of distributional shapes for interarrival and service times. We find that the average waiting time is monotonically increasing in the positively autocorrelated interarrival times and/or service times. For example, average waiting time of the M/M/1 system at 50% utilization increases from 0.99935 (independent and identically distributed interarrival times) to 2.17093 (+0.7 autocorrelated interarrival times). This increase can be explained by the burstiness caused by positively autocorrelated interarrival times. However, the average waiting time is not monotonically decreasing in the negatively autocorrelated interarrival times and/or service times. For example, the average waiting time of the M/M/1 system at 50% utilization increases from 0.85313 (-0.5 autocorrelated service times, and independent and identically distributed interarrival times) to 0.86746 (-0.7 autocorrelated service times, and independent and identically distributed interarrival times). This result is consistent with the literature. However, nonmonotonic behavior of the average waiting time in negatively autocorrelated interarrival times is a new observation to the literature. This observation happens in high utilization levels and negative autocorrelations close to minus one. For instance, the average waiting time of the M/M/1 system at 80% utilization increases from 3.03475 (-0.5 autocorrelated interarrival times, and independent and identically distributedservice times) to $3.03867 (-0.7 \text{ autocorrelated interarrival times, and independent and iden$ tically distributed interarrival times). This increase is even higher for -0.99 autocorrelated interarrival times, in which the average waiting time increases to 10.19664 for this specific M/M/1 system with 80% utilization level. We investigate this nonmonotonic behavior by

the analysis of the sample paths of the interarrival and service times. Since we experiment lag-two autocorrelations, we investigate the impact of the pattern of dependence on the average waiting time of single-server queue. Moreover, our experimental study shows that the impact of dependence in interarrival times and/or service demands increase monotonically as the utilization of the queueing system. Additionally, we investigate the impact of bivariate and temporal dependent interarrival times and/or service times on the performance of M/G/1 and G/M/1 systems with Johnson marginal distributions.

The rest of the paper is organized as follows. The motivation behind our study is discussed in Section 3.2. Section 3.3 gives a comprehensive survey of the related literature, while Section 3.4 introduces the VARTA model for capturing the bivariate and temporal dependencies among interarrival times and service demands with marginal distributions from the Johnson translation system. Section 3.5 provides a summary of simulation results and key findings. Finally, the conclusion of our study is presented in Section 3.6, and the related appendices are followed by the section of conclusions.

3.2. Motivation

Interarrival and service times of service and telecommunication systems are often assumed to be independent and identically distributed. However, this assumption leads to a poor representation of the real system when there are autocorrelations within and across interarrival and service times. One example of the problem of dependence in queueing systems is given in call centers: Strongly autocorrelated interarrival times cause burstiness in the call centers [16]. Accounting dependence is also important in data and voice transfers, such as in ISDN (Integrated Service Digital Network) and ATM (Asynchronous Transfer Mode) technologies [29]. Also within the WWW (World Wide Web) environment, the arrival of internet users to web sites and their "think time" exhibit strong autocorrelation and burstiness [25].

Strong dependencies are also observed among job arrivals, machining times, machine failure and down times, vendor lead times, and material handling times of manufacturing systems. However, these dependencies are typically ignored in both production and planning problems. Altiok [3] states that down and failure times in manufacturing environments are positively correlated. He reports that observing 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 common machine failures that occur less frequently and have longer down times [23]. In other manufacturing applications, the dependence is significant in the pressure variable of a continuous-flow production line [9], and the correlation among different parts' processing times in parallel production line using ATO (Assemble-to-Order) manufacturing[91], and it also affects the performance of sequential and ordering systems in JIT (Just-in-Time) manufacturing systems [82].

In service environments, dependence has also been shown to be important in modeling customer demand for airline tickets, where demand is price sensitive and changes over time [31]; 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 exhibits strong temporal dependence.

3.3. Literature Review

In the queueing literature, dependence is typically ignored, despite the fact that there exists well-known temporal dependencies within as well as between interarrival times and service demands of computer, telecommunication, manufacturing, and service systems (see Section 2 for industry applications). 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, simulation studies use various input generation models such as TES (Transfer-Expand-Sample) method [62], Minification [54], etc. for generating dependent interarrival and service times.

The analytical studies are often motivated by telecommunication systems and call centers, where autocorrelated arrivals in data and voice transfer have significant impact on performance [29]. In those studies, the dependence is either short-range or long-range. Only lag-one autocorrelation in interarrival and service times, or correlation between interarrival and service times is studied in short range. The most common result in those analytical papers is the demonstration of dramatic performance reduction due to positive autocorrelation in interarrival times. In addition, positive temporal dependence in interarrival times or positive bivariate dependence between interarrival times and service demands increase the average waiting time [35].

Markov arrival and service processes are commonly used to model dependence in analytical studies; see Runnenburg [74], Hadidi [35], Langaris [52], Heffes et al. [38], Fendick et al. [29], Szekli et al. [81], Patuwo et al. [69], Boucherie et al. [14], and Shioda [76]. The common result in these studies is that the performance of the single server queue decreases as the bivariate dependence approaches plus one. Additionally, Chao [21] uses a bivariate exponential distribution to represent dependent interarrival times and service demands. He shows that average waiting time is monotonically decreasing in the bivariate dependency. Iver and Manjunath [41] investigate the impact of only bivariate dependence between interarrival times and service demands by using various heavy tailed distributions. They use finite mixture of bivariate distributions to model the joint density of interarrival times and service demands; thus, this model allows them to capture non-linear dependencies while specifying marginal distributions. They present a numerical study to capture the effect of numerous parameters in the model on the waiting time. In a recent study, Xu [91] makes a structural analysis of a queueing system with multiple classes of autocorrelated 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 induces the autocorrelation structure. The paper uses a simple queueing model of Poisson arrivals and exponential service times with parallel servers and considers only autocorrelated interarrival times. Xu [91] concludes that more positively autocorrelated arrivals improve the worst component performance, which has the longest queue among parallel servers, by reducing the diversity among the servers. Moreover, the impact of dependencies is considered under heavy-traffic conditions. Jacobs [43] investigates the effect of bivariate and temporal dependence within interarrival times and service demands on the average waiting time of a M/M/1 queue under heavy traffic conditions. The interarrival times and service demands are generated by a mixed exponential moving average method, called exponential mixed autoregressive moving average with both autoregression and moving average of order. After deriving the heavy traffic limiting distribution of the average waiting time, they conclude that positive bivariate dependence between service and interarrival times decreases the average waiting time and positive autocorrelation within interarrival times or service demands increases the average waiting time.

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. Two widely used methods are the TES (Transfer-Expand-Sample) method [62] and the Minification technique [54] to generate short-range dependent interarrival times and service demands for the queueing simulations. It is found that autocorrelation in interarrival times has 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 the autocorrelation in service times at a specific utilization level. The impact of the positively autocorrelated interarrival and service times is monotonically increasing in the utilization of the queue.

Livny et. al. [57] is the most comprehensive simulation study in the literature. They analyze the impact of lag-one autocorrelation generated by TES (Transfer-Expand-Sample) method [62] and Minification technique [54] on an M/M/1 system. They conclude that introducing autocorrelated arrivals has a greater impact than 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 the 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. In addition to bivariate and temporal dependencies within as well as between interarrival times and/or service demands, another simulation research stream focuses on time-dependent behavior of interarrival times and/or service times (i.e., the mean of interarrival times and/or service times varies across time). Thus, there is no bivariate or temporal dependencies in the queueing system. Nelson and Taaffe [67] study this time-dependent queueing system in single-server case and multi-server case Taaffe [66]. They develop a numerical method to evaluate the time-dependent mean, variance, and higher order moments of the number of jobs via finite sets of differential equations which are integrated numerically. For algorithmic purposes, Iravani et. al. [42] develop a decomposition algorithm for parallel queues in which interarrival times are autocorrelated. They extend their algorithm to large systems as an approximation of the performance measures, and perform numerical examples to test the accuracy of their decomposition algorithm.

Our work fits in the category of simulation studies of the problem. We use a comprehensive input modeling framework, VARTA, for the first time in the literature of dependent queues. This allows us to represent dependence both in time sequence and among interarrival times and service demands. We work on a single-server queue using a flexible system of distributions known as the Johnson translation system as opposed to assuming exponentially distributed inputs. VARTA allows us to introduce cross-correlations between interarrival and service times of different jobs. In addition to the novelty of our study, we perform simulations to observe the impact of pattern of dependence in the autocorrelations of the first two lags. Our main results verify the other simulation studies' findings by using VARTA input modeling and extend these observations to new cases. Moreover, we explain the nonmonotonic behavior of the average waiting time with negatively autocorrelated service times. We also observe that the impact of temporal dependencies in interarrival times and/or service demands is monotonically increasing in server utilization. In addition to our results, we show that positive correlation between interarrival times and service demands increases the performance of the queue and the impact of correlation is monotonically increasing in both the utilization of the system and the magnitude of the bivariate dependency.

3.4. VARTA for Modeling Interarrival and Service Times

In this section, we describe how we capture the joint distributional properties of interarrival times and service demands using VARTA. Specifically, VARTA is a comprehensive multivariate input model for representing and generating multivariate time-series input processes with marginal distributions from the Johnson translation system, and autocorrelation matrices represented in product-moment correlations. It achieves 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 introduced for representing a stationary k-variate time-series input process $\{\mathbf{X}_t; t = 0, 1, 2, ...\}$ has the following properties:

- (1) Each component time series $\{X_{i,t}; t = 0, 1, 2, ...\}$ 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, ... and i = 1, 2, ..., k.
- (2) The dependence structure is specified via Pearson product-moment correlations $\rho_{\mathbf{X}}(i, j, h) =$ Corr $[X_{i,t}, X_{j,t-h}]$, for h = 0, 1, ..., p and i, j = 1, 2, ..., k. Equivalently, the lag-h correlation matrices are defined by $\Sigma_X(h) =$ Corr $[\mathbf{X}_t, \mathbf{X}_{t-h}] = [\rho_{\mathbf{X}}(i, j, h)]_{(k \times k)}$, for h = 0, 1, ..., p, where $\rho_{\mathbf{X}}(i, i, 0) = 1$.

For instance, in a single-server queueing system with correlated interarrival and service times, the random vector X_t is two-dimensional (i.e., k = 2) with one component representing the interarrival time while the other corresponding to the service time. However, if there is additional lag-one autocorrelation between interarrival and service times, then the random vector of interest is four-dimensional; e.g., $X_{1,0}$ and $X_{1,1}$ represent interarrival times at times zero and one, respectively, while $X_{2,0}$ and $X_{2,1}$ represent service times for times zero and one. In this model, 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

$$\mathbf{Z}_{t} = \alpha_{1}\mathbf{Z}_{t-1} + \alpha_{2}\mathbf{Z}_{t-2} + \dots + \alpha_{p}\mathbf{Z}_{t-p} + \mathbf{u}_{t}, \ t = 0, \pm 1, \pm 2, \dots,$$
(3.1)

where $\mathbf{Z}_t = (Z_{1,t}, Z_{2,t}, \dots, Z_{k,t})'$ is a $(k \times 1)$ random vector of the observations recorded at time t and the α_i , $i = 1, 2, \dots, p$, are fixed $(k \times k)$ autoregressive coefficient matrices. Finally, $\mathbf{u}_t = (u_{1,t}, u_{2,t}, \dots, u_{k,t})'$ is a k-dimensional white noise vector representing the part of \mathbf{Z}_t that is not linearly dependent on past observations; it has a positive definite $(k \times k)$ covariance matrix Σ_u such that

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

We generate multivariate time series from this bivariate Gaussian vector autoregressive process of any required length, say T. In our study, we experiment with lag-one and lag-two for bivariate and temporal dependencies in interarrival and service times. The algorithm for generating dependent interarrival and service times are presented for p order of dependence and length T:

- First, we obtain the starting values, z_{-p+1}, z_{-p+2},..., z₀, using the autocorrelation structure, Σ_Z(h), h = 0, 1, ..., p, and the implied system parameters, α₁,..., α_p and Σ_u. We also obtain a series of Gaussian white noise vectors, u₁, u₂,..., u_T. Then we generate the time series z₁, z₂,..., z_T recursively as z_t = α₁z_{t-1}+···+ α_pz_{t-p}+u_t for t = 1, 2, ..., T.
- To generate $\mathbf{z}_{-p+1}, \mathbf{z}_{-p+2}, \ldots, \mathbf{z}_0$ as realizations of $\mathbf{Z}_{-p+1}, \mathbf{Z}_{-p+2}, \ldots, \mathbf{Z}_0$ whose joint distribution is given by a nonsingular 2p-dimensional multivariate normal distribution, we choose a $(2p \times 2p)$ matrix \mathbf{Q} such that $\mathbf{Q}\mathbf{Q}' = \boldsymbol{\Sigma}_{\mathbf{Z}}$. Then we obtain the

starting-value vector as $(\mathbf{z}_0, \mathbf{z}_{-1}, \cdots, \mathbf{z}_{-p+1})' = \mathbf{Q} (v_1, \cdots, v_{2p})'$, where the v_i 's are independent standard normal random variates. Therefore, this way ensures that the process starts stationary.

• To obtain the series of independent Gaussian white noise vectors, $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_T$, we first choose two-independent univariate standard normal variates v_1 and v_2 for interarrival and service times, and then multiply by a (2×2) matrix **P** for which $\mathbf{PP'} = \Sigma_u$; that is, $\mathbf{u}_t = \mathbf{P} (v_1, v_2)'$. We repeat this process **T** times.

We use Johnson family [44] distributions in our simulations. A cumulative distribution function of any Johnson-type random variable \boldsymbol{X} is specified through

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

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:

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

We refer the reader to Figure 3.1 for the partition of the two-dimensional plot of β_1 and β_2 into regions in which a different Johnson family is used to match the third and fourth moments.

3.5. Implementation

In this section, we discuss how we select the distributional properties of the interarrival and service times we experiment with as well as the performance metrics. We additionally provide the experimental setup and discuss the selection of the simulation run length, warmup period, and the number of replications that ensures a prespecified level of error. We end this section by presenting the simulation results. We use C++ programming language for our simulations. The VARTA multivariate input generation software is developed in C++ environment by Biller and Nelson [8]. In our experiments, the autocorrelated data for processes like interarrival and service times, are generated by this software. We refer



Figure 3.1: The two-dimensional region of the square of skewness β_1 and kurtosis β_2 any legitimate random variable can have and its partition among the Johnson families.

the reader to Biller and Nelson [8] for further technical details about the code. 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 for determining initial conditions of our simulations. We refer the reader to Law and Kelton [53] for the estimation of the distributional properties of the performance measures of interest and a detailed description of the replication/deletion method. We determine the number of replications needed for a **95%** confidence interval of the average waiting time by following a two-stage procedure. We present simulation results for M/M/1 with lag-one autocorrelation in interarrival times and/or service times (no bivariate dependence), M/M/1 with lag-two autocorrelation in interarrival and service times (effect of dependence pattern), G/M/1 with lag-one autocorrelation in interarrival times and/or service times, effect of bivariate dependence on M/M/1, G/M/1 and M/G/1, M/M/1 with both bivariate and temporal dependence (effect of cross-correlation), and the effect of temporal dependent interarrival and service times in M/M/k systems for $k \geq 2$.

3.5.1. First-Order Autocorrelated, Exponentially Distributed Interarrival and Service Times

We perform experiments assuming four different utilization levels; 25%, 50%, 80%, and 99%. We fix the service demand process as exponential with mean 1 and change mean of the exponentially distributed interarrival times to reach the prespecified utilization levels. For example, the mean of the arrival process is 4 at the utilization of 25%. In each experiment with respect to the utilization level, the lag-one autocorrelation values for interarrival and service times are selected from {-0.99, -0.70, -0.50, -0.30, 0.00, 0.30, 0.50, 0.70, 0.99}. In the tables presenting experiment results, $\rho_A(1)$ and $\rho_S(1)$ represent the lag-one autocorrelation in interarrival times and service times, respectively. The horizontal autocorrelation values represent $\rho_S(1)$ and vertical autocorrelations represent $\rho_A(1)$. Thus, there are 81 different experiments per utilization level. For example in Table 3.1, 0.391 represents the result obtained from the M/M/1 system that has interarrival times with lag-one autocorrelation of 0.30 and service times with lag-one autocorrelation of -0.50. We refer reader to Tables 3.19, 3.20, and 3.21 in the appendix section for experiment results of 50%, 80%, and 99% utilizations of the M/M/1 system.

					$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	0.273	0.204	0.203	0.205	0.210	0.227	0.251	0.309	3.668
-0.70	0.250	0.216	0.217	0.220	0.231	0.253	0.280	0.333	2.471
-0.50	0.268	0.231	0.232	0.237	0.251	0.278	0.307	0.364	2.477
-0.30	0.291	0.252	0.254	0.260	0.277	0.308	0.343	0.405	2.492
0.00	0.357	0.300	0.302	0.310	0.333	0.374	0.419	0.501	2.714
0.30	0.482	0.390	0.391	0.402	0.434	0.489	0.554	0.668	3.182
0.50	0.659	0.513	0.514	0.527	0.567	0.639	0.721	0.868	3.763
0.70	1.065	0.797	0.794	0.814	0.868	0.971	1.083	1.295	4.999
0.99	22.858	19.765	20.021	19.612	20.199	20.587	20.642	22.216	45.340

Table 3.1: First-Order Autocorrelated, Exponentially Distributed Interarrival and Service Times, 25% utilization

The average waiting time is monotonically increasing in the utilization level with autocorrelation in interarrival times and/or service times. Additionally, the average waiting time is monotonically decreasing as a function of the autocorrelation autocorrelation in interarrival times and/or service times. These two results confirm the literature. However, the impact of dependence is interesting in negatively autocorrelated interarrival times and/or service demands. In the literature, nonmonotonic behavior of the average waiting time in negatively autocorrelated service demands was observed, but nonmonotonic behavior of the performance in negatively autocorrelated interarrival times is a new observation in the literature. Because nonmonotonic behavior of the average waiting time in negatively autocorrelated interarrivals exist when lag-one autocorrelation is close to -1 and/or utilization is close to 100%. For instance, the average waiting time is 0.273 in 25% utilization with $\rho_A(1) = -0.99$ and $\rho_S(1) = -0.99$, which is greater than 0.250 with $\rho_A(1) = -0.70$ and $\rho_S(1) = -0.99$ at both 50% and 80% utilizations. However, we observe this nonmonotonic behavior of the average waiting time at $\rho_A(1) = -0.50$ and all $\rho_S(1)$ values in 99% utilization.

In the literature, nonmonotonic behavior of the average waiting time in negatively autocorrelated service demands was observed, but nonmonotonic behavior of the performance in negatively autocorrelated interarrival times is a new observation in the literature. Because nonmonotonic behavior of the average waiting time in negatively autocorrelated interarrivals exist when lag-one autocorrelation is close to -1 and/or utilization is close to 100%. We focus our analysis on the nonmonotonic behavior of the average waiting time in negatively autocorrelated service demands. When service times are negatively autocorrelated, there



Figure 3.2: Waiting times of a single sample path for lag-one=-0.9 and lag-one=-0.5 autocorrelated service times

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 negatively autocorrelated service times. This is indeed the case for negative service-time autocorrelations that are close to zero. However, as the negative autocorrelation in service times approaches -1, 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 autocorrelation closer to -1 in service times.

We observe that the increase in mean waiting time at -0.9 (lag-one) autocorrelated service times for 80% and 50%, lag-one=-0.99 (lag-one) autocorrelated service times for 25%. We pick 25% utilization level, identical and independently distributed interarrival times, and set lag-one autocorrelation for the service times from -0.9, -0.5, 0, 0.5, 0.9. We observe that the mean waiting time increases from 0.30089 to 0.31076 as the lag-one temporal dependence decreases from -0.5 to -0.9. In order to analyze this nonmonotonic behavior, we investigate the sample paths of the average waiting time and service times. For example, when we compare the waiting time time sample path of -0.90 autocorrelated service times and -0.50 case in Figure 3.2, we observe that there are higher peaks in -0.90 case, which



Figure 3.3: Histogram of the frequencies of waiting times of (lag-one=-0.9 - lag-one=-0.5) autocorrelated service time cases

might be the cause of the nonmonotonic behavior. Therefore, 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.

Figure 3.3 shows the histogram created by subtracting the number of cases of waiting times at specific range for the system with -0.5 autocorrelated service times from corresponding cases of waiting times for the system with -0.9 autocorrelated service times. The number of jobs with zero waiting time in -0.5 cases is significantly larger than -0.9 case. We conjecture that the large amount of zero-waiting time incidents outweigh other ones, so the performance of the queue is better in the system with -0.5 autocorrelation in service times than the system with -0.9 autocorrelation service times. Therefore, the nonmonotonic result, in which average waiting time of -0.5 case is lower than -0.9 case, might be explained by this far left-tail of the waiting time distribution in -0.5 case.

3.5.2. Second-Order Autocorrelated, Exponentially Distributed Interarrival Times or Service Times

In this experiment, we analyze the impact of dependence pattern on the performance of the single-server system for lag-h, h = 2. The interarrival times and service demands' lag-one and lag-two autocorrelations are defined. We set **0.3** as the lag-one and lag-two autocorrelations. For instance, (+, +) represents that both lag-one and lag-two autocorrelations are equal to 0.3. Our conclusions are robust to this value. Note that the average waiting times

u	chi and iden	uncany v	distributed		muos			
	Utilization	iid	(+, +)	(+, -)	(-,+)	(-, -)	(-)	(+)
	25%	0.333	0.500	0.364	0.291	0.260	0.277	0.433
	50%	1.005	1.694	1.020	0.909	0.709	0.811	1.358
_	80%	4.016	7.328	3.849	3.800	2.838	3.264	5.509

Table 3.2: Second-order autocorrelated, exponentially distributed interarrival times, and independent and identically distributed service times

Table 3.3: Second-order autocorrelated, exponentially distributed service times, and independent and identically distributed interarrival times

Utilization	iid	(+,+)	(+, -)	(-,+)	(-, -)	(-)	(+)
25%	0.333	0.391	0.349	0.318	0.302	0.309	0.372
50%	1.005	1.421	1.023	0.959	0.836	0.891	1.225
80%	4.016	6.881	3.891	3.882	3.074	3.425	5.294

in Tables 3.2 and 3.3 are calculated with 95% confidence interval. In the tables (+) means that lag-one autocorrelation is 0.3 and (-) means lag-one autocorrelation of -0.3. (+) and (-) are benchmark cases like independent system. In lag-two simulations, (+, +) means lag-one is 0.3 and lag-two is 0.3; (+, -) means lag-one is 0.3 and lag-two is -0.3. (-, +) and (-, -) are defined similarly.

In all utilization levels, the average waiting times of different dependence patterns of autocorrelated interarrival times or service demands satisfy the following order:

$$(+,+) > (+) > (+,-) > (-,+) > (-) > (-,-).$$

This result is intuitive because of the temporal dependence decay (see Table 3.4). For instance, the positive autocorrelation doesn't vanish in (+, +) through time, which gets close to zero at lag-ten; however it becomes zero at lag-seven for (+). Thus, (+, +) has more

	Table 5.4. Temporal Dependence Decay									
Lag	(+0.3, +0.3)	(+0.3, -0.3)	(-0.3, +0.3)	(-0.3, -0.3)	(-0.3)	(+0.3)				
1	0.300	0.300	-0.300	-0.300	-0.300	0.300				
2	0.300	-0.300	0.300	-0.300	0.090	0.090				
3	0.138	-0.257	-0.138	0.257	-0.027	0.027				
4	0.101	0.018	0.101	0.018	0.008	0.008				
5	0.055	0.118	-0.055	-0.118	-0.002	0.002				
6	0.036	0.043	0.036	0.043	0.001	0.001				
7	0.021	-0.032	-0.021	0.032	0.000	0.000				
8	0.013	-0.032	0.013	-0.032	0.000	0.000				
9	0.008	0.000	-0.008	0.000	0.000	0.000				
10	0.005	0.014	0.005	0.014	0.000	0.000				

Table 3.4: Temporal Dependence Decay


Figure 3.4: Lognormal distributions (a) $\delta = 1$ with mean 1.65, (b) $\delta = 2$ with mean 1.13

negative impact on the performance than (+). Similar argument works for (-, -) versus (-). As for (+, -) versus (-, +), there is slight difference in the decay of temporal dependence except lag-one and lag-two autocorrelations. The sign of the lag-one autocorrelation has significant impact on the average waiting time than lag-two autocorrelations; hence, (+, -) increases the average waiting time more. However, the performance of (+, -) and (-, +) are almost equal to each other in high utilizations because of the diminishing effect of the pattern of temporal dependence in lag-two autocorrelations.

3.5.3. First-Order Autocorrelated, Exponentially Distributed Service Times and Lognormal Interarrival Times

Considering nonexponential marginals, we perform G/M/1 queue simulations relaxing the assumption of exponential interarrival times at 50% and 80% utilization levels. In deviating from exponential distribution of interarrival times (M/M/1), we use lognormal distributions such that the Johnson translation parameters are $\boldsymbol{\xi} = \boldsymbol{0}$, $\boldsymbol{\lambda} = \boldsymbol{1}$, $\boldsymbol{\gamma} = \boldsymbol{0}$ and $\boldsymbol{\delta} = \boldsymbol{1}, \boldsymbol{2}$. Figure 3.4 shows these two lognormal distributions with different tails comparing to the exponential distribution. First plot represents the lognormal distribution with $\boldsymbol{\delta} = \boldsymbol{1}$ that has a mean of **1.65** and variance of **4.67**. This lognormal distribution is positive skewed and has higher variance than the exponential distribution with mean **1.65**. On the other hand, the second plot represents the lognormal distribution with $\boldsymbol{\delta} = \boldsymbol{2}$ that has a mean of **1.13** and variance of **0.36**, and it is negative skewed and has lower variance than the exponential distribution with mean **1.13**.

In order to investigate the impact of nonexponential interarrival times with temporal

1 mico, 0070	a a maa a a a a a a a a a a a a a a a a								
M/M/1	50% Utilization				$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	3.24	0.44	0.42	0.44	0.55	0.79	1.13	1.88	38.12
-0.70	2.19	0.46	0.45	0.47	0.56	0.73	0.95	1.47	31.60
-0.50	2.28	0.51	0.50	0.52	0.60	0.77	0.99	1.47	28.31
-0.30	2.37	0.57	0.56	0.58	0.67	0.84	1.08	1.57	29.40
0.00	2.69	0.72	0.70	0.73	0.82	1.01	1.26	1.78	34.21
0.30	2.93	0.99	0.98	1.01	1.12	1.34	1.61	2.23	34.99
0.50	3.59	1.37	1.36	1.39	1.52	1.77	2.08	2.76	38.90
0.70	5.14	2.25	2.25	2.28	2.43	2.74	3.13	3.90	47.08
0.99	72.36	62.57	63.35	64.65	64.95	66.59	68.64	71.08	135.65

Table 3.5: First-Order Autocorrelated, Exponentially Distributed Service and Interarrival Times, 50% utilization

Table 3.6: First-Order Autocorrelated, Exponentially Distributed Service Times and Lognormal Interarrival Times ($\delta = 1$), 50% utilization

G/M/1	50% Utilization				$\rho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	5.05	0.60	0.56	0.60	0.78	1.18	1.72	2.86	47.43
-0.70	2.56	0.54	0.52	0.56	0.67	0.88	1.16	1.77	32.79
-0.50	2.55	0.58	0.56	0.59	0.70	0.91	1.17	1.73	33.95
-0.30	2.58	0.64	0.62	0.66	0.76	0.97	1.24	1.80	30.75
0.00	2.94	0.78	0.77	0.80	0.92	1.14	1.43	2.04	35.82
0.30	3.29	1.07	1.05	1.08	1.22	1.48	1.81	2.49	36.95
0.50	3.93	1.45	1.43	1.48	1.64	1.93	2.30	3.08	40.30
0.70	5.64	2.33	2.32	2.37	2.57	2.93	3.35	4.28	49.83
0.99	73.79	63.86	63.68	64.87	65.00	67.27	73.72	77.57	148.49

dependence, we experiment with the M/M/1 system as a benchmark to the G/M/1 system. Table 3.5 shows the simulation results for the M/M/1 system at 50% utilization level. The exponential interarrival times have a mean of **1.65** and the exponential service times have a mean of **0.825**. Table 3.6 shows the results for the G/M/1 system with lag-one autocorrelation in interarrival and service times. The average waiting times of all temporal dependence cases of interarrival and service times at the G/M/1 system is larger than the M/M/1 system. This result is intuitive since the lognormal distribution with $\delta = 1$ has significantly larger variance than the exponential distribution, and it has a longer right tail than the exponential distribution.

Regarding the impact of dependence with utilization levels on the average waiting time, the effect increases in utilization. We refer the reader to Table 3.22 for the M/M/1 at 80% utilization and Table 3.23 for the G/M/1 with lognormal interarrival distributions ($\delta = 1$) at 80% utilization in the appendix. In addition the utilization effect, the nonmonotonic behavior of negatively autocorrelated interarrival and service times in the M/M/1 system occurs in this G/M/1 system as well. For example, the average waiting time of the G/M/1 system at 50% utilization (Table 3.6) with -0.50 lag-one autocorrelation in interarrival times and 0.70 lag-one autocorrelation in service times is 1.73. However, it increases to 1.77 when the negative autocorrelation in interarrival times becomes -0.70. Therefore, this nonmonotonic behavior doesn't change by deviating from the assumption of exponential distribution in interarrival times.

In addition to lognormal distribution with longer right tail, we experiment with a lognormal distribution with longer left tail than the exponential distribution. In this case, the Johnson parameter of δ is equal to two and is plotted in Figure 3.4. Similarly, we need the simulation results of the M/M/1 system for benchmark purposes. Table 3.7 shows the simulation results for the M/M/1 system at 50% utilization level. The exponential interarrival times have a mean of **1.13** and the exponential service times have a mean of **0.565**. Table 3.8 shows the results for the G/M/1 system with lag-one autocorrelation in interarrival and service times. The average waiting times of all temporal dependence cases of interarrival and service times at the G/M/1 system is smaller than the M/M/1 system. This result is intuitive since the lognormal distribution with $\delta = 2$ has significantly smaller variance than the exponential distribution, and it has a longer left tail than the exponential distribution.

111100, 0070	atimetron								
M/M/1	50% Utilization				$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	2.25	0.30	0.29	0.30	0.38	0.54	0.77	1.30	26.63
-0.70	1.48	0.32	0.31	0.33	0.38	0.50	0.66	0.99	21.93
-0.50	1.55	0.35	0.34	0.36	0.41	0.53	0.68	1.02	21.70
-0.30	1.58	0.39	0.38	0.40	0.46	0.58	0.73	1.07	22.68
0.00	1.74	0.49	0.48	0.50	0.57	0.69	0.87	1.23	23.05
0.30	2.14	0.68	0.67	0.69	0.77	0.92	1.10	1.52	24.44
0.50	2.49	0.94	0.93	0.95	1.04	1.21	1.43	1.87	24.64
0.70	3.48	1.56	1.54	1.56	1.67	1.88	2.13	2.67	27.16
0.99	48.52	42.98	44.29	44.33	44.54	45.24	46.17	47.89	92.72

Table 3.7: First-Order Autocorrelated, Exponentially Distributed Service and Interarrival Times, 50% utilization

Table 3.8: First-Order Autocorrelated, Exponentially Distributed Service Times and Lognormal Interarrival Times ($\delta = 2$), 50% utilization

	(//							
G/M/1	50% Utilization				$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	1.31	0.16	0.15	0.16	0.20	0.29	0.42	0.75	21.85
-0.70	1.16	0.16	0.16	0.17	0.20	0.29	0.41	0.70	20.92
-0.50	1.26	0.17	0.16	0.17	0.21	0.30	0.42	0.71	20.85
-0.30	1.27	0.18	0.17	0.18	0.22	0.31	0.44	0.73	21.66
0.00	1.24	0.20	0.19	0.20	0.24	0.34	0.47	0.78	22.09
0.30	1.38	0.22	0.21	0.23	0.28	0.38	0.53	0.85	23.16
0.50	1.39	0.25	0.25	0.26	0.32	0.44	0.59	0.94	23.66
0.70	1.71	0.32	0.31	0.32	0.39	0.53	0.71	1.11	23.77
0.99	6.43	2.75	2.73	2.75	2.94	3.31	3.76	4.88	39.98

3.5.4. First-Order Autocorrelated, Exponentially Distributed Interarrival Times and Lognormal Service Times

Considering nonexponential marginals for the service times, we perform M/G/1 queue simulations relaxing the assumption of exponential service times at 50% and 80% utilization levels. Similar to G/M/1 experiments, we use lognormal distributions such that the Johnson translation parameters are $\boldsymbol{\xi} = 0, \lambda = 1, \gamma = 0$ and $\delta = 1, 2$. Figure 3.4 shows these two lognormal distributions with different tails comparing to the exponential distribution. In order to investigate the impact of nonexponential service times with temporal dependence, we need to experiment with the M/M/1 system as a benchmark. Table 3.9 shows the simulation results for the M/M/1 system at 50% utilization level. The exponential interarrival times have a mean of **3.3** and the exponential service times have a mean of **1.65**. Table 3.10 shows the results for the M/G/1 system with lag-one autocorrelation in interarrival and service times. The average waiting times of all temporal dependence cases of interarrival and service times at the M/G/1 system is significantly larger than the M/M/1 system. This result is intuitive since the lognormal distribution with $\delta = 1$ has significantly larger variance than the exponential distribution, and it has a longer right tail than the exponential distribution. This result coincides with our previous result in the G/M/1 system. Therefore, larger variance in either the interarrival times or service times increases the average waiting time. Similar to the M/M/1 and the G/M/1 systems, the impact of dependence increases in higher utilization levels. We refer the reader to Table 3.26 for the M/M/1 at 80% utilization and Table 3.27 for the M/G/1 with lognormal interarrival distributions ($\delta = 1$) at 80% utilization in the appendix. In addition the utilization effect, the nonmonotonic behavior of negatively autocorrelated interarrival and service times in the M/M/1 and G/M/1 systems occurs in this M/G/1 system as well.

In addition to lognormal distribution with longer right tail, we experiment with a lognormal distribution with longer left tail than the exponential distribution. In this case, the Johnson parameter of δ is equal to two and is plotted in Figure 3.4. Similarly, we need the simulation results of the M/M/1 system for benchmark purposes. Table 3.11 shows the simulation results for the M/M/1 system at 50% utilization level. The exponential interarrival times have a mean of 2.26 and the exponential service times have a mean of 1.13. Table 3.12 shows the results for the M/G/1 system with lag-one autocorrelation in interarrival and service times. The average waiting times of all temporal dependence cases of interarrival

0.99
0.99
0.00
74.7
58.4
65.2
63.6
68.6
69.4
69.7
71.9
263.7

Table 3.9: First-Order Autocorrelated, Exponentially Distributed Interarrival and Service times, 50% utilization

Table 3.10: First-Order Autocorrelated, Exponentially Distributed Interarrival Times and Lognormal Service Times ($\delta = 1$), 50% utilization

	(//							
M/G/1	50% Utilization				$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	32.8	1.9	1.7	1.6	1.9	2.6	3.78	6.5	136.6
-0.70	24.5	1.8	1.5	1.5	1.7	2.3	3.10	5.1	145.5
-0.50	25.4	1.9	1.6	1.6	1.8	2.3	3.11	5.0	133.6
-0.30	28.3	1.9	1.7	1.7	1.9	2.5	3.23	5.2	139.9
0.00	31.1	2.3	2.0	2.0	2.2	2.8	3.64	5.6	149.4
0.30	29.3	2.9	2.6	2.6	2.9	3.5	4.38	6.4	158.8
0.50	30.1	3.7	3.4	3.4	3.7	4.4	5.29	7.5	165.1
0.70	31.4	5.6	5.3	5.3	5.6	6.3	7.42	9.7	183.6
0.99	180.7	131.3	132.5	135.1	135.2	130.5	138.4	142.6	337.7

1 mes, 3070	utilization								
M/M/1	50% Utilization				$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	4.49	0.61	0.57	0.61	0.75	1.08	1.55	2.59	59.09
-0.70	3.08	0.63	0.62	0.65	0.77	1.00	1.31	2.01	43.58
-0.50	3.12	0.69	0.68	0.71	0.83	1.06	1.36	2.04	42.69
-0.30	3.42	0.78	0.77	0.80	0.92	1.14	1.46	2.13	43.74
0.00	3.35	0.98	0.96	1.00	1.12	1.39	1.72	2.48	45.95
0.30	4.08	1.37	1.35	1.39	1.54	1.84	2.23	3.01	46.04
0.50	4.98	1.88	1.86	1.91	2.09	2.40	2.85	3.74	49.61
0.70	6.70	3.13	3.06	3.14	3.34	3.76	4.27	5.27	58.54
0.99	92.48	88.38	87.22	86.24	90.19	88.15	89.32	90.63	201.19

Table 3.11: First-Order Autocorrelated, Exponentially Distributed Service and Interarrival Times, 50% utilization

Table 3.12: First-Order Autocorrelated, Exponentially Distributed Interarrival Times and Lognormal Service Times ($\delta = 2$), 50% utilization

M/G/1	50% Utilization				$ ho_S(1)$				
$ ho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	0.60	0.31	0.31	0.32	0.35	0.40	0.49	0.69	8.69
-0.70	0.51	0.37	0.37	0.39	0.41	0.46	0.53	0.65	5.60
-0.50	0.55	0.43	0.43	0.44	0.48	0.52	0.59	0.71	5.93
-0.30	0.71	0.50	0.51	0.52	0.55	0.61	0.68	0.81	5.70
0.00	0.81	0.67	0.68	0.69	0.73	0.80	0.88	1.03	7.02
0.30	1.21	1.01	1.02	1.03	1.09	1.17	1.27	1.46	7.55
0.50	1.78	1.48	1.50	1.53	1.58	1.66	1.79	2.02	8.22
0.70	3.06	2.67	2.65	2.70	2.74	2.88	3.03	3.33	10.48
0.99	85.29	87.43	86.66	85.77	88.83	86.27	86.97	86.05	117.26

and service times at the M/G/1 system is smaller than the M/M/1 system. This result is intuitive since the lognormal distribution with $\delta = 2$ has significantly smaller variance than the exponential distribution, and it also has a longer left tail. Similar to the M/M/1 and the G/M/1 systems, the impact of dependence increases in higher utilization levels. We refer the reader to Table 3.28 for the M/M/1 at 80% utilization and Table 3.29 for the M/G/1 with lognormal interarrival distributions ($\delta = 2$) at 80% utilization in the appendix. In addition the utilization effect, the nonmonotonic behavior of negatively autocorrelated interarrival and service times in the M/M/1 and G/M/1 systems occurs in this M/G/1 system as well.

					Correlation				
System	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
M/M/1, 50%	2.52	2.27	2.09	1.91	1.64	1.40	1.25	1.13	1.01
$\delta = 1$									
M/G/1, 50%	3.13	2.91	2.72	2.52	2.22	1.93	1.77	1.64	1.51
M/M/1, 80%	10.71	9.65	8.71	7.90	6.62	5.12	4.14	3.28	2.31
$\delta = 1$									
$M/G/1, \ 80\%$	13.05	12.06	11.26	10.37	8.88	7.24	6.12	5.06	3.92
M/M/1, 50%	1.73	1.55	1.43	1.31	1.13	0.96	0.86	0.78	0.69
$\delta=2$									
M/G/1, 50%	1.04	0.95	0.89	0.82	0.73	0.65	0.60	0.56	0.51
M/M/1, 80%	7.28	6.54	6.05	5.45	4.49	3.51	2.83	2.24	1.59
$\delta=2$									
$M/G/1, \ 80\%$	4.52	4.12	3.82	3.49	2.97	2.43	2.08	1.76	1.38
M/M/1, 50%	1.26	1.13	1.05	0.96	0.82	0.70	0.63	0.57	0.50
$\delta = 1$									
G/M/1, 50%	1.37	1.25	1.16	1.06	0.91	0.77	0.69	0.62	0.53
M/M/1, 80%	8.51	7.68	7.01	6.38	5.26	4.07	3.33	2.61	1.85
$\delta = 1$									
$G/M/1, \ 80\%$	9.89	9.12	8.50	7.80	6.58	5.23	4.30	3.41	2.32
M/M/1, 50%	0.86	0.77	0.72	0.66	0.56	0.48	0.43	0.39	0.34
$\delta=2$									
G/M/1, 50%	0.36	0.33	0.30	0.28	0.24	0.21	0.19	0.18	0.17
M/M/1, 80%	5.83	5.22	4.83	4.35	3.60	2.80	2.28	1.79	1.27
$\delta=2$									
G/M/1, 80%	3.23	2.91	2.68	2.44	2.05	1.66	1.41	1.18	0.91

 Table 3.13: Bivariate Dependence Between Exponentially Distributed Interarrival and Service Times

 Correlation

3.5.5. Bivariate Dependence Between Exponentially Distributed Interarrival and Service Times

In this experimental setup, there is only bivariate dependence between interarrival and service times of the same customer. In this simulation, we verify the results of the literature for the M/M/1 system, in which positively correlated interarrival and service times are expected to decrease the average waiting time and negatively correlated interarrival and service times increase the average waiting time. However, we observe same type of result for M/G/1 and G/M/1 systems, which is new in the literature. Table 3.13 presents the results for M/G/1 and G/M/1 results for 50% and 80% utilization levels with their benchmark results from the corresponding M/M/1 system. For example, the first row (M/M/1, 50%) shows the average waiting time of the M/M/1 system at 50% utilization level that the interarrival times are exponentially distributed with mean **1.65** and service times are exponentially distributed with mean **3.3**. Hence, the second row (M/G/1, 50%, $\delta = 1$) shows the average waiting time of the M/G/1 system at 50% utilization level that the interarrival time is exponentially distributed with mean **3.3** and the service time has a lognormal distribution with Johnson δ parameter equal to one. We refer the reader to Section 5.3 for more details about the lognormal distribution used in the M/G/1 and G/M/1 experiments.

The impact of correlation is higher in high utilizations for each experimental setup. 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 affects the system performance more. Regarding the nonexponential marginals for interarrival or service times, the impact is similar to temporal dependence results in Sections 5.3 and 5.4, so that the average waiting time increases in both M/G/1 and G/M/1 systems if the Johnson parameter δ is one for the lognormal distribution because of the higher variance. On the other hand, average waiting time decreases in both M/G/1 and G/M/1 systems if the Johnson parameter δ is two for the lognormal distribution because of the smaller variance. In addition to results for nonexponential marginals, the impact of bivariate dependence is smaller than temporal dependence on the average waiting of the single-server systems. The main difference between the impact of bivariate and temporal dependence in interarrival and service times on the average waiting time is the opposite effect of positive and negative dependencies. Thus, positive temporal dependence in interarrival and service times increases the average waiting time; on the other hand, positive bivariate dependence decreases the average waiting time; on the other hand, positive bivariate dependence decreases the average waiting time; on the other hand, positive bivariate dependence decreases the average waiting time; on the other hand, positive bivariate dependence decreases the average waiting time; on the other hand, positive bivariate dependence decreases the average waiting time; on the other hand, positive bivariate dependence decreases the average waiting time; on the other hand, positive bivariate dependence decreases the average waiting time; on the other hand, positive bivariate dependence decreases the average waiting time; on the other hand, positive bivariate dependence decreases the average waiting time; o

		Utilization		
Cross-correlation	25%	50%	66.7%	80%
-0.20	1.4092	5.6263	12.5837	26.8565
0.00	0.6285	2.3783	5.2879	11.3116
0.20	0.3219	1.2514	2.8245	6.1180
0.40	0.1548	0.6617	1.5531	3.4551
0.95	0.0001	0.0030	0.0207	0.0890

Table 3.14: First-Order Autocorrelated, Bivariate Dependent and Exponentially Distributed Interarrival and <u>Service Times</u>

time. This result is not new to the literature for the M/M/1 systems, but it is new for nonexponential interarrival or service times.

3.5.6. First-Order Autocorrelated, Bivariate Dependent and Exponentially Distributed Interarrival and Service Times

In this simulation, we investigate the impact of the cross-correlation between the interarrival and service times of two different jobs. The cross-correlation represents the temporal dependence between interarrival time of job i and service time of job j for $i \neq j$. This situation is a proxy for feedback systems in a telecommunication, manufacturing or service systems, where the manager can affect the service of a customer by observing the interarrival time of the previous customer. For example, if a server realizes really short interarrival times between customers, shorter service times are beneficial for the server in increasing the customer satisfaction. Thus, the server can adjust the service rate depending on how short or long the interarrival times are.

Since there are both temporal and bivariate dependent interarrival and service times in the M/M/1 system, lag-one autocorrelation in interarrival times, lag-one autocorrelation in service times and the bivariate dependence between the interarrival and service times are all equal to -0.40. Our results are robust to this initialization of temporal and bivariate dependencies in the interarrival and service times. In this M/M/1 system, the service time is exponentially distributed with mean 1 for each utilization level, and the interarrival time is exponentially distributed with mean 4 for utilization of 25%, 2 for utilization of 50%, 1.5 for utilization of 66.7% and 1.25 for utilization of 80%. We vary the lag-one cross-correlation from -0.20 to 0.95. Note that the VARTA model cannot generate simulation input for some extreme cross-correlation values such as -0.99 because of the failure. We refer the reader to Biller and Civelek [13] for the discussion about the failure probability of

					$ ho_S(1)$				
$ ho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	1.225	0.688	0.662	0.689	0.751	0.841	0.904	0.922	1.381
-0.70	0.846	0.235	0.020	0.028	0.094	0.150	0.169	0.194	1.072
-0.50	0.830	0.022	0.017	0.020	0.093	0.147	0.172	0.200	1.017
-0.30	0.851	0.023	0.018	0.020	0.132	0.186	0.217	0.250	1.068
0.00	0.863	0.026	0.020	0.024	0.191	0.213	0.250	0.274	1.101
0.30	0.928	0.258	0.239	0.262	0.328	0.382	0.427	0.484	1.171
0.50	0.943	0.417	0.392	0.418	0.467	0.566	0.590	0.671	1.282
0.70	0.971	0.690	0.673	0.700	0.760	0.821	0.930	0.977	1.301
0.99	1.982	1.348	1.312	1.422	1.621	1.714	1.863	1.926	4.722

Table 3.15: First-Order Autocorrelated, Exponentially Distributed Interarrival and Service Times for M/M/2, 40% utilization

this method.

Table 3.14 shows the simulation results for the impact of cross-correlation on the average waiting time of the M/M/1 system. Positive cross-correlation significantly decreases the average waiting time. For example, the average waiting time is very close to zero at 25% utilization level with **0.95** cross-correlation. Similar to both impacts of temporal and bivariate dependencies on the average waiting time, the impact of cross-correlation increases as a function of the utilization level of the single-server system. This result might gives a managerial insight for of a feedback system at the server such that positively correlated interarrival and service times of different jobs might increase the performance of the server. Note that the impact of the cross-correlation in the literature of dependence modeling in the queueing systems is new and novel, facilitated by of our input modeling scheme, the VARTA.

3.5.7. First-Order Autocorrelated Exponentially Distributed Interarrival and Service Times for Multi-Servers

In this simulation study, we investigate the impact of temporal dependence in interarrival and service times on the average waiting time of a multi-server system (M/M/k for $k \geq 2$). Note that there is no bivariate dependence between interarrival and service times. Considering upgrading the single-server system (M/M/1) to multi-server system (i.e., M/M/2), we can add either another identical server that decreases the utilization of the system by half or another server with half service rate by also reducing the original server's rate in order to keep the utilization level of the system constant. Therefore, we perform experiments in these two ways at 80% utilization level. Our initial system is an M/M/1 system, in

					$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	0.241	0.125	0.102	0.104	0.127	0.133	0.140	0.162	0.214
-0.70	0.180	0.018	0.017	0.018	0.020	0.020	0.029	0.033	0.183
-0.50	0.172	0.015	0.011	0.012	0.012	0.013	0.016	0.022	0.170
-0.30	0.183	0.015	0.015	0.016	0.016	0.017	0.018	0.024	0.172
0.00	0.190	0.018	0.018	0.019	0.024	0.025	0.030	0.033	0.191
0.30	0.227	0.041	0.030	0.038	0.040	0.049	0.054	0.061	0.200
0.50	0.233	0.051	0.041	0.044	0.053	0.062	0.075	0.084	0.230
0.70	0.248	0.060	0.055	0.061	0.071	0.084	0.091	0.103	0.246
0.99	0.253	0.148	0.138	0.149	0.159	0.163	0.179	0.190	0.358

Table 3.16: First-Order Autocorrelated, Exponentially Distributed Interarrival and Service Times for M/M/3, 26.67% utilization

which interarrival times are exponentially distributed with mean 1.25 and service times are exponentially distributed with mean 1.25. The results for impact of temporal dependent interarrival and service times for this system is shown in Table 3.20.

Tables 3.15 and 3.16 show the result for M/M/2 and M/M/3 with temporal dependent interarrival and service times. Note that, in these experiments we add identical servers such that the utilization drops to 40\$ for the M/M/2 system and 26.67% for the M/M/3 system. The theoretical average waiting times for independent and identically distributed interarrival and service times are 0.1905 and 0.0237 for the M/M/2 and M/M/3 systems, respectively. The impact of temporal dependence in both the M/M/2 and M/M/3 is similar to the M/M/1 system, but it is weaker. For instance, the average waiting time is 715.131 for the M/M/1 system, in which both interarrival and service times have lag-one autocorrelation of 0.99 (Table 3.20). However, the average waiting time is 4.722 for the M/M/2 system (Table 3.15) and 0.358 for the M/M/3 system (Table 3.16). This is intuitive since adding identical servers decreases the utilization significantly.

Additionally, we can keep the utilization level constant while increasing the number of the servers in the system. In the next experiments for the M/M/2 and M/M/3 systems, the interarrival times are exponentially distributed with mean 1.25. However, the service times are now exponentially distributed with mean 2 for the M/M/2 system and 3 for M/M/3 system. Note that the utilization level is 80% in both M/M/2 and M/M/3 systems. The theoretical average waiting time for independent and identically distributed interarrival and service times are 3.5556 and 3.236 for these M/M/2 and M/M/3 systems, respectively. Similar to previous experiments for the multi-server systems, Tables 3.17 and 3.18 show that

					$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	9.280	5.812	5.554	5.710	5.823	5.913	6.320	7.812	25.801
-0.70	6.103	2.810	2.619	3.118	3.242	3.318	3.446	3.672	19.300
-0.50	5.952	2.881	2.750	3.027	3.229	3.252	3.302	3.417	18.928
-0.30	6.012	2.910	2.816	3.157	3.414	3.438	3.469	3.631	19.322
0.00	6.038	3.117	2.991	3.218	3.558	4.527	5.278	8.138	20.025
0.30	8.128	3.718	3.450	3.671	4.612	4.691	5.491	9.553	20.912
0.50	9.618	5.016	4.881	5.439	6.222	6.710	6.966	10.157	21.357
0.70	10.392	8.159	7.714	8.557	9.172	9.437	9.865	12.221	23.183
0.99	19.610	18.358	18.273	19.002	19.832	19.920	20.155	21.442	37.001

Table 3.17: First-Order Autocorrelated, Exponentially Distributed Interarrival and Service Times for M/M/2, 80% utilization

Table 3.18: First-Order Autocorrelated, Exponentially Distributed Interarrival and Service Times for M/M/3, 80% utilization

					$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	6.038	4.689	3.926	4.115	4.462	4.719	5.102	7.119	14.062
-0.70	4.820	2.739	2.551	2.710	3.183	3.223	3.742	5.528	11.339
-0.50	4.005	2.691	2.400	2.631	2.918	3.038	3.618	4.929	10.018
-0.30	4.017	2.855	2.539	2.957	3.005	3.282	3.990	5.302	10.431
0.00	4.482	2.973	2.687	3.016	3.200	3.517	4.371	7.002	12.093
0.30	4.917	3.192	2.917	3.257	3.326	3.648	4.581	8.013	13.073
0.50	5.400	3.821	3.519	3.953	4.038	4.513	4.956	8.151	14.103
0.70	6.211	4.383	4.014	4.719	5.350	6.410	7.104	8.518	15.821
0.99	13.891	11.038	10.832	11.018	11.984	12.281	12.911	13.431	19.719

the impact of the temporal dependence in both M/M/2 and M/M/3 systems is weaker than the M/M/1 system. However, the impact of temporal dependence in interarrival and services times on the average waiting time of multi-server systems is greater than the previous case with identical servers because of higher utilization levels.

3.6. Conclusion

In this paper, we use an advanced simulation input modeling (VARTA) to study the impact of bivariate and temporal dependencies among interarrival and service times on the performance of a single-server queue. The distinguishing feature of our study from those in the literature is to consider a wide variety of distributional shapes for the probability density functions of the interarrival and service times and the patterns that arise in the temporal dependencies of the interarrival and service times. We investigate the impact of dependent interarrival and service times on the average waiting time of M/M/1, M/G/1, G/M/1, M/M/2 and M/M/3 systems. Our main contribution is to combine this advanced input modeling method with queueing theory for investigating the impacts of dependent interarrival and service demands on the performance of a single-server queue.

In our simulation studies, we observe that positively autocorrelated interarrival times and/or service times always increase the average waiting time and the impact of this positive temporal dependence is monotonically increasing in utilization of the system. The average waiting time is monotonically increasing in the value of autocorrelation in interarrival times and/or service times. These results confirm the literature. However, the impact of dependence is interesting in negatively autocorrelated interarrival times and/or service demands. In the literature, nonmonotonic behavior of the average waiting time in negatively autocorrelated service demands was observed, but nonmonotonic behavior of the performance in negatively autocorrelated interarrival times is a new observation in the literature. Because nonmonotonic behavior of the average waiting time in negatively autocorrelated interarrivals exist when lag-one autocorrelation is close to minus one and/or utilization is close to 100%.

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.

One major contribution of our simulation study is to investigate the impact of temporal and bivariate dependencies on the average waiting time of single-server queue with nonexponential marginals for interarrival and service times. We perform simulations for the M/G/1and G/M/1 systems and use M/M/1 as a benchmark. We use two different lognormal distributions with two different tails comparing to the exponential distribution. The average waiting times of all temporal and bivariate dependencies of interarrival and service times in both G/M/1 and M/G/1 systems is larger than the M/M/1 system for the lognormal distribution that has longer right tail and higher variance than the exponential distribution. On the other hand, the average waiting time decreases in both G/M/1 and M/G/1 systems for the lognormal distribution that has longer left tail and smaller variance than the exponential distribution. Furthermore, our study is differentiated from other simulation studies in the literature of dependent queues by combining 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 temporal dependence between interarrival and service times of different jobs at the server. 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.

					$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	4.018	0.536	0.506	0.535	0.666	0.959	1.362	2.293	48.786
-0.70	2.695	0.561	0.548	0.576	0.676	0.883	1.153	1.772	38.937
-0.50	2.663	0.616	0.604	0.633	0.731	0.934	1.200	1.798	39.424
-0.30	2.858	0.691	0.679	0.709	0.811	1.017	1.294	1.897	39.951
0.00	3.161	0.867	0.853	0.886	0.999	1.229	1.525	2.171	40.984
0.30	3.701	1.205	1.186	1.227	1.361	1.626	1.967	2.677	41.721
0.50	4.430	1.663	1.645	1.692	1.843	2.145	2.521	3.325	44.450
0.70	6.073	2.741	2.716	2.766	2.958	3.322	3.780	4.714	45.909
0.99	87.840	77.596	77.484	77.916	78.235	79.369	79.364	82.887	167.239

Table 3.19: First-Order Autocorrelated, Exponentially Distributed Interarrival and Service Times, 50% utilization

Table 3.20: First-Order Autocorrelated, Exponentially Distributed Interarrival and Service Times, 80% utilization

					$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	46.02	8.59	8.29	8.75	10.19	13.14	16.69	23.75	361.24
-0.70	29.19	2.41	2.24	2.40	3.04	4.39	6.23	10.55	322.03
-0.50	28.77	2.45	2.28	2.43	3.04	4.34	6.15	10.41	316.78
-0.30	29.91	2.69	2.52	2.67	3.27	4.58	6.38	10.66	321.69
0.00	30.42	3.41	3.24	3.39	4.00	5.34	7.17	11.49	319.54
0.30	32.26	4.92	4.74	4.89	5.54	6.94	8.82	13.28	321.06
0.50	35.26	7.00	6.82	6.99	7.67	9.16	11.08	15.62	330.26
0.70	39.93	11.98	11.76	12.00	12.69	14.19	16.26	21.14	330.74
0.99	398.19	362.89	361.87	365.63	366.38	368.03	368.85	380.219	715.13

Table 3.21: First-Order Autocorrelated, Exponentially Distributed Interarrival and Service Times, 99% utilization

					$ ho_S(1)$				
$ ho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	807.9	535.7	538.4	532.5	534.3	554.2	572.8	603.9	2126.2
-0.70	535.1	65.6	61.2	63.9	79.2	113.8	153.4	238.2	2089.1
-0.50	545.3	59.9	56.2	59.2	74.3	105.4	149.7	240.1	2081.8
-0.30	538.3	63.9	59.9	63.0	79.2	109.2	155.8	237.1	2118.2
0.00	551.5	80.8	74.1	78.3	99.2	124.9	164.3	252.5	2056.6
0.30	563.5	114.5	104.5	113.6	124.9	155.4	193.5	273.29	2096.3
0.50	592.1	157.1	149.9	153.6	167.6	194.6	224.1	309.40	2030.9
0.70	636.4	242.6	242.1	229.4	249.2	283.1	307.5	368.18	2084.2
0.99	2222.7	2051.8	2118.8	2061.4	2114.9	2144.1	2189.8	2074.04	3007.5

M/M/1	80% Utilization				$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	59.2	11.5	10.9	11.6	13.2	17.2	22.1	31.4	407.7
-0.70	39.7	3.2	2.9	3.2	4.0	5.8	8.2	14.3	387.9
-0.50	41.1	3.2	3.0	3.2	4.0	5.7	8.2	13.6	340.6
-0.30	40.5	3.6	3.3	3.5	4.3	6.0	8.5	14.0	361.3
0.00	40.6	4.5	4.3	4.5	5.3	6.9	9.6	14.9	438.9
0.30	41.4	6.5	6.3	6.5	7.4	9.2	11.5	17.7	442.1
0.50	45.4	9.2	8.9	9.2	10.1	12.1	14.5	20.8	448.7
0.70	54.6	15.5	15.6	15.8	16.4	18.7	21.7	28.1	501.6
0.99	506.7	456.8	465.6	474.4	500.7	502.5	512.6	521.8	891.0

Table 3.22: First-Order Autocorrelated, Exponentially Distributed Service and Interarrival Times, 80% utilization

Table 3.23: First-Order Autocorrelated, Exponentially Distributed Service Times and Lognormal Interarrival Times ($\delta = 1$), 80% utilization

G/M/1	80% Utilization				$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	99.9	44.2	42.1	43.6	43.6	49.9	56.4	68.0	479.4
-0.70	43.6	4.7	4.5	4.7	5.8	7.8	10.5	17.1	395.9
-0.50	43.2	4.5	4.2	4.4	5.4	7.3	9.9	15.6	397.8
-0.30	41.8	4.7	4.4	4.7	5.6	7.4	10.1	15.7	365.9
0.00	42.1	5.7	5.5	5.7	6.6	8.6	11.1	17.3	453.0
0.30	46.1	8.2	7.9	8.1	9.2	11.1	13.7	20.3	454.7
0.50	47.8	11.6	11.4	11.7	12.7	14.8	17.6	24.2	455.4
0.70	61.4	19.8	19.8	20.2	21.1	23.3	26.4	33.3	610.3
0.99	606.7	585.0	617.7	592.4	590.8	597.8	609.6	617.3	1055.8

Table 3.24: First-Order Autocorrelated, Exponentially Distributed Service and Interarrival Times, 80% utilization

M/M/1	80% Utilization				$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	40.7	7.8	7.5	7.8	9.3	11.9	14.8	21.5	301.2
-0.70	25.2	2.2	2.0	2.2	2.8	3.9	5.7	9.4	280.6
-0.50	26.7	2.2	2.0	2.2	2.8	3.9	5.6	9.4	274.7
-0.30	26.6	2.4	2.3	2.4	2.9	4.2	5.8	9.7	275.8
0.00	26.5	3.1	2.9	3.0	3.6	4.8	6.6	10.5	300.7
0.30	29.6	4.4	4.3	4.4	4.9	6.3	7.8	12.1	303.4
0.50	31.1	6.4	6.2	6.3	6.9	8.2	9.9	14.0	304.5
0.70	37.5	10.9	10.7	10.8	11.4	12.9	14.7	19.1	308.3
0.99	357.8	318.4	313.9	314.6	319.1	325.8	334.5	352.7	717.6

G/M/1	80% Utilization				$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	25.99	1.56	1.37	1.51	2.15	3.54	5.38	9.84	266.26
-0.70	23.46	1.32	1.18	1.29	1.78	2.87	4.48	8.08	264.99
-0.50	25.34	1.35	1.20	1.32	1.80	2.90	4.42	8.14	265.66
-0.30	25.04	1.43	1.28	1.39	1.87	2.98	4.51	8.20	267.60
0.00	23.75	1.59	1.45	1.56	2.07	3.18	4.78	8.60	294.46
0.30	25.29	1.93	1.76	1.89	2.41	3.56	5.19	8.90	295.53
0.50	24.76	2.33	2.20	2.33	2.89	4.05	5.72	9.49	296.35
0.70	27.87	3.28	3.13	3.28	3.88	5.18	6.91	11.17	298.83
0.99	98.30	65.05	64.95	68.18	66.43	68.34	69.94	77.18	391.22

Table 3.25: First-Order Autocorrelated, Exponentially Distributed Service Times and Lognormal Interarrival Times ($\delta = 2$), 80% utilization

Table 3.26: First-Order Autocorrelated, Exponentially Distributed Service and Interarrival Times, 80% utilization

M/M/1	80% Utilization				$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	74.3	14.3	13.7	14.3	17.0	21.2	27.0	39.9	574.8
-0.70	44.5	4.0	3.7	3.9	5.0	7.3	10.3	17.3	468.9
-0.50	46.6	4.1	3.8	4.0	5.0	7.2	10.2	17.3	407.7
-0.30	49.2	4.4	4.2	4.4	5.4	7.5	10.5	17.5	434.5
0.00	52.3	5.6	5.3	5.6	6.6	8.8	11.9	18.8	459.1
0.30	52.3	8.1	7.9	8.0	9.2	11.5	14.4	22.1	509.7
0.50	60.9	11.6	11.3	11.5	12.6	15.2	18.3	25.8	548.8
0.70	72.2	19.9	19.0	19.8	20.9	23.5	27.1	34.6	582.8
0.99	685.4	633.5	565.0	576.5	580.5	597.7	603.1	609.6	1039.7

Table 3.27: First-Order Autocorrelated, Exponentially Distributed Interarrival Times and Lognormal Service Times ($\delta = 1$), 80% utilization

M/G/1	80% Utilization				$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	188.4	21.7	18.9	19.2	21.2	26.4	34.8	49.9	821.7
-0.70	142.4	7.9	6.4	6.3	7.5	10.5	14.9	25.9	724.2
-0.50	153.2	7.7	6.3	6.2	7.4	10.4	14.8	24.8	709.3
-0.30	155.2	8.2	6.7	6.7	7.8	10.6	15.1	25.8	748.4
0.00	179.2	9.2	7.9	7.9	8.9	12.2	16.2	26.7	756.9
0.30	162.7	12.1	10.4	10.3	11.5	14.5	19.3	29.8	764.5
0.50	182.5	15.5	13.9	13.6	15.2	18.3	22.8	33.8	815.5
0.70	197.7	24.1	21.8	22.0	23.2	26.6	30.8	42.8	820.8
0.99	795.9	642.9	569.6	579.3	584.2	603.4	607.3	614.04	1373.3

111100, 0070	aumzauon								
M/M/1	80% Utilization				$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	50.6	9.8	9.3	9.8	11.4	15.0	18.7	26.7	396.2
-0.70	33.4	2.7	2.5	2.7	3.4	4.9	7.0	12.0	351.8
-0.50	31.6	2.7	2.5	2.7	3.4	4.9	7.0	11.6	336.9
-0.30	32.6	3.0	2.8	3.0	3.6	5.1	7.1	11.9	340.0
0.00	38.3	3.8	3.6	3.8	4.6	6.0	8.2	12.8	344.2
0.30	34.5	5.5	5.3	5.5	6.2	7.8	9.8	15.1	379.1
0.50	40.2	7.8	7.6	7.8	8.6	10.3	12.5	17.2	398.9
0.70	44.6	13.4	13.3	13.6	14.4	16.0	18.2	23.7	414.0
0.99	448.3	434.5	429.6	397.0	400.9	413.0	424.5	442.6	842.6

Table 3.28: First-Order Autocorrelated, Exponentially Distributed Service and Interarrival Times, 80% utilization

Table 3.29: First-Order Autocorrelated, Exponentially Distributed Interarrival Times and Lognormal Service Times ($\delta = 2$), 80% utilization

$\frac{1}{2} = \frac{1}{2}, \frac{1}{2}, \frac{1}{2} = \frac{1}{2}, \frac{1}{2}, \frac{1}{2} = \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{$									
M/G/1	80% Utilization				$ ho_S(1)$				
$\rho_A(1)$	-0.99	-0.70	-0.50	-0.30	0.00	0.30	0.50	0.70	0.99
-0.99	14.2	7.0	7.0	7.1	7.9	8.9	10.7	13.7	115.8
-0.70	5.4	1.4	1.4	1.5	1.7	2.2	2.8	4.1	78.6
-0.50	5.1	1.5	1.5	1.6	1.8	2.3	2.8	4.1	70.2
-0.30	5.6	1.8	1.8	1.9	2.1	2.5	3.1	4.3	72.3
0.00	7.1	2.6	2.6	2.7	2.9	3.3	3.9	5.3	79.6
0.30	8.2	4.3	4.2	4.3	4.6	5.1	5.7	7.1	88.9
0.50	11.2	6.5	6.6	6.7	6.9	7.4	8.1	9.6	91.8
0.70	16.7	12.2	12.3	12.5	12.5	13.2	13.7	15.3	102.9
0.99	417.8	430.3	426.7	393.7	399.5	411.2	421.6	437.86	565.9

Bibliography

- AABB. 2005. American Association of Blood Banks. Nationwide Blood Collection and Utilization Survey Report. www.aabb.org.
- [2] AABB. 2007. American Association of Blood Banks. Nationwide Blood Collection and Utilization Survey Report. www.aabb.org.
- [3] Altiok, T. 2001. The case for modeling correlation in manufacturing systems. *IIE Trans*actions, Vol. 33, 779–791.
- [4] Altman, L. K. 2001. Donors flood blood banks, but a steady stream is what's needed. The New York Times, 18 Sept 2001.
- [5] American Red Cross homepage for blood donation. 2008. http://www.givelife2.org/donor /default.asp.
- [6] Arslan, H, S. C. Graves, T. A. Roemer. 2007. A single-product inventory model for multiple demand classes. *Management Science*, Vol. 53, No. 9, 1486-1500.
- [7] Balter, M. H., A. Downey. 1996. Exploiting process lifetime distributions for dynamic load balancing. SIGMETRICS'96, Philedelphia, Pennsylvania, USA, 13–24.
- [8] Biller, B., B. L. Nelson. 2003. Modeling and generating multivariate time-series input processes using a vector autoregressive technique. ACM TOMACS, Vol. 13, No. 3, 211– 237.
- [9] Biller, B., B. L. Nelson. 2005. Fitting time-series input processes for simulation. Operations Research, 53, 549–559.
- [10] Biller, B. 2005. Multivariate time-series input processes for simulation. Working Paper. Tepper School of Business, Carnegie Mellon University.

- [11] Biller, B. and S. Ghosh. 2006. Multivariate input processes. In Handbooks in Operations Research and Management Science: Simulation, ed. B. L. Nelson and S. G. Henderson. Elsevier Science, Amsterdam.
- [12] Biller, B. 2007. Copula-based multivariate input models for stochastic simulation. Tepper School of Business Working Paper, Carnegie Mellon University, Pittsburgh, PA.
- [13] Biller, B., I. Civelek. 2009. Failure probability of VARTA in high-dimensional settings. In revision at *INFORMS Journal of Computing*.
- [14] Boucherie, R. J., T. Huisman. 2002. The sojourn time distribution in an infinite server resequencing queue with dependent interarrival and service times. J. Appl. Prob., Vol. 39, 590–603.
- [15] Brodheim, E. C. Derman, G. P. Prastacos. 1975. On the evaluation of a class of inventory policies for perishable products such as blood. *Management Science*, Vol. 21, 1320-1326.
- [16] Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2002. Statistical analysis of a telephone call center: A queueing-science perspective. *Technical report*, The Wharton School of Business, University of Pennsylvania, Philadelphia, PA.
- [17] Buzacott, J. A., J. G. Shanthikumar. 1993. Stochastic models of manufacturing systems. Prentice Hall, Inc.
- [18] Cario, M. C. and B. L. Nelson. 1996. Autoregressive to anything: Time-series input processes for simulation. Operations Research Letters, 19, 51–58.
- [19] Cario, M. C. and Nelson, B. L. 1997. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Working Paper, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.
- [20] Cario, M. C., B. L. Nelson. 1998. Numerical methods for fitting and simulating autoregressive-to-anything processes. *INFORMS J. of Computing*, Vol. 10, 72–81.
- [21] Chao, X. 1995. Monotone effect of dependency between interarrival and service times in a simple queueing system. Operations Research Letters, Vol. 17, 47–51.

- [22] Chen, H. 2001. Initialization of NORTA: Generation of random vectors with specified marginals and correlations. *INFORMS J. of Computing*, Vol. 13, 312–331.
- [23] Cole, G. 1994. Pharmaceutical production facilities. Ellis Horwood Publishers, London.
- [24] Condon, D. 2006. Demand for platelets up 50% in 5yrs. www.irishhealth.com, 18 Oct 2006.
- [25] Crovella, M. E., A. Bestavros. 1996. Self-similarity in world wide web traffic evidence and possible causes. *SIGMETRICS'96*, Philedelphia, Pennsylvania, USA, 160–168.
- [26] Dekker, R., R. M. Hill, M. J. Kleijn, R. H. Teunter. 2002. On the (S-1,S) lost sales inventory model with priority demand classes. *Naval Res. Logist.*, Vol. 49, 593-610.
- [27] Deniz, B., I. Karaesmen, A. Scheller-Wolf. 2009. Managing Perishables with Substitution: Issuance and Replenishment Heuristics. To appear in *Manufacturing and Service Operations Management*.
- [28] Deshpande, V., M. A. Cohen, K. Donohue. 2003. A threshold rationing policy for service differentiated demand classes. *Management Science*, Vol. 49, No. 6, 683-703.
- [29] Fendick, K. W., V. R. Saksena, W. Whitt. 1989. Dependence in packet queues. *IEEE Transactions on Communications*, Vol. 37, No. 11, 1173–1183.
- [30] Fontaine, M. J., F., Y.T. Chung, W. M. Rogers, H. D. Sussmann, P. Quach, S. A. Galel, L. T. Goodnough, F. Erhun. 2009. Improving Platelet Supply Chains through Collaborations between Blood Centers and Transfusion Services. *Transfusion*, Vol. 49(10), 2040-2047.
- [31] Gallego, G., G. Van Ryzin. 1994. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science*, Vol. 40, No. 8, 999–1020.
- [32] Ghosh, S., S. G. Henderson. 2003. Behavior of NORTA Method for Correlated Random Vector Generation as the Dimension Increases. ACM TOMACS, 13, 276–294.
- [33] Ghosh, S., S. G. Henderson. 2002. Chessboard distributions and random vectors with specified marginals and covariance matrix. *Operations Research*, 50, 820-834.

- [34] Guttman, L. 1946. Enlargement methods for computing the inverse matrix. Annals of Mathematical Statistics, 17, 336-343.
- [35] Hadidi, N. 1985. Further results on queues with partial correlation. Operations Research, Vol. 33, No. 1, 203–209.
- [36] Haijema, R., van der Wal, J., van Dijk, N. 2009. Blood platelet production: a novel approach for practical optimization. *Transfusion*, Vol. 49, 411-419.
- [37] Haijema, R., van der Wal, J., van Dijk, N. 2007. Blood platelet production: Optimization by dynamic programming and simulation. *Computers and Operations Research*, Vol. 34, 760-779.
- [38] Heffes, H. D. M. Lucantoni. 1986. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE J. on Selected Areas in Communications*, SAC-4, 856–868.
- [39] Hersh, E. M., G. P. Bodey, B. A. Nies, E. J. Freireich. 1965. Causes of death in acute leukemia: a ten-year study of 414 patients from 1954-1963. JAMA, Vol. 193, 105-109.
- [40] Hopp, W. J., M. L. Spearman. 1996. Factory physics. Irwin McGraw-Hill, Inc.
- [41] Iyer, S. K., D. Manjunath. 2006. Queues with dependency between interarrival and service times using mixtures of bivariates. *Stochastic Models*, Vol. 22, No. 1, 3–20.
- [42] Iravani, S. M. R., K. L. Luangkesorn, D.Simchi-Levi. 2004. A general decomposition algorithm for parallel queues with correlated arrivals. *Queueing Systems*, Vol. 47, 313– 344.
- [43] Jacobs, P. A. Heavy traffic results for single server queues with dependent (EARMA) service and interarrival times. Advances in Applied Probability, Vol. 12, 517–529.
- [44] Johnson, N. L. 1949. Systems of frequency curves generated by methods of translation. Biometrika, Vol. 36, 149176.
- [45] Karaesmen, I. Z., A. Scheller-Wolf, B. Deniz. 2007. Managing perishable and aging inventories: Review and future research directions. To appear in *Handbook of Production Planning* (K. Kempf, P. Keskinocak, R. Uzsoy, eds), Kluwer International Series in Operations Research and Management Science, Kluwer Academic Publishers.

- [46] Kelly, F. P. 1979. Reversibility and Stochastic Networks. Wiley.
- [47] Kendall, M. G. 1961. A Course in the Geometry of n Dimensions. Charles Griffin and Co., London, England.
- [48] Kopach, R., B. Balcioglu, M. Carter. 2008. Tutorial on constructing a red blood cell inventory management system with two demand rates. *EJOR*, Vol. 185, 1051-1059.
- [49] Kranenburg, A. A., G. J. van Houtum. 2007. Cost optimization inn the (S-1,S) lost sales inventory model with multiple demand classes. *Operations Res. Let.*, Vol. 35, 493-502.
- [50] Kurowicka, D. and R. Cooke. 2006. Uncertainty Analysis with High Dimensional Dependence Modeling. Wiley Series in Probability and Statistics.
- [51] Landro, L. 2009. New Swine Flu Victim: Blood Supply. The Wall Street Journal, November 10, 2009.
- [52] Langaris, C. 1986. A correlated queue with infinitely many servers. J. App. Prob., 23, 155–165.
- [53] Law, A. M., W. D. Kelton. 2000. Simulation Modeling and Analysis, 3PrdP ed. McGraw Hill, Boston, MA.
- [54] Lewis, P. A. W., E. McKenzie. 1991. Minification processes and their transformations. J. Appl. Prob., Vol. 28, 45–57.
- [55] Li, H., S. H. Xu. 2000. On the dependence structure and bounds of correlated parallel queues and their applications to synchronized stochastic systems. J. Appl. Prob., Vol. 37, 1020–1043.
- [56] Li, S. T., J. L. Hammond. 1975. Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. *IEEE Trans. Syst. Man. and Cybernet*, 5, 557–561.
- [57] Livny, M., B. Melamed and A. K. Tsiolis. 1993. The impact of autocorrelation on queuing systems. *Management Science*, Vol. 39, No. 3, 322–339.
- [58] Lurie, P. M., M. S. Goldberg. 1998. An approximate method for sampling correlated random variables from partially-specified distributions. *Management Science*, 44, 203– 218.

- [59] Lütkepohl, H. 1993. Introduction to Multiple Time Series Analysis. Springer-Verlag, New York.
- [60] Mallows, C. L. 1967. Linear processes are nearly Gaussian. Journal of Applied Probability, 4, 313–329.
- [61] Marsaglia G., I. Olkin. 1984. Generating correlation matrices. SIAM Journal of Scientific and Statistical Computing, 5, 470-475.
- [62] Melamed, B., J. R. Hill and D. Goldsman. 1992. The TES methodology: Modeling empirical stationary time series. In *Proceedings of the 1992 Winter Simulation Conference*, ed. J. J. Swain, D. Goldsman, R. C. Crain and J. R. Wilson, 135–144.
- [63] Moroff, G. 2008. Transfusion of Platelets: Current Issues. American Red Cross Biomedical Services, Medical and Scientific Updates, Number 8-6.
- [64] Moss, M. 2009. Hospitals facing blood platelet crisis. Scotland on Sunday, 29 Sept 2009.
- [65] Nahmias, S. 1982. Perishable inventory theory problem. Operations Research, Vol. 30, No. 4, 680-708.
- [66] Nelson, B., M. R. Taaffe. 2004. The Ph_t/Ph_t/∞ queueing system: Part I–Single node. INFORMS J. of Computing, Vol. 16, No. 3, 266–274.
- [67] Nelson, B., M. R. Taaffe. 2004. The [Ph_t/Ph_t/∞]^K queueing system: Part II–The multiclass network. INFORMS J. of Computing, Vol. 16, No. 3, 275–283.
- [68] Ouellette, D. V. 1981. Schur complements and statistics. Linear Algebra and Its Applications, 36, 187-295.
- [69] Patuwo, B. E., R. L. Disney, D. C. McNickle. 1993. The effect of correlated arrivals on queues. *IIE Transactions*, Vol. 25, No. 3, 105–110.
- [70] Pereboom, I. T. A., T. Lisman, R. J. Porte. 2008. Platelets in liver transplantation: friend or foe? *Liver Transplantation*, Vol. 14, 923-931.
- [71] Pierskalla, W. P. 2004. Supply Chain Management of Blood Products. In Operations Research and Health Care. Edited by M. L. Brandeau, F. Sainfort, and W. P. Pierskalla. International Series in Operations Research and Management Science, Springer New York, 2004.

- [72] Prastacos, G. P. 1981. Allocation of a perishable product inventory. Operations Research, Vol. 29, No. 1, 95-107.
- [73] Prastacos, G. P. 1982. Blood inventory management problem: An overview of theory and practice. *Management Science*, Vol. 30, No. 7, 777-799.
- [74] Runnenburg, J. Th. 1961. An example illustrating the possibilities of renewal theory and waiting-time theory for Markov-dependent arrival intervals. Proc. Ser. A. Kon. Neder. Akad. Weten., Vol. 64, 560–576.
- [75] Runnenburg, J. Th. 1962. Some numerical results on waiting-time distributions for dependent arrival-intervals. *Statistica Neerlandica*, Vol. 16, No. 4, 337–348.
- [76] Shioda, S. 2003. Departure process of the MAP/SM/1 queue. Queueing Systems, Vol. 44, 31–50.
- [77] Song, W. T., L. Hsiao, Y. Chen. 1996. Generating pseudorandom time series with specified marginal distributions. *European Journal of Operations Research*, Vol. 93, 1– 12.
- [78] Standridge, C. R. 2004. How factory physics helps simulation. Proceedings of the 2004 Winter Simulation Conference, ed. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, pp. 1103–1108.
- [79] Stroncek, D. F., P. Rebulla. 2007. Transfusion Medicine 2: Platelet transfusions. The Lancet, Vol. 370, No. 9585, 427-438.
- [80] Sullivan, M. T., E. L. Wallace. 2005. Blood collection and transfusion in the United States in 1999. *Transfusion*, Vol. 45, No. 2, 141-148.
- [81] Szekli, R., R. L. Disney, S. Hur. 1994. MR/GI/1 queues with positively correlated arrival stream. J. Appl. Prob., Vol. 31, 497–514.
- [82] Takahashi, K., N. Nakamura. 1998. The effect of autocorrelated demand in JIT production systems. International Journal of Production Research, 36, 5, 1159–1176.
- [83] Tong, Y. L. 1990. The Multivariate Normal Distribution. New York: Springer-Verlag.

- [84] Topkis, D. M. 1968. Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. *Management Science*, Vol. 15, 160176.
- [85] Veinott, A. F. 1965. Optimal policy in a dynamic, single product, non-stationary inventory model with several demand classes. *Operations Research*, Vol. 13, 761-778.
- [86] Vincent, S. 1998. Input data analysis. In Handbook of Simulation, ed. J. Banks, 55–91. New York: John Wiley & Sons.
- [87] 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.
- [88] Wolff, R. W. 1989. Stochastic modeling and the theory of queues. Prentice Hall, Inc. 293–296.
- [89] Wei, W. W. S. 1990. Time series analysis, univariate and multivariate methods. Redwood City, CA: Addison-Wesley.
- [90] Wilson, K., P. Hebert. 2003. The challenge of an increasingly expensive blood system. Canadian Med. Assoc. J., Vol. 168, 1149-1150.
- [91] Xu, S. H. 1999. Structural analysis of a queueing system with multiclass of correlated arrivals and blocking. Operations Research, Vol. 47, No. 2, 264–276.
- [92] Zhao, H., J. K. Ryan, V. Deshpande. 2008. Optimal Dynamic Production and Inventory Transshipment Policies for a Two-Location Make-to-Stock System. *Operations Research*, Vol. 56, No. 2, 400-410.