

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE Jan 96		3. REPORT TYPE AND DATES COVERED
4. TITLE AND SUBTITLE Multivariate Sampling With Explicit Correlation Induction For Simulation and Optimization Studies				5. FUNDING NUMBERS
6. AUTHOR(S) Raymond R. Hill Jr., Ph.D.				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AFIT Student Attending: Ohio State University				8. PERFORMING ORGANIZATION REPORT NUMBER 96-003D
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) DEPARTMENT OF THE AIR FORCE AFIT/CI 2950 P STREET, BLDG 125 WRIGHT-PATTERSON AFB OH 45433-7765				10. SPONSORING / MONITORING AGENCY REPORT NUMBER
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release IAW AFR 190-1 Distribution Unlimited BRIAN D. GAUTHIER, MSgt, USAF Chief Administration				12b. DISTRIBUTION CODE
13. ABSTRACT (Maximum 200 words)				
14. SUBJECT TERMS				15. NUMBER OF PAGES 115
				16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT		18. SECURITY CLASSIFICATION OF THIS PAGE		19. SECURITY CLASSIFICATION OF ABSTRACT
20. LIMITATION OF ABSTRACT				

19960531 074

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to *stay within the lines* to meet *optical scanning requirements*.

Block 1. Agency Use Only (Leave blank).

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C - Contract	PR - Project
G - Grant	TA - Task
PE - Program Element	WU - Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. (If known)

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."

DOE - See authorities.

NASA - See Handbook NHB 2200.2.

NTIS - Leave blank.

Block 12b. Distribution Code.

DOD - Leave blank.

DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.

NASA - Leave blank.

NTIS - Leave blank.

Block 13. Abstract. Include a brief (*Maximum 200 words*) factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code (*NTIS only*).

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

MULTIVARIATE SAMPLING WITH EXPLICIT CORRELATION INDUCTION FOR SIMULATION AND OPTIMIZATION STUDIES

By

Raymond R. Hill Jr., Ph.D.

The Ohio State University, 1996

Professor Charles H. Reilly, Advisor

Composite distributions based on specified marginal distributions and a specified Pearson product-moment correlation structure are formed by mixing extreme-correlation distributions of a multivariate random variable and the joint distribution under independence. Closed-form expressions are provided for the composition probabilities for composite distributions for trivariate random variables, and a simple algorithm for finding composition probabilities in the case of quadrivariate random variables is presented. A linear program provides a general approach for finding composition probabilities. For all but the extreme correlation structures a range of composite distributions is provided. Composite distributions are used to generate coefficients for 1120 two-dimensional knapsack problems based on a variety of Pearson correlation structures. An equal number of problems is generated based on Spearman rank correlation structures. The computational results with a branch-and-bound procedure and a well-known heuristic indicate that the type of

correlation structure induced (Pearson or Spearman) can affect the performance of solution procedures. The correlation structure specified matters, as do the values specified for each correlation term. There is a noticeable interaction between the correlation structure induced and the constraint slackness settings. Finally, the interconstraint correlation is found to affect solution procedure performance more than either of the objective-constraint correlations.

Maj Dietrich,

2 Jan 96

Here's the final, approved dissertation. I passed my final exam on 7 Dec 95. You should be receiving my grades along with the rest of the OSU grades (direct from the ROTC Det 645). My official commencement is 15 Mar 96.

Thanks for your help.

I'm DONE !!

Maj Ray Hill

AFSAA/SAGW

1570 Air Force Pentagon

Washington DC 20330-1570

DSN: 227-5677

MULTIVARIATE SAMPLING WITH EXPLICIT
CORRELATION INDUCTION FOR SIMULATION AND
OPTIMIZATION STUDIES

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By

Raymond R. Hill Jr., B.S., M.S.

* * * * *

The Ohio State University

1996

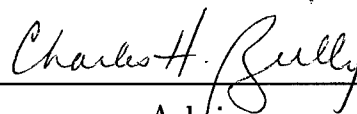
Dissertation Committee:

Charles H. Reilly

Marc E. Posner

Barry L. Nelson

Approved by:



Advisor

Industrial and Systems
Engineering Graduate
Program

To Christy - Love and Thanks

ACKNOWLEDGEMENTS

I express my sincere appreciation to Dr. Charles H. Reilly for his guidance and insight throughout my studies and research at The Ohio State University. I also wish to thank the other members of my committee, Drs. Barry L. Nelson and Marc E. Posner for their assistance and insight in both class work and this research effort. A special thanks goes to the Operational Research Department of the Air Force Institute of Technology for providing me the computational resources I required. I especially thank my wife, Christy for enduring the ups and downs associated with my endeavors and giving me the support I needed to eventually succeed.

VITA

~~January 1984~~ ~~Eastern Connecticut State University~~

May, 1983 B.S., Mathematics
Eastern Connecticut State University
Willimantic, Connecticut

December, 1988 M.S., Operations Research
Air Force Institute of Technology
Wright-Patterson Air Force Base, Ohio

1984-1987 Communications Officer - United States Air Force
Standard Systems Center
Montgomery, Alabama

1987-1991 Research Scientist - United States Air Force
Air Force Materiel Command
Wright-Patterson Air Force Base, Ohio

1994-Present Weapons and Tactics Analyst
Air Force Studies and Analyses Agency
Pentagon, Washington DC

1991-Present Graduate Student
Industrial, Welding, and Systems Engineering
The Ohio State University
Columbus, Ohio

FIELDS OF STUDY

Major Field: Industrial and Systems Engineering

Table of Contents

DEDICATION	ii
ACKNOWLEDGEMENTS	iii
VITA	iv
LIST OF TABLES	viii
LIST OF FIGURES	x
 CHAPTER	 PAGE
I INTRODUCTION	1
1.1 Dissertation Format	2
1.1.1 Overview of Chapter 2	3
1.1.2 Overview of Chapter 3	3
1.2 Contributions of the Research	4
 II MULTIVARIATE COMPOSITE DISTRIBUTIONS FOR CO- EFFICIENTS IN SYNTHETIC OPTIMIZATION PROBLEMS	 6
2.1 Introduction	6
2.2 Background	9
2.2.1 Implicit correlation induction	10
2.2.2 Explicit correlation induction	11
2.2.3 Explicit rank correlation induction	14
2.3 Explicit Correlation Induction for Multivariate Random Variables Using Composition	16
2.3.1 Extreme-correlation distributions for Y	16
2.3.2 Constructing composite distributions	21

2.4	Explicit Correlation Induction For Trivariate Random Variables	27
2.4.1	Type L composite distributions for trivariate random variables	27
2.4.2	Other composite distributions for trivariate random variables	32
2.4.3	Feasible correlation points for trivariate random variables	37
2.5	Extensions for General Multivariate Random Variables .	38
2.6	Composite Distributions for Quadravariate Random Variables	41
2.6.1	Adjusting Composition Weight Vectors	41
2.6.2	Choosing feasible quadravariate correlation points	44
2.7	Summary and Discussion	45
III THE EFFECTS OF COEFFICIENT CORRELATION STRUCTURE IN TWO-DIMENSIONAL KNAPSACK PROBLEMS ON SOLUTION PROCEDURE PERFORMANCE		52
3.1	Introduction	52
3.2	Background	54
3.2.1	The multidimensional knapsack problem	55
3.2.2	Empirical studies involving problems with correlated coefficients	55
3.2.3	Some empirical studies involving MKP	57
3.3	The Test Problem Generation Methods	59
3.3.1	Pearson product-moment-based correlation induction method	59
3.3.2	Spearman rank correlation-based correlation induction method	60
3.4	The Experiment Design and Analysis Methods	61
3.4.1	Definition of the experiment design settings . . .	62
3.4.2	Methods for analyzing results	67
3.5	Comparing Samples From The Correlation Induction Methods	69
3.6	Influence of Population Correlation Measure	75
3.7	Analysis of CPLEX performance	78
3.7.1	Correlation structure influence	79
3.7.2	Individual correlation term influence	82
3.7.3	Constraint slackness influences	84
3.7.4	The interaction between correlation structure and constraint slackness	85

3.7.5	Regression models for NODES	89
3.8	Analysis of heuristic performance	91
3.8.1	Correlation structure influence	91
3.8.2	Individual correlation term influence	92
3.8.3	Constraint slackness influence	96
3.8.4	The interaction between correlation structure and constraint slackness	97
3.8.5	Regression models for REL	102
3.9	Analysis of LP-IP Gap	104
3.10	Discussion and Conclusions	105
IV	CONCLUSIONS AND DISCUSSION	110
	BIBLIOGRAPHY	112

List of Tables

TABLE		PAGE
2.1	Extreme-correlation points when $k = 3$	18
2.2	Extreme-correlation points when $k = 4$	18
2.3	Values of δ_{ij}^ℓ when $k = 3$	19
2.4	Values of δ_{ij}^ℓ when $k = 4$	19
2.5	Upper Bound on λ_ℓ as k increases	40
3.1	Factors and Measures Used in Previous Empirical Studies of MKP Heuristics	58
3.2	Experiment Design Correlation Structures	64
3.3	Slackness Settings From Previous MKP studies	64
3.4	Sample Correlations by Target and Induction Type	73
3.5	Sample Correlations by Target, Method, and Alternate Measure	74
3.6	Performance Measure Averages by Correlation Measure	75
3.7	Results of Sign Test on Performance Measures	75
3.8	Sign Test Results On Each Correlation Structure	77
3.9	Sign Test Results For Type L Versus Type U Distributions	78
3.10	CPLEX Results using Pearson Correlation Induction . .	80
3.11	CPLEX Results using Spearman Correlation Induction .	81
3.12	Results of Kruskal-Wallis Tests on Each Correlation Term	83

3.13 Mean NODES by Constraint Slackness Setting	84
3.14 Sign Test Results for Performance Differences Between Mixed Constraint Slackness Levels	85
3.15 Interaction of Constraint Slackness and Correlation Type on Average NODES	86
3.16 Average NODES by Inter-Constraint Slackness and Cor- relation	87
3.17 Design Points Requiring Most and Least Average NODES for Pearson Correlation Problems	88
3.18 Design Points Requiring Most and Least Average NODES for Spearman Correlation Problems	88
3.19 Regression Model of CPLEX Results	90
3.20 TOYODA Results using Pearson Correlation Induction .	93
3.21 TOYODA Results using Spearman Correlation Induction	94
3.22 Results of Kruskal-Wallis Tests on Each Type of Corre- lation for REL	95
3.23 Mean REL by Constraint Slackness Setting	97
3.24 Sign Test Results for Performance Differences Between Mixed Constraint Slackness Levels	97
3.25 Design Points with Extreme REL Averages for Pearson Correlation Problems	101
3.26 Design Points with Extreme REL Averages for Spearman Correlation Problems	101
3.27 Regression Model of TOYODA Results	103
3.28 KW Test results for LP-IP Gap	105

List of Figures

FIGURE	PAGE
2.1 Example of P for $k = 3$	21
2.2 Example Type L composite pmf, $\rho = 0.6$	24
2.3 Example Type U composite pmf, $\rho = 0.6$	24
2.4 Type L composite pdf, exponential marginals, $\rho = 0.4$. .	25
2.5 Type U composite pdf, exponential marginals, $\rho = 0.4$. .	25
3.1 Three Dimensional Plot of Experiment Design Correlation Structures	65
3.2 Sample Distributional Form From Pearson Induction Method, $\rho = 0.49876$	71
3.3 Sample Distributional Form From Spearman Induction Method, $\rho = 0.49876$	72
3.4 Pearson Correlation Measure - REL $\times$ Correlation Setting	99
3.5 Inter-Constraint Correlation - REL $\times$ Correlation Setting	99
3.6 Spearman Correlation Measure - REL $\times$ Correlation Setting	100
3.7 Inter-Constraint Correlation - REL $\times$ Correlation Setting	100

CHAPTER I

INTRODUCTION

It is easy to envision a customer who requires a long service time at one service station in a tandem queueing system also requiring long service times at subsequent stations. Similar arguments apply to products requiring long processing times at one station in a serial production line. In thinking about optimization problems, one can imagine a direct, though imperfect, relationship between the cost of a product and the number of resource units that are needed to produce that product.

Users of simulation techniques need the ability to generate values of multivariate random variables with specified dependencies, or population correlation structure, to accurately simulate real phenomena. This is true for simulations of serial manufacturing systems, as well as empirical evaluations of optimization solution methods.

Synthetic optimization problems are (most often) randomly generated optimization problems that provide test cases for empirical evaluations of optimization solution methods. These studies are of greatest value when the synthetic problems have characteristics similar to those of problems encountered in practice or a variety of characteristics so that the range of a solution procedure's performance may

be determined. In many studies of optimization methods, researchers assume that all coefficients are mutually independent, but as with other practical simulations, this independence assumption may not reflect real-life dependencies.

The results of some empirical studies on the performance of algorithms and heuristics indicate that the correlation between objective function and constraint coefficients influences the performance of solution methods. Multivariate sampling would facilitate the generation of values of multivariate random variables with realistic or prescribed population correlation structures to represent the coefficients in the objective function and more than one constraint. Such sampling would lead to a deeper understanding of solution procedure performance when there are dependencies among the problem coefficients.

The goals of this research are: (1) to develop a methodology for generating values of multivariate random variables with specified marginal distributions and a specified population correlation structure, (2) to demonstrate the use of this methodology in generating synthetic optimization problems, and (3) to conduct an empirical study to assess the influence of the population correlation structure on the performance of optimization solution methods.

1.1 Dissertation Format

This dissertation contains two self-contained papers. The first paper, provided here as Chapter 2, presents a methodology for constructing composite distributions for multivariate random variables with specified marginal distributions and a specified correlation structure. The second paper, presented here as Chapter 3, presents the

results of an empirical study of the influence of population correlation structure of the coefficients in synthetic two-dimensional knapsack problems on solution procedure performance.

1.1.1 Overview of Chapter 2

Extreme-correlation distributions are joint distributions in which all pairwise population correlations have either their most positive or their most negative possible values. These distributions are the building blocks for a class of multivariate composite distributions. Composite distributions constructed from the extreme-correlation distributions and the joint distribution under independence form an even richer class of distributions. Both classes of composite distributions apply to multivariate discrete and continuous random variables. They facilitate more realistic simulations of practical systems, such as manufacturing and other tandem queueing systems, as well as more comprehensive computational experiments on optimization methods.

1.1.2 Overview of Chapter 3

Chapter 3 presents an empirical study of the effects on solution methods of the population correlation structure among the coefficient types in the two-dimensional knapsack problem (2KP). The composite distributions presented in Chapter 2 provide the requisite foundation for multivariate sampling for generating values of coefficients with specified Pearson product-moment correlation structures for synthetic 2KPs. Additional instances of 2KPs with specified Spearman rank cor-

relation structures are also generated using a known sampling technique. A total of 2240 2KP instances are generated based on two correlation measures, various population correlation structures (matrices), and four different constraint slackness settings. All of the problems are solved with a commercially available branch-and-bound code and a well-known heuristic.

1.2 Contributions of the Research

This research makes two principal contributions. The first is the characterization of composite distributions for multivariate random variables. Straightforward procedures for generating values of multivariate random variables with a specified population correlation structure make these characterizations an easy way to induce realistic dependencies in simulated data. The second contribution is an increased understanding of the effect on solution procedure performance of the correlation structure among the coefficient types in 2KP instances.

Chapter 2 contains theory providing a foundation for characterizing multivariate composite distributions. More specifically, Chapter 2 presents

- characterizations of extreme-correlation distributions for both discrete and continuous multivariate random variables,
- methods for characterizing multivariate distributions, based on a specified Pearson product-moment correlation structure, as a composition of extreme-correlation distributions and the joint distribution under independence,
- closed-form methods for characterizing composite distributions for trivariate random variables, and a simple procedure for finding composition probabilities for quadrivariate random variables, and
- methods for selecting feasible correlation structures for both trivariate and quadrivariate random variables.

Chapter 3 demonstrates the practicality of using composite distributions to induce correlation explicitly among three types of coefficients in 2KP. More specifically, Chapter 3 presents

- an experiment design for investigating solution procedure performance that takes advantage of multivariate explicit correlation induction, and treats the population correlation structure and the correlation measure as factors in the experiment, and
- insights regarding the synergistic effect between population correlation structure and constraint slackness on both the characteristics of the synthetic 2KP and the ability of solution procedures to solve the problem.

CHAPTER II

MULTIVARIATE COMPOSITE DISTRIBUTIONS FOR COEFFICIENTS IN SYNTHETIC OPTIMIZATION PROBLEMS

2.1 Introduction

This chapter presents a characterization of composite distributions for multivariate random variables with specified marginal distributions and a specified Pearson product-moment population correlation structure. A composite distribution for a multivariate random variable $\mathbf{Y} = (Y_1, Y_2, \dots, Y_k)$ is a distribution that may be represented as a convex combination of other valid distributions for $\mathbf{Y}$. The Pearson product-moment correlation between random variables Y_i and Y_j , where $i \neq j$,

$$\text{Corr}(Y_i, Y_j) = \frac{E(Y_i Y_j) - E(Y_i)E(Y_j)}{(\text{Var}(Y_i)\text{Var}(Y_j))^{\frac{1}{2}}}, \quad (2.1)$$

is a measure of the strength of the linear relationship between Y_i and Y_j .

The principal motivation for this research is the generation of synthetic optimization problems, which is too infrequently viewed as an application of multivariate sampling. (However, this research is applicable to many other simulation

applications, e.g., simulations of manufacturing systems.) Synthetic optimization problems are used to test and compare algorithms and heuristics because of limited supplies of real-life test problems. A common practice is generating the coefficients for these problems under mutual independence. This approach alone is inadequate when there may be dependencies among the coefficients in real-life problems. An alternative approach is to induce several structured dependencies between some types of coefficients to provide a greater variety of test problems, from easy ones to difficult ones, and a higher degree of realism in the test problems. The composite distributions described in this chapter facilitate the generation of synthetic optimization problems with a dependence structure represented by a Pearson product-moment population correlation matrix.

Some researchers have induced correlation between objective function and constraint coefficients in synthetic optimization problems and found that the level of correlation is related to the performance of solution methods (Martello and Toth, 1979, 1981, 1988; Balas and Zemel, 1980; Balas and Martin, 1980; Potts and Van Wassenhove, 1988; Guignard and Rosenwein, 1989; John, 1989; Reilly, 1991; Rushmeier and Nemhauser, 1993; Moore and Reilly, 1993; Amini and Racer, 1994; Cario *et al.*, 1995). The correlations between the coefficients in different constraints may also be related to solution method performance, but this possibility has only been systematically investigated by Hill (Chapter 3). He uses the characterizations of multivariate distributions presented in this paper to investigate the effects of interconstraint correlation, as well as those of the correlations between the objective function and constraint coefficients, on the performance of standard solution methods on two-dimensional knapsack problems. He observes

that the interconstraint correlation term has at least as significant a relationship to solution method performance as the correlations between the objective function coefficients and the coefficients in either of the constraints.

There are usually an infinite number of ways to characterize a joint distribution for $\mathbf{Y}$ when the marginal distributions for Y_i , $i = 1, 2, \dots, k$, and a population correlation structure are specified. Composite distributions, and in particular those whose constituent joint distributions have a simple form, are easy to sample from. Consequently, attention here is restricted to multivariate composite distributions that are composed of the joint distribution under independence and the extreme-correlation distributions, the 2^{k-1} distributions of $\mathbf{Y}$ for which $\text{Corr}(Y_i, Y_j)$ has either its most positive or most negative possible value for all $i < j \leq k$.

This paper is organized as follows. In §2.2, implicit and explicit correlation induction methods for generating coefficients for synthetic optimization problems are discussed. Basic concepts for composite distributions for multivariate random variables are presented in §2.3. Extreme-correlation distributions for multivariate random variables are characterized and used in conjunction with the joint distribution under independence to construct multivariate composite distributions. Composite distributions for trivariate random variables are constructed using closed-form formulas for the composition probabilities in §2.4. The limitations of extending the composition probability formulas for trivariate random variables to multivariate

random variables are discussed in §2.5. A composition weight adjustment technique for constructing composite distributions for quadrivariate random variables is presented in §2.6. The contributions of this research and areas of further investigation are summarized in §2.7.

Some of the early results of this research appear in Hill and Reilly (1994).

2.2 Background

How the correlation structure of multivariate random variables is modeled and varied can affect the results and conclusions in computational experiments and other simulation applications. For certain classes of discrete optimization problems, randomly generated instances with high, positive correlation between objective function and constraint coefficients are relatively hard to solve with enumerative procedures. Such results are reported by Martello and Toth (1979, 1988), Balas and Zemel (1980), and Reilly (1991) for knapsack problems; Balas and Martin (1980) for capital budgeting problems; Rushmeier and Nemhauser (1993) and Moore and Reilly (1993) for set covering problems; and Potts and Van Wassenhove (1988, 1992) and John (1989) for scheduling problems. Instances of the generalized assignment problem with high, negative correlation between the objective function and capacity constraint coefficients are relatively hard to solve with enumerative procedures (Martello and Toth, 1981; Fisher, Jaikumar, and Van Wassenhove, 1986; Guignard and Rosenwein, 1989; Trick, 1992; Mazzola and Neebe, 1993; Amini and Racer, 1994; Cario *et al.*, 1995).

In the rest of this section, correlation-induction methods that have been used to generate coefficients for synthetic optimization problems are discussed

2.2.1 Implicit correlation induction

Moore and Reilly (1993) discuss three ways to generate bivariate random variables as coefficients for synthetic optimization problems: mutual independence, implicit correlation induction, and explicit correlation induction. They classify as implicit correlation induction any generation method for which the parameters of the method imply the population correlation between two random variables.

The implicit correlation induction method in Martello and Toth (1979), which has been widely mimicked by others, induces dependence between two random variables Y_1 and Y_2 , by generating a value for Y_1 and then a value of $Y_2 = Y_1 + W$, where W is an independently generated noise term. For instance, they let $Y_1 \sim U\{1, 2, \dots, 100\}$ (i.e., Y_1 has a discrete uniform distribution over the integers from 1 to 100) and $W \sim U\{-10, -9, \dots, 10\}$ when generating objective function (Y_2) and constraint (Y_1) coefficients for knapsack problems. The implied value of $\rho = \text{Corr}(Y_1, Y_2)$ is above 0.97 in this case. Martello and Toth (1981) let $Y_2 = 111 - Y_1 + W$ for generalized assignment problems, and the implied value of ρ is below -0.97. Martello and Toth, as well as other authors, call such population correlation levels “weak.” But, it is not clear whether any of these authors knows the magnitudes of the values of ρ implied by the parameters used in their problem generation methods.

One may change the support of Y_1 or of W , or multiply either of these random variables by a constant, and systematically vary the implied population correlation. Such an approach was used by Cario *et al.* (1995).

Other implementations of implicit correlation induction include Balas and Zemel (1980), Balas and Martin (1980), Potts and Van Wassenhove (1988), Guignard and Rosenwein (1989), John (1989), Rushmeier and Nemhauser (1993), and Amini and Racer (1994). Balas and Martin (1980) generate capital budgeting problems with implicitly induced interconstraint correlations, as well as objective function-constraint correlations.

2.2.2 Explicit correlation induction

According to Moore and Reilly (1993), an explicit correlation induction method is one where the user specifies the population correlation structure.

Fréchet (1951) characterized bounds on joint probability distributions for (Y_1, Y_2) as

$$H^-(y_1, y_2) = \max \{F_1(y_1) + F_2(y_2) - 1, 0\}, \quad (2.2)$$

and

$$H^+(y_1, y_2) = \min \{F_1(y_1), F_2(y_2)\}, \quad (2.3)$$

where $F_1(y_1)$ is the cumulative distribution function (cdf) for Y_1 , and $F_2(y_2)$ is the cdf for Y_2 . $H^-(y_1, y_2)$ and $H^+(y_1, y_2)$ are, respectively, the minimum- and maximum-correlation joint cdfs for (Y_1, Y_2) . Fréchet shows that

$$H^-(y_1, y_2) \leq H(y_1, y_2) \leq H^+(y_1, y_2) \quad (2.4)$$

for all (y_1, y_2) and all possible joint distributions $H(y_1, y_2)$.

Fréchet uses the bounding distributions (2.2) and (2.3), along with the joint distribution under independence, $F_1(y_1)F_2(y_2)$, to characterize two classes of composite distributions for (Y_1, Y_2) :

$$\lambda H^-(y_1, y_2) + (1 - \lambda)H^+(y_1, y_2), \quad 0 \leq \lambda \leq 1, \quad (2.5)$$

and

$$(1 - a - b)F_1(y_1)F_2(y_2) + aH^-(y_1, y_2) + bH^+(y_1, y_2), \quad a, b \geq 0, a + b \leq 1. \quad (2.6)$$

The weights λ and $(1 - \lambda)$ in (2.5) and a, b , and $(1 - a - b)$ in (2.6) are referred to as composition probabilities. Nelsen (1987) also describes composite distributions (2.6).

A class of joint distributions for (Y_1, Y_2) is comprehensive if the class includes the boundary distributions (2.2) and (2.3) and $F_1(y_1)F_2(y_2)$ (Devroye, 1986). The class of composite distributions (2.6) is comprehensive, while the class (2.5) is not.

Extreme mixtures

The composite distributions (2.5) are sometimes referred to as extreme mixtures because they are composed of just the extreme-correlation distributions for (Y_1, Y_2) , $H^-(y_1, y_2)$ and $H^+(y_1, y_2)$. Suppose ρ is specified and $\lambda = (\rho^+ - \rho)/(\rho^+ - \rho^-)$, where ρ^+ and ρ^- are, respectively, the maximum and minimum possible values of ρ . Then there is a unique extreme mixture for (Y_1, Y_2) for each value of ρ such that $\rho^- \leq \rho \leq \rho^+$. Extreme mixtures are easy to use but cannot generate observations of (Y_1, Y_2) with Y_1 and Y_2 independent because extreme mixtures do not form a comprehensive class of distributions for (Y_1, Y_2) . Extreme mixtures apply to both discrete and continuous random variables.

Conventional mixtures

The composite distributions (2.6) are sometimes referred to as conventional mixtures when either $a = 0$ or $b = 0$. Suppose ρ is specified. Then, $a = 0$ and $b = \rho/\rho^+$ if $\rho \geq 0$, and $a = \rho/\rho^-$ and $b = 0$ if $\rho \leq 0$. Conventional mixtures form a comprehensive class of distributions for (Y_1, Y_2) , are easy to use, and provide a unique distribution for all values of ρ such that $\rho^- \leq \rho \leq \rho^+$. However, a conventional mixture cannot generate values of (Y_1, Y_2) with Y_1 and Y_2 uncorrelated but dependent. Conventional mixtures apply to both discrete and continuous random variables. Conventional mixtures are also discussed by Schmeiser and Lal (1982) and Nelsen (1987).

Moore and Reilly (1993) generate set covering problem coefficients based on conventional mixtures.

Parametric mixtures

For finite discrete random variables Y_1 and Y_2 , Peterson and Reilly (1993) describe a special case of the distributions (2.6) which they refer to as parametric mixtures.

Define θ to be the minimum joint probability for any value (y_1, y_2) in the support of (Y_1, Y_2) . Let $f_1(y_1)$ and $f_2(y_2)$ be the pmfs for Y_1 and Y_2 , respectively. Also let $i^* = \arg \min_i \{f_1(y_{1i})\}$; $j^* = \arg \min_j \{f_2(y_{2j})\}$; and $\theta^+ = f_1(y_{1i^*})f_2(y_{2j^*})$. Suppose that (ρ, θ) is a point such that

$$0 \leq \theta \leq \theta^+ \quad (2.7)$$

and

$$(1 - \theta/\theta^+)\rho^- \leq \rho \leq (1 - \theta/\theta^+)\rho^+. \quad (2.8)$$

Suppose also that $h^+(y_{1i^*}, y_{2j^*}) = h^-(y_{1i^*}, y_{2j^*}) = 0$, where $h^+(y_1, y_2)$ and $h^-(y_1, y_2)$ are the maximum- and minimum-correlation probability mass functions (pmfs) for (Y_1, Y_2) , respectively. Peterson and Reilly show that $\text{Corr}(Y_1, Y_2) = \rho$ and the minimum joint probability is θ for the following composite distribution:

$$(1 - a - b)f_1(y_1)f_2(y_2) + ah^-(y_1, y_2) + bh^+(y_1, y_2), \quad (2.9)$$

where

$$a = ((1 - \theta/\theta^+)\rho^+ - \rho) / (\rho^+ - \rho^-), \quad (2.10)$$

and

$$b = (\rho - (1 - \theta/\theta^+)\rho^-) / (\rho^+ - \rho^-). \quad (2.11)$$

There are an infinite number of parametric mixtures (2.9) for each value of ρ such that $\rho^- < \rho < \rho^+$, but a unique parametric mixture for each point (ρ, θ) that makes either inequality in (2.8) active. For bivariate discrete random variables, extreme and conventional mixtures are special cases of parametric mixtures.

Reilly (1991) generates knapsack problems based on parametric mixtures. Yang (1994) generates knapsack problems and Cario *et al.* (1995) generate generalized assignment problems based on parametric mixtures, including extreme mixtures (2.5) and conventional mixtures.

2.2.3 Explicit rank correlation induction

Iman and Conover (1982) describe a method for generating n samples of a k -variate random variable $\mathbf{Y}$ with specified marginal distributions and a specified Spearman

rank population correlation structure. Their method shuffles n independently generated components of a multivariate random variable across k vectors so that the sample Spearman correlation structure approximates a specified Spearman rank population correlation structure, $\mathbf{M}$. First generate two matrices $\mathbf{R}$ and $\mathbf{V}$ such that $\mathbf{R}$ is an $(n \times k)$ matrix of van der Waerden scores, randomized within each of the k columns, and $\mathbf{V}$ is an $(n \times k)$ matrix of n independent observations of each of the k random variables. Consider each column of $\mathbf{R}$ as n observations of k random variables and compute $\mathbf{T}$, the corresponding sample rank correlation matrix. Compute the Choleski factorizations $\mathbf{A}$ and $\mathbf{Q}$ such that $\mathbf{T} = \mathbf{A}\mathbf{A}'$ and $\mathbf{M} = \mathbf{Q}\mathbf{Q}'$. Compute

$$\mathbf{S} = \mathbf{R}(\mathbf{A}\mathbf{Q}^{-1})', \quad (2.12)$$

which is a transformed matrix of scores. The k columns of n values in $\mathbf{S}$ have a sample rank correlation structure that approximates $\mathbf{M}$. The entries in each column of $\mathbf{V}$ are reordered so that their rankings are the same as the rankings in the corresponding columns of $\mathbf{S}$. The sample Spearman rank correlation structure of the shuffled matrix of observations, $\mathbf{V}$, approximates the specified correlation structure, $\mathbf{M}$.

Hill (Chapter 3) compares the performance of an algorithm and a heuristic on two-dimensional knapsack problems generated based on composite distributions with specified Pearson product-moment population correlation structures and with Iman and Conover's method for the same Spearman rank population correlation structures.

2.3 Explicit Correlation Induction for Multivariate Random Variables Using Composition

In this section, extreme-correlation distributions for a multivariate random variable $\mathbf{Y}$ are characterized and used to construct composite distributions with a specified correlation structure. A procedure for generating samples from composite distributions of multivariate random variables is presented.

2.3.1 Extreme-correlation distributions for $\mathbf{Y}$

Assume there are k random variables, $Y_i, i = 1, 2, \dots, k$. Each Y_i has support S_i , is distributed according to $f_i(y_i)$, and has cdf $F_i(y_i)$. Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_k)$ and $\mathbf{y} = (y_1, y_2, \dots, y_k)$ be a value of $\mathbf{Y}$. Let $S = S_1 \times S_2 \times \dots \times S_k$ be the support of $\mathbf{Y}$. Each feasible joint distribution for $\mathbf{Y}$, $h(\mathbf{y})$, has a Pearson product-moment correlation structure whose correlation terms comprise a $\binom{k}{2}$ -vector, $\boldsymbol{\rho} = (\rho_{12}, \rho_{13}, \dots, \rho_{k,k-1})$, where $\rho_{ij} = \text{Corr}(Y_i, Y_j)$ for all $i < j \leq k$.

Let Ψ_{ij} be the set of all feasible bivariate distributions $h_{ij}(y_i, y_j)$ for (Y_i, Y_j) , for all $i < j \leq k$. Let $K_{ij}^+ = \max_{h_{ij} \in \Psi_{ij}} \{E(Y_i Y_j)\}$ and $K_{ij}^- = \min_{h_{ij} \in \Psi_{ij}} \{E(Y_i Y_j)\}$ for all $i < j \leq k$. The maximum and minimum values of each ρ_{ij} are

$$\rho_{ij}^+ = (K_{ij}^+ - E(Y_i)E(Y_j)) / (\text{Var}(Y_i)\text{Var}(Y_j))^{\frac{1}{2}} \quad (2.13)$$

and

$$\rho_{ij}^- = (K_{ij}^- - E(Y_i)E(Y_j)) / (\text{Var}(Y_i)\text{Var}(Y_j))^{\frac{1}{2}}, \quad (2.14)$$

respectively. Peterson (1990) presents a factored transportation problem to find K_{ij}^+ or K_{ij}^- for finite discrete random variables Y_i and Y_j . He finds K_{ij}^+ with the

Northwest Corner Rule (NWCR) and uses the Southwest Corner Rule (SWCR) to find K_{ij}^- . For a ge

$i \quad j$

$$K_{ij}^+ = \int_0^1 F_i^{-1}(u) F_j^{-1}(u) du \quad (2.15)$$

and

$$K_{ij}^- = \int_0^1 F_i^{-1}(u) F_j^{-1}(1-u) du. \quad (2.16)$$

Define a *correlation point*, $\boldsymbol{\rho} = (\rho_{12}, \rho_{13}, \dots, \rho_{k-1,k})$, as the $\binom{k}{2}$ -vector of correlation values associated with some $h(\mathbf{y})$. An *extreme-correlation point* for $\mathbf{Y}$ is a correlation point where either $\rho_{ij} = \rho_{ij}^+$ or $\rho_{ij} = \rho_{ij}^-$ for all $i < j \leq k$. Each extreme-correlation point is associated with a feasible assignment of ρ_{1j}^+ or ρ_{1j}^- to each ρ_{1j} , $j = 2, 3, \dots, k$. So there are 2^{k-1} extreme-correlation points for $\mathbf{Y}$. Denote the extreme-correlation points as $\mathbf{q}_\ell$, $\ell = 1, 2, \dots, 2^{k-1}$.

For each extreme-correlation point, define the $\binom{k}{2}$ -vector $\boldsymbol{\delta}^\ell = (\delta_{12}^\ell, \dots, \delta_{k-1,k}^\ell)$, where

$$\delta_{ij}^\ell = \begin{cases} 1 & \text{if } \rho_{ij} = \rho_{ij}^+; \\ 0 & \text{if } \rho_{ij} = \rho_{ij}^-. \end{cases} \quad (2.17)$$

To determine the components of each vector $\boldsymbol{\delta}^\ell$, the appropriate values are assigned to δ_{1j}^ℓ , $j = 2, 3, \dots, k$, and then the formula

$$\delta_{ij}^\ell = 1 - |\delta_{1i}^\ell - \delta_{1j}^\ell| \quad (2.18)$$

is used to find the remaining δ_{ij}^ℓ values. Tables 2.1 and 2.2 characterize the extreme-correlation points for $k = 3$ and $k = 4$, respectively. Tables 2.3 and 2.4 provide the corresponding values for δ_{ij}^ℓ .

neral random variable (Y, Y) ,

Table 2.1: Extreme-correlation points when $k = 3$

ℓ	ρ_{ij}		
	ρ_{12}	ρ_{13}	ρ_{23}
1	ρ_{12}^+	ρ_{13}^-	ρ_{23}^-
2	ρ_{12}^-	ρ_{13}^+	ρ_{23}^-
3	ρ_{12}^-	ρ_{13}^-	ρ_{23}^+
4	ρ_{12}^+	ρ_{13}^+	ρ_{23}^+

Table 2.2: Extreme-correlation points when $k = 4$

ℓ	ρ_{ij}					
	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}
1	ρ_{12}^-	ρ_{13}^-	ρ_{14}^-	ρ_{23}^+	ρ_{24}^+	ρ_{34}^+
2	ρ_{12}^-	ρ_{13}^-	ρ_{14}^+	ρ_{23}^+	ρ_{24}^-	ρ_{34}^-
3	ρ_{12}^-	ρ_{13}^+	ρ_{14}^-	ρ_{23}^-	ρ_{24}^+	ρ_{34}^-
4	ρ_{12}^-	ρ_{13}^+	ρ_{14}^+	ρ_{23}^-	ρ_{24}^-	ρ_{34}^+
5	ρ_{12}^+	ρ_{13}^-	ρ_{14}^-	ρ_{23}^-	ρ_{24}^-	ρ_{34}^+
6	ρ_{12}^+	ρ_{13}^-	ρ_{14}^+	ρ_{23}^-	ρ_{24}^+	ρ_{34}^-
7	ρ_{12}^+	ρ_{13}^+	ρ_{14}^-	ρ_{23}^+	ρ_{24}^-	ρ_{34}^-
8	ρ_{12}^+	ρ_{13}^+	ρ_{14}^+	ρ_{23}^+	ρ_{24}^+	ρ_{34}^+

Table 2.3: Values of δ_{ij}^ℓ when $k = 3$.

ℓ	$1, j$		i, j
	1,2	1,3	2,3
1	1	0	0
2	0	1	0
3	0	0	1
4	1	1	1

Table 2.4: Values of δ_{ij}^ℓ when $k = 4$.

ℓ	$1, j$			i, j		
	1,2	1,3	1,4	2,3	2,4	3,4
1	0	0	0	1	1	1
2	0	0	1	1	0	0
3	0	1	0	0	1	0
4	0	1	1	0	0	1
5	1	0	0	0	0	1
6	1	0	1	0	1	0
7	1	1	0	1	0	0
8	1	1	1	1	1	1

Define the *zero-correlation point* as the zero vector of dimension $\binom{k}{2}$ and denote it $\mathbf{q}_0$. Also define P to be the convex hull of the correlation points $\mathbf{q}_\ell, \ell = 0, 1, \dots, 2^{k-1}$. If $\boldsymbol{\rho} \in P$, then there exist values $\lambda_\ell \geq 0, \ell = 0, 1, \dots, 2^{k-1}$, such that

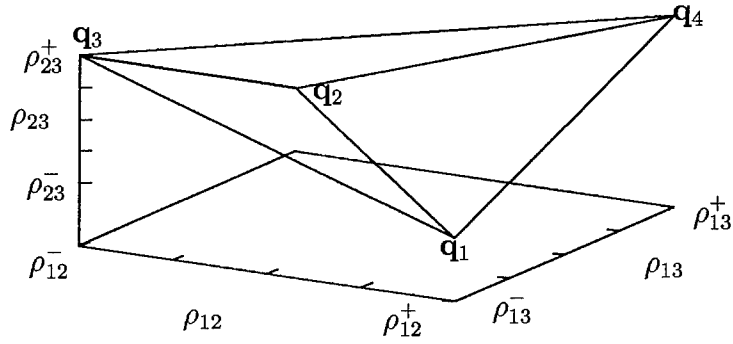
$$\boldsymbol{\rho} = \sum_{\ell=0}^{2^{k-1}} \lambda_\ell \mathbf{q}_\ell, \quad (2.19)$$

and

$$\sum_{\ell=0}^{2^{k-1}} \lambda_\ell = 1. \quad (2.20)$$

Figure 2.1 depicts an example set P where $k = 3$. According to Rousseeuw and Molenberghs (1994), all feasible correlation structures, $(\rho_{12}, \rho_{13}, \rho_{23})$, for such a trivariate random variable are contained in an elliptical tetrahedron. P is a proper subset of this elliptical tetrahedron; the extreme points of P in Figure 2.1 are the extreme-correlation points $\mathbf{q}_\ell, \ell = 1, 2, 3, 4$, characterized in Table 2.1 and are the extreme points of Rousseeuw and Molenberghs' elliptical tetrahedron.

An *extreme-correlation distribution* for $\mathbf{Y}$ is a joint distribution for which either $\rho_{ij} = \rho_{ij}^+$ or $\rho_{ij} = \rho_{ij}^-$, for all $i < j \leq k$. Denote the 2^{k-1} extreme-correlation distributions as $h_\ell(\mathbf{y}), \ell = 1, 2, \dots, 2^{k-1}$. There is a one-to-one correspondence between the extreme-correlation points in P and the extreme-correlation distributions of $\mathbf{Y}$.

Figure 2.1: Example of P for $k = 3$

2.3.2 Constructing composite distributions

Let

$$h_0(\mathbf{y}) = \prod_{i=1}^k f_i(y_i) \quad (2.21)$$

be the joint distribution of $\mathbf{Y}$ under independence. A comprehensive class of composite distributions for $\mathbf{Y}$ is given by:

$$\Omega = \left\{ h(\mathbf{y}) \left| h(\mathbf{y}) = \sum_{\ell=0}^{2^k-1} \lambda_{\ell} h_{\ell}(\mathbf{y}), \sum_{\ell=0}^{2^k-1} \lambda_{\ell} = 1, \lambda_{\ell} \geq 0, \forall \ell \right. \right\}. \quad (2.22)$$

The class Ω generalizes the comprehensive class (2.6) of composite distributions for bivariate random variables introduced by Fréchet (1951) in the sense that Ω includes each of $h_{\ell}(\mathbf{y})$, $\ell = 0, 1, \dots, 2^k-1$. It is clear from the definitions of Ω

and P that $\boldsymbol{\rho} \in P$ if and only if there is some joint distribution $h(\mathbf{y}) \in \Omega$ whose correlation structure is given by $\boldsymbol{\rho}$. P contains correlation points that correspond to correlation structures that are expressible as convex combinations of $\mathbf{q}_\ell$, $\ell = 0, 1, \dots, 2^{k-1}$, but does not necessarily contain all feasible correlation structures for $\mathbf{Y}$, as demonstrated in Rousseeuw and Molenberghs (1994).

Let $\boldsymbol{\rho} \in P$ represent the desired population correlation structure. Constructing a composite distribution $h(\mathbf{y}) \in \Omega$ associated with $\boldsymbol{\rho} \in P$ requires a composition probability vector, $\boldsymbol{\lambda}$, that satisfies the following conditions:

$$\sum_{\ell=1}^{2^{k-1}} \lambda_\ell [\delta_{ij}^\ell \rho_{ij}^+ + (1 - \delta_{ij}^\ell) \rho_{ij}^-] = \rho_{ij} \quad \forall i < j \leq k, \quad (2.23)$$

$$\sum_{\ell=0}^{2^{k-1}} \lambda_\ell = 1, \quad (2.24)$$

$$\lambda_\ell \geq 0 \quad \ell = 0, 1, \dots, 2^{k-1}. \quad (2.25)$$

A composite distribution $h(\mathbf{y}) \in \Omega$ with a minimum value of λ_0 is referred to here as a Type L distribution. In many cases, $\lambda_0 = 0$ for Type L distributions, meaning $h_0(\mathbf{y})$ is not included in the composition. It is easily shown for bivariate random variables that extreme mixtures are Type L composite distributions (see the Appendix to this chapter for details).

A composite distribution $h(\mathbf{y}) \in \Omega$ with a maximum value of λ_0 is referred to here as a Type U distribution. For bivariate random variables, conventional mixtures are Type U distributions (see the Appendix to this chapter for details).

For any $\boldsymbol{\rho} \in P$, the corresponding Type L and Type U composite distributions define a range of composite distributions in Ω with the correlation structure $\boldsymbol{\rho}$. While the correlation structure for the distributions within this range of composite

distributions does not change as λ_0 changes, the distributions themselves can be strikingly different. Therefore, λ_0 may be considered an index for the composite distributions in Ω that are associated with each $\rho \in P$.

Two examples for bivariate random variables are used to illustrate the range of possible composite distributions defined by the limiting Type L and Type U distributions. The first example illustrates how the value of λ_0 affects the nature of the pmf for $\mathbf{Y}$ for discrete random variables. The second example illustrates a similar point when $\mathbf{Y}$ is continuous.

Example 1 (Hill and Reilly, 1994). Let $Y_1 \sim U\{1, 2, 3, 4, 5\}$ and let Y_2 be a binomial random variable with 3 independent trials and success probability 0.4. Suppose that the desired population correlation value is $\rho = 0.6$. The pmf shown in Figure 2.2 is the Type L composite distribution with $\lambda_0 = 0$, $\lambda_1 = 0.1715$ and $\lambda_2 = 0.8285$. The pmf shown in Figure 2.3 is the Type U composite distribution with $\lambda_0 = 0.3431$, $\lambda_1 = 0$, $\lambda_2 = 0.6569$. Note in Figure 2.2 that the joint probability for 7 of 20 members of $S_1 \times S_2$ is zero, while each member of $S_1 \times S_2$ has positive probability with the pmf in Figure 2.3. $\square$

Example 2 (Hill and Reilly, 1994). Let Y_1 and Y_2 be exponential random variables with unit mean and $\rho = 0.4$. From Page (1965) it is known that $\rho^+ = 1.0$ and $\rho^- = 1 - \pi^2/6$. Figure 2.4 shows 1000 observations based on the Type L composite distribution with $\lambda_0 = 0$, $\lambda_1 = 0.36$, and $\lambda_2 = 0.64$. Figure 2.5 shows 1000 observations based on the Type U composite distribution with $\lambda_0 = 0.6$, $\lambda_1 = 0$, and $\lambda_2 = 0.4$. Including the independent pdf in the composition, as in the Type U distribution, leads to a greater variety of possible realizations of $\mathbf{Y}$. $\square$

Y_1	Y_2			
	0	1	2	3
1	0.1657	0	0.0233	0.0110
2	0.0132	0.1607	0.0261	0
3	0	0.2000	0	0
4	0.0028	0.0713	0.1259	0
5	0.0343	0	0.1127	0.0530

Figure 2.2: Example Type L composite pmf, $\rho = 0.6$

Y_1	Y_2			
	0	1	2	3
1	0.1462	0.0296	0.0198	0.0044
2	0.0254	0.1504	0.0198	0.0044
3	0.0148	0.1610	0.0198	0.0044
4	0.0148	0.0614	0.1194	0.0044
5	0.0148	0.0296	0.1092	0.0464

Figure 2.3: Example Type U composite pmf, $\rho = 0.6$

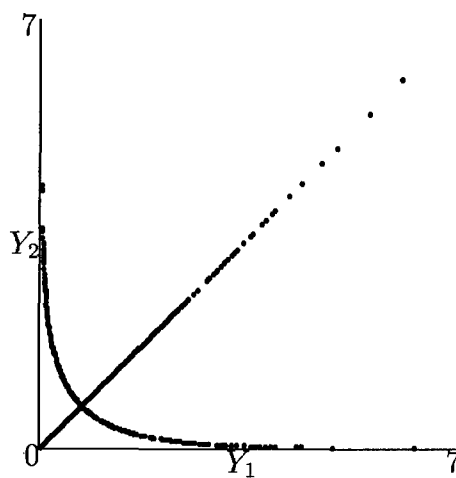


Figure 2.4: Type L composite pdf, exponential marginals, $\rho = 0.4$

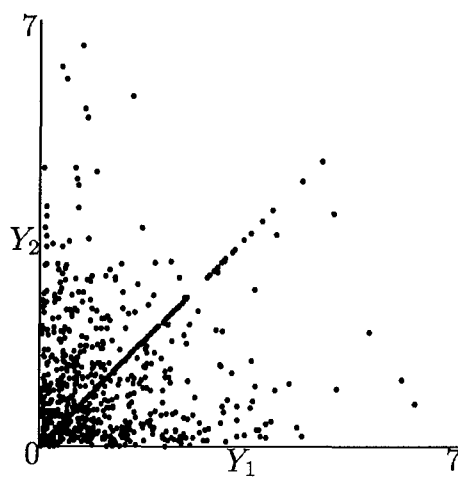


Figure 2.5: Type U composite pdf, exponential marginals, $\rho = 0.4$

Given a valid composition probability vector, the following procedure generates values of $\mathbf{Y}$ based on a composite distribution in Ω with the correlation structure associated with the correlation point $\boldsymbol{\rho}$.

Procedure RVAR1

1. Generate $u_1, u_2, \dots, u_{k+1} \sim U(0, 1)$.
2. If $u_{k+1} \leq \lambda_0$, then for $i = 1, 2, \dots, k$,
 $y_i = F_i^{-1}(u_i)$ and go to Step 6.
 Otherwise, set $m = 1, \Gamma = \lambda_0 + \lambda_1$.
3. If $u_{k+1} > \Gamma$, go to Step 4. Otherwise, go to Step 5.
4. $m \leftarrow m + 1, \Gamma \leftarrow \Gamma + \lambda_m$. Go to Step 3.
5. Generate $\mathbf{y}$ with u_1 based on $g_m(\mathbf{y})$.
 - (a) $y_1 \leftarrow F_1^{-1}(u_1)$.
 - (b) $y_i = F_i^{-1}(1 + u_1(2\delta_{ij}^m - 1) - \delta_{ij}^m), i = 2, 3, \dots, 2^{k-1}$.
6. Return $\mathbf{y}$.

RVAR1 uses $k+1$ random numbers per observation of $\mathbf{Y}$. One random number, u_{k+1} , is used to select a constituent distribution of the composite distribution. Another random number, u_1 , is used to generate a value of Y_1 , and the δ_{ij}^ℓ values determine whether u_1 or $1 - u_1$ is used for sampling values of $Y_2, Y_3, \dots, Y_k$ for any extreme-correlation distribution. The remaining $k - 1$ random numbers are used only for independent sampling, i.e., if $u_{k+1} \leq \lambda_0$. RVAR1 is designed to facilitate synchronized sampling, which is useful in many computational experiments. An alternative to RVAR1 is a more efficient procedure that generates values of $\mathbf{Y}$ using an expected number of $2 + (k - 1)\lambda_0$ random numbers per observation, rather than the constant $k + 1$ random numbers.

2.4 Explicit Correlation Induction For Trivariate Random Variables

This section introduces closed-form formulas for the unique composition probabilities for Type L distributions for trivariate random variables. These formulas are extended to other composite distributions, including Type U composite distributions. Throughout this section, it is assumed that $k = 3$, $\mathbf{Y} = (Y_1, Y_2, Y_3)$, $\mathbf{y} = (y_1, y_2, y_3)$, and $f_i(y_i) > 0$ for all y_i , $i = 1, 2, 3$.

2.4.1 Type L composite distributions for trivariate random variables

Define $\bar{\rho}_{ij} = (\rho_{ij}^+ + \rho_{ij}^-)/2$ for $i = 1, 2$, $j = i + 1, 3$. Let $\boldsymbol{\rho} \in P$, $\lambda_0 = 0$, and

$$\lambda_\ell = \frac{1 + \sum_{i=1}^2 \sum_{j=i+1}^3 (2\delta_{ij}^\ell - 1) \frac{2(\rho_{ij} - \bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-}}{4}, \quad \ell = 1, 2, 3, 4. \quad (2.26)$$

Consider the sets

$$T_\ell = \left\{ \boldsymbol{\rho} \left| t_\ell(\boldsymbol{\rho}) = 1 + \sum_{i=1}^2 \sum_{j=i+1}^3 (2\delta_{ij}^\ell - 1) \frac{2(\rho_{ij} - \bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-} = 0 \right. \right\}, \quad \ell = 1, 2, 3, 4, \quad (2.27)$$

and refer to Figure 2.1. All convex combinations of $\mathbf{q}_2, \mathbf{q}_3$, and, $\mathbf{q}_4$ belong to T_1 . Similarly, all convex combinations of $\mathbf{q}_1, \mathbf{q}_3$, and, $\mathbf{q}_4$ belong to T_2 , of $\mathbf{q}_1, \mathbf{q}_2$, and, $\mathbf{q}_4$ belong to T_3 and of $\mathbf{q}_1, \mathbf{q}_2$, and, $\mathbf{q}_3$ belong to T_4 . The next proposition and the following corollary establish the relationship between the points in P and the formulas (2.26).

Proposition 2.1 *For any random variable $\mathbf{Y}$,*

$$1 + \sum_{i=1}^2 \sum_{j=i+1}^3 (2\delta_{ij}^\ell - 1) \frac{2(\rho_{ij} - \bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-} \geq 0, \quad \ell = 1, 2, 3, 4, \quad (2.28)$$

are valid inequalities for P .

Proof: The set P contains all correlation points that are convex combinations of the extreme-correlation points, $\mathbf{q}_1$, $\mathbf{q}_2$, $\mathbf{q}_3$, and $\mathbf{q}_4$. It may be verified that $t_1(\mathbf{q}_1) > 0$, $t_2(\mathbf{q}_2) > 0$, $t_3(\mathbf{q}_3) > 0$, and $t_4(\mathbf{q}_4) > 0$. Let $\bar{\mathbf{q}}$ be any convex combination of $\mathbf{q}_1$, $\mathbf{q}_2$, $\mathbf{q}_3$, and $\mathbf{q}_4$. Then $t_1(\bar{\mathbf{q}}) \geq 0$, $t_2(\bar{\mathbf{q}}) \geq 0$, $t_3(\bar{\mathbf{q}}) \geq 0$, and $t_4(\bar{\mathbf{q}}) \geq 0$. $\square$

Corollary 2.1 *For any random variable $\mathbf{Y}$, the valid inequalities (2.28) are facets of P .*

Proof: P has dimension 3. Each T_ℓ , $\ell = 1, 2, 3, 4$, has dimension 2. It follows from Proposition 2.1 and the definition of the sets T_ℓ that each T_ℓ is a face of P and therefore a facet of P . $\square$

The following result establishes that there is a unique solution to (2.23)-(2.24) for any $\boldsymbol{\rho} \in P$ and any value of λ_0 such that $0 \leq \lambda_0 \leq 1$.

Proposition 2.2 *For any value of λ_0 , such that $0 \leq \lambda_0 \leq 1$, and any $\boldsymbol{\rho} \in P$, there is a unique solution to (2.23)-(2.24).*

Proof: For any value of λ_0 such that $0 \leq \lambda_0 \leq 1$, the equations (2.23)-(2.24) reduce to:

$$\lambda_1 \rho_{12}^+ + \lambda_2 \rho_{12}^- + \lambda_3 \rho_{12}^- + \lambda_4 \rho_{12}^+ = \rho_{12} \quad (2.29)$$

$$\lambda_1 \rho_{13}^- + \lambda_2 \rho_{13}^+ + \lambda_3 \rho_{13}^- + \lambda_4 \rho_{13}^+ = \rho_{13} \quad (2.30)$$

$$\lambda_1 \rho_{23}^- + \lambda_2 \rho_{23}^- + \lambda_3 \rho_{23}^+ + \lambda_4 \rho_{23}^+ = \rho_{23} \quad (2.31)$$

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1 - \lambda_0. \quad (2.32)$$

In matrix terms this can be written as $\mathbf{A}\mathbf{x} = \mathbf{b}$, where

$$\mathbf{A} = \begin{pmatrix} \rho_{12}^+ & \rho_{12}^+ & \rho_{12}^- & \rho_{12}^- \\ \rho_{13}^+ & \rho_{13}^- & \rho_{13}^+ & \rho_{13}^- \\ \rho_{23}^+ & \rho_{23}^- & \rho_{23}^+ & \rho_{23}^- \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad (2.33)$$

$\mathbf{x} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)^T$, and $\mathbf{b} = (\rho_{12}, \rho_{13}, \rho_{23}, 1 - \lambda_0)^T$. If $\det(\mathbf{A}) \neq 0$, then $\mathbf{A}^{-1}$ exists and $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ is the unique solution. Performing elementary row and column operations on $\mathbf{A}$ yields the following matrix:

$$\mathbf{A}' = \begin{pmatrix} 0 & 0 & \rho_{12}^- - \rho_{12}^+ & \rho_{12}^- - \rho_{12}^+ \\ 0 & \rho_{13}^- - \rho_{13}^+ & 0 & \rho_{13}^- - \rho_{13}^+ \\ 0 & \rho_{23}^- - \rho_{23}^+ & \rho_{23}^- - \rho_{23}^+ & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \quad (2.34)$$

It follows that $\det(\mathbf{A}) = \det(\mathbf{A}') = -2(\rho_{12}^- - \rho_{12}^+)(\rho_{13}^- - \rho_{13}^+)(\rho_{23}^- - \rho_{23}^+)$. Since $\rho_{ij}^- < 0$ and $\rho_{ij}^+ > 0$ for all $i < j$, $\det(\mathbf{A}) < 0$. $\square$

The next two propositions establish that the values for λ_ℓ , $\ell = 1, 2, 3, 4$, given in (2.26) and $\lambda_0 = 0$ satisfy (2.23)-(2.25), and therefore constitute unique composition probabilities.

Proposition 2.3 *The values for λ_ℓ , $\ell = 1, 2, 3, 4$, given in (2.26) and $\lambda_0 = 0$, satisfy (2.23).*

Proof: Without loss of generality, consider equation (2.23) for any $i < j \leq k$:

$$\sum_{\ell=1}^4 [\delta_{ij}^\ell \lambda_\ell \rho_{ij}^+ + (1 - \delta_{ij}^\ell) \lambda_\ell \rho_{ij}^-] = \sum_{\ell=1}^4 \delta_{ij}^\ell \lambda_\ell \rho_{ij}^+ + \sum_{\ell=1}^4 (1 - \delta_{ij}^\ell) \lambda_\ell \rho_{ij}^-$$

$$\begin{aligned}
&= \rho_{ij}^+ \sum_{\ell=1}^4 \delta_{ij}^\ell \left[\frac{1 + \sum_{i=1}^2 \sum_{j=i+1}^3 (2\delta_{ij}^\ell - 1) \frac{2(\rho_{ij}^+ - \bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-}}{4} \right] + \\
&\quad \rho_{ij}^- \sum_{\ell=1}^4 (1 - \delta_{ij}^\ell) \left[\frac{1 + \sum_{i=1}^2 \sum_{j=i+1}^3 (2\delta_{ij}^\ell - 1) \frac{2(\rho_{ij}^+ - \bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-}}{4} \right] \\
&= 2\rho_{ij}^+ \left[\frac{1 + \frac{2(\rho_{ij}^+ - \bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-}}{4} \right] + 2\rho_{ij}^- \left[\frac{1 - \frac{2(\rho_{ij}^+ - \bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-}}{4} \right] \\
&= \rho_{ij}^+ \left[\frac{1}{2} + \frac{\rho_{ij}^+ - \bar{\rho}_{ij}}{\rho_{ij}^+ - \rho_{ij}^-} \right] + \rho_{ij}^- \left[\frac{1}{2} - \frac{\rho_{ij}^+ - \bar{\rho}_{ij}}{\rho_{ij}^+ - \rho_{ij}^-} \right] \\
&= \frac{(\rho_{ij}^+)^2 - (\rho_{ij}^-)^2 + 2\rho_{ij}(\rho_{ij}^+ - \rho_{ij}^-) - 2\bar{\rho}_{ij}(\rho_{ij}^+ - \rho_{ij}^-)}{2(\rho_{ij}^+ - \rho_{ij}^-)} \\
&= \frac{(\rho_{ij}^+ + \rho_{ij}^-)(\rho_{ij}^+ - \rho_{ij}^-) + (\rho_{ij}^+ - \rho_{ij}^-)(2\rho_{ij} - 2\bar{\rho}_{ij})}{2(\rho_{ij}^+ - \rho_{ij}^-)} \\
&= \frac{(\rho_{ij}^+ + \rho_{ij}^-) + 2(\rho_{ij} - \bar{\rho}_{ij})}{2} \\
&= \bar{\rho}_{ij} + \rho_{ij} - \bar{\rho}_{ij} \\
&= \rho_{ij}. \quad \square
\end{aligned}$$

Proposition 2.4 For any $\boldsymbol{\rho} \in P$, the values of λ_ℓ , $\ell = 1, 2, 3, 4$, in (2.26) and $\lambda_0 = 0$ satisfy (2.24) and (2.25).

Proof: Let $\boldsymbol{\rho} \in P$. By Proposition 2.1, the numerator of λ_ℓ , $\ell = 1, 2, 3, 4$, is nonnegative. Therefore, $\lambda_\ell \geq 0$ for $\ell = 1, 2, 3, 4$. Consider $\sum_{\ell=1}^4 \lambda_\ell$. For any $i < j$,

$$\sum_{\ell=1}^4 (2\delta_{ij}^\ell - 1) \frac{2(\rho_{ij}^+ - \bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-} = 0, \tag{2.35}$$

so that $\sum_{\ell=1}^4 \lambda_\ell = \sum_{\ell=1}^4 1/4 = 1$. $\square$

The next two propositions establish more results regarding Type L distributions. The first result provides the conditions under which a Type L distribution exists for $\boldsymbol{\rho} = \mathbf{q}_0 \in P$, and the second result indicates that there are many Type L distributions with $\lambda_0 = 0$ if there is such a distribution for $\boldsymbol{\rho} = \mathbf{q}_0 \in P$.

Proposition 2.5 *There always exists a Type L composite distribution for $\rho = \mathbf{q}_0 \in P$ for which $\lambda_0 = 0$ whenever*

$$\sum_{i=1}^2 \sum_{j=i+1}^3 (2\delta_{ij}^\ell - 1) \frac{\rho_{ij}^+ + \rho_{ij}^-}{\rho_{ij}^+ - \rho_{ij}^-} \leq 1 \quad \ell = 1, 2, 3, 4. \quad (2.36)$$

Proof: Let $\rho = (0, 0, 0)$ and $\lambda_0 = 0$. Condition (2.36) ensures that $\lambda_\ell \geq 0, \ell = 1, 2, 3, 4$. $\square$

For many distributions used in practical applications, condition (2.36) is easily satisfied. For instance, if every marginal distribution is uniform, then $\rho_{ij}^+ + \rho_{ij}^- = 0$ for all $i < j \leq 3$.

Proposition 2.6 *If there is a Type L distribution with $\lambda_0 = 0$ associated with the correlation point $\mathbf{q}_0 \in P$, then there is a Type L distribution with $\lambda_0 = 0$ associated with every $\rho \in P$.*

Proof: Suppose there is a Type L distribution with $\lambda_0 = 0$ associated with $\mathbf{q}_0 \in P$. Then there exist $\alpha_\ell, \ell = 0, 1, 2, 3, 4$, with $\alpha_0 = 0$ such that

$$\mathbf{q}_0 = \sum_{\ell=1}^4 \alpha_\ell \mathbf{q}_\ell, \quad (2.37)$$

$\sum_{\ell=1}^4 \alpha_\ell = 1$, and $\alpha_\ell \geq 0, \ell = 1, 2, 3, 4$.

Select any $\rho \in P$. There exists a composition probability vector, λ , such that

$$\begin{aligned} \rho &= \sum_{\ell=0}^4 \lambda_\ell \mathbf{q}_\ell \\ &= \lambda_0 \mathbf{q}_0 + \sum_{\ell=1}^4 \lambda_\ell \mathbf{q}_\ell \end{aligned}$$

$$\begin{aligned}
&= \lambda_0 \sum_{\ell=1}^4 \alpha_\ell \mathbf{q}_\ell + \sum_{\ell=1}^4 \lambda_\ell \mathbf{q}_\ell \\
&= \sum_{\ell=1}^4 (\lambda_0 \alpha_\ell + \lambda_\ell) \mathbf{q}_\ell \\
&= \sum_{\ell=1}^4 \lambda'_\ell \mathbf{q}_\ell,
\end{aligned}$$

where $\lambda'_\ell = (\lambda_0 \alpha_\ell + \lambda_\ell)$, $\ell = 1, 2, 3, 4$. Since

$$\sum_{\ell=1}^4 \lambda'_\ell = \sum_{\ell=1}^4 (\lambda_0 \alpha_\ell + \lambda_\ell) = \lambda_0 \sum_{\ell=1}^4 \alpha_\ell + \sum_{\ell=1}^4 \lambda_\ell = \lambda_0 + (1 - \lambda_0) = 1, \quad (2.38)$$

one can set $\lambda'_0 = 0$ and there is a Type L composite distribution associated with $\boldsymbol{\rho} \in P$. $\square$

Additional composition probabilities for $\boldsymbol{\rho} \in P$ such that $\lambda_0 \geq 0$ are presented in the next subsection.

2.4.2 Other composite distributions for trivariate random variables

Let $\boldsymbol{\rho} \in P$,

$$d_\ell = 1 - \sum_{i=1}^2 \sum_{j=i+1}^3 (2\delta_{ij}^\ell - 1) \frac{2\bar{\rho}_{ij}}{\rho_{ij}^+ - \rho_{ij}^-}, \quad \ell = 1, 2, 3, 4, \quad (2.39)$$

and

$$0 \leq \gamma_0 \leq \gamma^* = \min_{\ell} \left\{ \frac{4\lambda_\ell}{d_\ell} \right\}. \quad (2.40)$$

The following three propositions establish that a composite distribution for $\mathbf{Y}$ with feasible correlation structure $\boldsymbol{\rho}$ is given by

$$h(\mathbf{y}) = \sum_{\ell=0}^4 \gamma_\ell h_\ell(\mathbf{y}), \quad (2.41)$$

where

$$\gamma_\ell = \frac{1 - \gamma_0 + \sum_{i=1}^2 \sum_{j=i+1}^3 (2\delta_{ij}^\ell - 1) \frac{2(\rho_{ij} - (1-\gamma_0)\bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-}}{4}, \quad \ell = 1, 2, 3, 4. \quad (2.42)$$

Proposition 2.7 *Let $\rho \in P$. The values of $\gamma_\ell, \ell = 1, 2, 3, 4$, given in (2.42), and γ_0 , given in (2.40), satisfy (2.23).*

Proof: Without loss of generality, consider equation (2.23) for any $i < j \leq k$:

$$\begin{aligned} \sum_{\ell=1}^4 [\delta_{ij}^\ell \gamma_\ell \rho_{ij}^+ + (1 - \delta_{ij}^\ell) \gamma_\ell \rho_{ij}^-] &= \gamma_0(0) + \sum_{\ell=1}^4 \delta_{ij}^\ell \gamma_\ell \rho_{ij}^+ + \sum_{\ell=1}^4 (1 - \delta_{ij}^\ell) \gamma_\ell \rho_{ij}^- \\ &= \rho_{ij}^+ \sum_{\ell=1}^4 \delta_{ij}^\ell \left[\frac{1 - \gamma_0 + \sum_{i=1}^2 \sum_{j=i+1}^3 (2\delta_{ij}^\ell - 1) \frac{2(\rho_{ij} - (1-\gamma_0)\bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-}}{4} \right] + \\ &\quad \rho_{ij}^- \sum_{\ell=1}^4 (1 - \delta_{ij}^\ell) \left[\frac{1 - \gamma_0 + \sum_{i=1}^2 \sum_{j=i+1}^3 (2\delta_{ij}^\ell - 1) \frac{2(\rho_{ij} - (1-\gamma_0)\bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-}}{4} \right] \\ &= \rho_{ij}^+ \left[\frac{1}{2} - \frac{\gamma_0}{2} + \frac{\rho_{ij} - (1-\gamma_0)\bar{\rho}_{ij}}{\rho_{ij}^+ - \rho_{ij}^-} \right] + \rho_{ij}^- \left[\frac{1}{2} - \frac{\gamma_0}{2} - \frac{\rho_{ij} - (1-\gamma_0)\bar{\rho}_{ij}}{\rho_{ij}^+ - \rho_{ij}^-} \right] \\ &= \frac{(\rho_{ij}^+ + \rho_{ij}^-)(\rho_{ij}^+ - \rho_{ij}^-) - \gamma_0(\rho_{ij}^+ + \rho_{ij}^-)(\rho_{ij}^+ - \rho_{ij}^-) + 2\rho_{ij}(\rho_{ij}^+ - \rho_{ij}^-)}{2(\rho_{ij}^+ - \rho_{ij}^-)} \\ &\quad - \frac{2(1-\gamma_0)\bar{\rho}_{ij}(\rho_{ij}^+ - \rho_{ij}^-)}{2(\rho_{ij}^+ - \rho_{ij}^-)} \\ &= \bar{\rho}_{ij} - (\gamma_0 \bar{\rho}_{ij} + (1-\gamma_0)\bar{\rho}_{ij}) + \rho_{ij} \\ &= \rho_{ij}. \quad \square \end{aligned}$$

Proposition 2.8 *If $\lambda_0 = 0$ and $\lambda_\ell > 0$ for $\ell = 1, 2, 3, 4$, then $d_\ell > 0$ for $\ell = 1, 2, 3, 4$.*

Proof: For any specified marginals, d_ℓ is a constant for all population correlation structures. Suppose that $\lambda_0 = 0$ and $\lambda_\ell \geq 0, \ell = 1, 2, 3, 4$. Consider any λ_ℓ for the

correlation point $\boldsymbol{\rho} = (0, 0, 0)$. Then,

$$\begin{aligned}\lambda_\ell &= \frac{1}{4} \left[1 + \sum_{i=1}^2 \sum_{j=i+1}^3 \frac{(2\delta_{ij}^\ell - 1)2\rho_{ij}}{\rho_{ij}^+ - \rho_{ij}^-} - \sum_{i=1}^2 \sum_{j=i+1}^3 \frac{(2\delta_{ij}^\ell - 1)2\bar{\rho}_{ij}}{\rho_{ij}^+ - \rho_{ij}^-} \right] \\ &= \frac{1}{4} d_\ell > 0.\end{aligned}\quad (2.43)$$

So $d_\ell > 0, \ell = 1, 2, 3, 4$. $\square$

Proposition 2.9 *If $\lambda_0 = 0$ and $\lambda_\ell > 0, \ell = 1, 2, 3, 4$, then the values of $\gamma_\ell, \ell = 0, 1, 2, 3, 4$, given by (2.40) and (2.42) satisfy (2.24) and (2.25).*

Proof: Suppose that $\lambda_0 = 0$ and $\lambda_\ell > 0, \ell = 1, 2, 3, 4$. For every $\ell, \ell = 1, 2, 3, 4$,

$$0 \leq \gamma_0 \leq \frac{4\lambda_\ell}{d_\ell}. \quad (2.44)$$

It follows that

$$\begin{aligned}\gamma_0 d_\ell &\leq 4\lambda_\ell \\ \gamma_0 \left[1 - \sum_{i=1}^2 \sum_{j=i+1}^3 \frac{(2\delta_{ij}^\ell - 1)2\bar{\rho}_{ij}}{\rho_{ij}^+ - \rho_{ij}^-} \right] &\leq 1 + \sum_{i=1}^2 \sum_{j=i+1}^3 \frac{(2\delta_{ij}^\ell - 1)2(\rho_{ij} - \bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-} \\ 1 - \gamma_0 + \sum_{i=1}^2 \sum_{j=i+1}^3 \frac{(2\delta_{ij}^\ell - 1)2(\rho_{ij} - (1 - \gamma_0)\bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-} &\geq 0 \\ 4\gamma_\ell &\geq 0.\end{aligned}$$

Therefore, $\gamma_\ell \geq 0, \ell = 1, 2, 3, 4$. Similar to arguments in the proof of Proposition 2.4, $\sum_{\ell=1}^4 \gamma_\ell = 1 - \gamma_0$ so that $\gamma_0 + \sum_{\ell=1}^4 \gamma_\ell = 1$. $\square$

Note that the composition probability formula (2.42) reduces to the Type L formula (2.26) when $\gamma_0 = 0$. The value of γ_0 may be thought of as an index for the composite distributions for a specified $\rho \in P$. For any $\rho \in P$, if $\lambda_\ell \geq 0$ for $\ell = 1, 2, 3, 4$, with at least one $\lambda_\ell = 0$, then $\gamma^* = 0$. In this case, the unique distribution for ρ is both Type L and Type U. The next result shows that the composite distribution is Type U when $\gamma_0 = \gamma^*$.

Proposition 2.10 *When $\gamma_0 = \gamma^*$, composite distributions (2.41) are Type U distributions.*

Proof: Let $\gamma_0 = \gamma^*$. Then using the composition probability formula (2.42) yields

$$\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4) = (\gamma^*, \lambda_1 - \frac{\gamma^*}{4}d_1, \lambda_2 - \frac{\gamma^*}{4}d_2, \lambda_3 - \frac{\gamma^*}{4}d_3, \lambda_4 - \frac{\gamma^*}{4}d_4). \quad (2.45)$$

From Propositions 2.7 and 2.9, the vector γ represents a feasible set of composition probabilities. Since $\gamma_0 = \gamma^*$, $\gamma_i = 0$ for at least one i , $i = 1, 2, 3, 4$. Without loss of generality, assume that $\gamma_1 = 0$. Then γ satisfies (2.23)-(2.25) which reduce to

$$\gamma_0 + \gamma_2 + \gamma_3 + \gamma_4 = 1, \quad (2.46)$$

$$\gamma_2\rho_{12}^- + \gamma_3\rho_{12}^- + \gamma_4\rho_{12}^+ = \rho_{12}^0, \quad (2.47)$$

$$\gamma_2\rho_{13}^+ + \gamma_3\rho_{13}^- + \gamma_4\rho_{13}^+ = \rho_{13}^0, \quad (2.48)$$

$$\gamma_2\rho_{23}^- + \gamma_3\rho_{23}^+ + \gamma_4\rho_{23}^+ = \rho_{23}^0, \quad (2.49)$$

$$\gamma_0, \gamma_2, \gamma_3, \gamma_4 \geq 0. \quad (2.50)$$

The dual of the linear program (LP) with the objective of maximizing λ_0 subject to constraints (2.46)-(2.50) is

$$\text{Minimize } w_1 + \rho_{12}^0 w_2 + \rho_{13}^0 w_3 + \rho_{23}^0 w_4 \quad (2.51)$$

subject to

$$w_1 \geq 1, \quad (2.52)$$

$$w_1 + \rho_{12}^- w_2 + \rho_{13}^+ w_3 + \rho_{23}^- w_4 \geq 0, \quad (2.53)$$

$$w_1 + \rho_{12}^- w_2 + \rho_{13}^- w_3 + \rho_{23}^+ w_4 \geq 0, \quad (2.54)$$

$$w_1 + \rho_{12}^+ w_2 + \rho_{13}^+ w_3 + \rho_{23}^+ w_4 \geq 0. \quad (2.55)$$

The complementary slackness conditions (CSC) are

$$(w_1 - 1)\gamma_0 = 0, \quad (2.56)$$

$$(w_1 + \rho_{12}^- w_2 + \rho_{13}^+ w_3 + \rho_{23}^- w_4)\gamma_2 = 0, \quad (2.57)$$

$$(w_1 + \rho_{12}^- w_2 + \rho_{13}^- w_3 + \rho_{23}^+ w_4)\gamma_3 = 0, \quad (2.58)$$

$$(w_1 + \rho_{12}^+ w_2 + \rho_{13}^+ w_3 + \rho_{23}^+ w_4)\gamma_4 = 0. \quad (2.59)$$

If $\gamma_0 = \gamma^* > 0$, then the CSC imply $w_1 = 1$. If $\gamma_0 = \gamma^* = 0$, then $w_1 \leq 1$. In either case, the constraints (2.53)-(2.55) and the CSC (2.57)-(2.59) are satisfied if

$$\mathbf{A} = \begin{bmatrix} \rho_{12}^- & \rho_{13}^+ & \rho_{23}^- \\ \rho_{12}^- & \rho_{13}^- & \rho_{23}^+ \\ \rho_{12}^+ & \rho_{13}^+ & \rho_{23}^+ \end{bmatrix} \quad (2.60)$$

is nonsingular. $\mathbf{A}^T$ is a square submatrix of the nonsingular matrix

$$\mathbf{D} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & \rho_{12}^- & \rho_{12}^- & \rho_{12}^+ \\ 0 & \rho_{13}^+ & \rho_{13}^- & \rho_{13}^+ \\ 0 & \rho_{23}^- & \rho_{23}^+ & \rho_{23}^+ \end{bmatrix} \quad (2.61)$$

derived from (2.46)-(2.49). Because $\mathbf{D}^{-1}$ exists, $(\mathbf{A}^T)^{-1}$ exists and $\mathbf{A}$ must be nonsingular. Since $\mathbf{A}$ is nonsingular, (2.45) is the composition probability vector for a Type U composite distribution. $\square$

2.4.3 Feasible correlation points for trivariate random variables

A correlation structure for a trivariate random variable has three interdependent correlation terms. The dependency among the three correlation terms means that specifying values for any two correlation terms limits the range of feasible values for the remaining correlation term (Olkin, 1981).

Assume $\mathbf{Y} = (Y_1, Y_2, Y_3)$ with $\rho_{13} = \rho_{13}^0$ and $\rho_{23} = \rho_{23}^0$. Equations (2.26) can be solved for the remaining correlation term, ρ_{12} . Two equations provide upper bounds and two equations provide lower bounds on the feasible range of ρ_{12} . The range for ρ_{12} is given by

$$\frac{\tau(\rho_{12}, \rho_{13}^0, \rho_{23}^0)(\rho_{12}^+ - \rho_{12}^-)}{2} + \bar{\rho}_{12} \leq \rho_{12} \leq \frac{\eta(\rho_{12}, \rho_{13}^0, \rho_{23}^0)(\rho_{12}^+ - \rho_{12}^-)}{2} + \bar{\rho}_{12}, \quad (2.62)$$

where

$$\tau(\rho_{12}, \rho_{13}^0, \rho_{23}^0) = -\max \left\{ 1 \pm \left(\frac{2(\rho_{13}^0 - \bar{\rho}_{13})}{\rho_{13}^+ - \rho_{13}^-} + \frac{2(\rho_{23}^0 - \bar{\rho}_{23})}{\rho_{23}^+ - \rho_{23}^-} \right) \right\}, \quad (2.63)$$

and

$$\eta(\rho_{12}, \rho_{13}^0, \rho_{23}^0) = \min \left\{ 1 \pm \left| \frac{2(\rho_{13}^0 - \bar{\rho}_{13})}{\rho_{13}^+ - \rho_{13}^-} + \frac{2(\rho_{23}^0 - \bar{\rho}_{23})}{\rho_{23}^+ - \rho_{23}^-} \right| \right\}. \quad (2.64)$$

Example 3: Suppose that $\mathbf{Y} = (Y_1, Y_2, Y_3)$ and each of the $Y_i, i = 1, 2, 3$ is a negative exponential random variable with $\rho_{12} = 0.6$ and $\rho_{13} = 0.65$ specified. Applying (2.62)-(2.64) gives the range $0.25 \leq \rho_{23} \leq 0.95$ to ensure $\boldsymbol{\rho} = (\rho_{12}, \rho_{13}, \rho_{23}) \in P$. $\square$

2.5 Extensions for General Multivariate Random Variables

In this section, limitations of extending the composition probability formulas (2.26) and (2.42) for trivariate random variables to general multivariate random variables are discussed. Examples are provided to demonstrate these limitations.

A convenient approach to finding composition probabilities for Type L composite distributions when $k \geq 4$ might involve an extended form of (2.26) with

$$\lambda_\ell = \frac{1 + \sum_{i=1}^{k-1} \sum_{j=i+1}^k (2\delta_{ij}^\ell - 1) \frac{2(\rho_{ij}^+ - \bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-}}{2^{k-1}}, \quad \ell = 1, 2, \dots, 2^{k-1}, \quad (2.65)$$

and $\lambda_0 = 0$. Proposition 2.3 may be extended to prove that for $k \geq 4$ values based on (2.65) satisfy (2.23), and it may be shown that $\sum_{\ell=1}^{2^{k-1}} \lambda_\ell = 1$ if $\lambda_0 = 0$. For some $\rho \in P$, the values suggested by (2.65) are valid composition probabilities. However, for some $\rho \in P$, the values suggested by (2.65) violate (2.25).

Example 4. Let $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$, where Y_i , $i = 1, 2, 3, 4$, are identical discrete uniform random variables so that $\rho_{ij}^+ = 1.0 = -\rho_{ij}^-$ for all $i < j \leq 4$. Suppose that the desired correlation matrix is

$$\mathbf{R}_1 = \begin{pmatrix} 1 & 0 & \rho_{13}^+/4 & \rho_{14}^-/16 \\ 0 & 1 & \rho_{23}^+/8 & 0 \\ \rho_{13}^+/4 & \rho_{23}^+/8 & 1 & \rho_{34}^-/8 \\ \rho_{14}^-/16 & 0 & \rho_{34}^-/8 & 1 \end{pmatrix}. \quad (2.66)$$

Suppose that $\lambda_0 = 0$. Then, the λ_ℓ values suggested by (2.65) are: $\lambda_1 = 13/128$; $\lambda_2, \lambda_4 = 15/128$; $\lambda_3 = 21/128$; $\lambda_5 = 9/128$; $\lambda_6 = 11/128$; $\lambda_7 = 25/128$; and $\lambda_8 = 19/128$. In this case, the application of (2.65) yields valid composition probabilities.

□

Example 5. Recall Example 4. Suppose that, instead of $\mathbf{R}_1$, the desired correlation matrix is represented by a correlation point nearer to an extreme-correlation point. For instance, suppose the correlation matrix is

$$\mathbf{R}_2 = \begin{pmatrix} 1 & 0.9 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 & 0.9 \\ 0.9 & 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 0.9 & 1 \end{pmatrix}. \quad (2.67)$$

Let $\lambda_0 = 0$. The λ_ℓ values suggested by (2.65) are: $\lambda_1, \lambda_4, \lambda_6, \lambda_7 = 1/8$; $\lambda_2, \lambda_3, \lambda_5 = -1/10$; and $\lambda_8 = 8/10$. Solving an LP with the objective of minimizing λ_0 to obtain a feasible λ for a Type L composite distribution yields: $\lambda_1, \lambda_4, \lambda_6, \lambda_7 = 1/40$; $\lambda_0, \lambda_2, \lambda_3, \lambda_5 = 0$; and $\lambda_8 = 9/10$. $\square$

The following generalization of Propositions 2.5 and 2.6 indicate when the formula (2.65) will provide valid composition probabilities.

Proposition 2.11 *There always exists a Type L composite distribution for $\rho = \mathbf{q}_0 \in P$ for which $\lambda_0 = 0$ whenever*

$$\sum_{i=1}^{k-1} \sum_{j=i+1}^k (2\delta_{ij}^\ell - 1) \frac{\rho_{ij}^+ + \rho_{ij}^-}{\rho_{ij}^+ - \rho_{ij}^-} \leq 1 \quad \ell = 1, 2, \dots, 2^{k-1}. \quad (2.68)$$

Proof: Let $\rho = (0, 0, \dots, 0)$ be the $\binom{k}{2}$ -component zero vector and $\lambda_0 = 0$. Condition (2.68) ensures that $\lambda_\ell \geq 0, \ell = 1, 2, \dots, 2^{k-1}$. $\square$

Proposition 2.12 *If there is a Type L distribution with $\lambda_0 = 0$ associated with the correlation point $\mathbf{q}_0 \in P$, then there is a Type L distribution with $\lambda_0 = 0$ associated with every $\rho \in P$.*

Table 2.5: Upper Bound on λ_ℓ as k increases

k	Number of ρ_{ij} s	Number of $\mathbf{q}_\ell$	Maximum λ_ℓ
2	1	2	1.0000
3	3	4	1.0000
4	6	8	0.8750
5	10	16	0.6875
6	15	32	0.5000

Proof: The proof of this proposition is virtually identical to the proof of Proposition 2.6. $\square$

The formula (2.65) does not always yield valid composition probabilities because, for $k > 3$, the numerators of (2.65) are not facets of P . Further, the number of extreme-correlation points $\mathbf{q}_\ell$, 2^{k-1} , grows faster with k than with the number of ρ_{ij} s, $\binom{k}{2}$. This means the denominator of (2.65) grows faster than the numerator, and the maximum value of any λ_ℓ decreases with increasing k as seen in Table 2.5. For example, when $k \geq 4$ and $\boldsymbol{\rho} = \mathbf{q}_\ell$, for any $\ell = 1, 2, \dots, 2^{k-1}$, it does not follow from application of (2.65) that $\lambda_\ell = 1$ as is expected.

Although the composition probabilities (2.26) are not readily extendable to distributions of general multivariate random variables, the values obtained through application of (2.65) can be used in conjunction with a composition weight adjustment method to provide valid composition probabilities when $k = 4$.

2.6 Composite Distributions for Quadrivariate Random Variables

In this section composite distributions for quadrivariate random variables are constructed by adjusting the composition weights obtained from (2.65). The adjustment process is described first and then a procedure implementing the adjustment process is presented.

2.6.1 Adjusting Composition Weight Vectors

Assume for $k \geq 4$ and population correlation structure $\rho \in P$, that a vector of composition weights λ from (2.65) is not nonnegative (i.e., λ violates (2.25)). One way to obtain a vector of composition probabilities is to adjust λ to satisfy (2.23)-(2.25).

Let $i^* = \arg \max\{\lambda_\ell\}$, and $j^* = \arg \min\{\lambda_\ell\}$. Suppose that $\lambda_{j^*} < 0$ so that (2.25) is violated. Assume that λ_0 will not be changed. One can partition $\{\lambda_1, \lambda_2, \dots, \lambda_8\}$ by defining index sets L and R such that the indices in L identify elements of λ that decrease in value while indices in R identify elements of λ that increase in value. Sets L and R effectively partition the rows in Table 2.4 in such a way that for any column in the table, an equal number of 1s (and thus 0s) are in L and R . Clearly, this requires $|L| = |R| = 4$. Any offsetting adjustments to λ based on the index sets L and R produce an alternative λ that still satisfies (2.23)-(2.24). To ensure that the alternative λ also satisfies (2.25), it must be that $j^* \in R$ and that each adjustment be at least $-\lambda_{j^*}$.

Define $\delta_{1\cdot}^\ell = (\delta_{12}^\ell, \delta_{13}^\ell, \delta_{14}^\ell)$ for each vector δ^ℓ and define the degree of agreement of $\delta_{1\cdot}^\ell$ as

$$D(\delta_{1\cdot}^\ell) = \sum_{j=2}^4 (1 - |\delta_{1j}^\ell - \delta_{1j}^{i*}|), \quad \ell = 1, 2, \dots, 8. \quad (2.69)$$

The degree of agreement function (2.69) provides a convenient method of determining L and R :

$$L = \{i | D(\delta_{1\cdot}^i) \in \{0, 2\}, i = 1, 2, \dots, 8\}, \quad (2.70)$$

and

$$R = \{i | i \notin L\}. \quad (2.71)$$

Conveniently, for $k = 4$, $L = (\lambda_1, \lambda_4, \lambda_6, \lambda_7)$ and $R = (\lambda_2, \lambda_3, \lambda_5, \lambda_8)$, or vice versa.

Given a vector λ from (2.65) that violates (2.25), one may use Procedure ADJUST to raise λ_{j^*} to zero.

Procedure ADJUST

1. Let $i^* = \arg \max_\ell \{\lambda_\ell\}$, $j^* = \arg \min_\ell \{\lambda_\ell\}$, and $\epsilon = -\lambda_{j^*}$.
2. Define L and R according to (2.70) and (2.71).
3. For $\ell = 1, 2, \dots, 8$, do
 - (a) If $\ell \in L$ then $\lambda_\ell = \lambda_\ell - \epsilon$
 - (b) If $\ell \in R$ then $\lambda_\ell = \lambda_\ell + \epsilon$
4. Return.

In some cases, the adjusted set of weights are the same as would be obtained using a LP to find a Type L distribution.

Example 6. Recall Example 5. Procedure ADJUST returns: $\lambda_1, \lambda_4, \lambda_6, \lambda_7 = 1/40$; $\lambda_2, \lambda_3, \lambda_5 = 0$; and $\lambda_8 = 9/10$ with $\lambda_0 = 0$ based on $\epsilon = 1/10$. $\square$

If λ violates (2.25) after ADJUST, then we conclude that $\rho \notin P$. This does not imply that ρ does not represent a valid correlation structure, but rather that ρ may not be expressed as a convex combination of the points $\mathbf{q}_\ell$, $\ell = 0, 1, \dots, 2^{k-1}$.

Example 7. Consider normal random variables, $Y_i, i = 1, 2, 3, 4$ and

$$\boldsymbol{\rho} = (0.959, 0.979, -0.904, 0.951, -0.819, -0.879). \quad (2.72)$$

The determinant of the corresponding correlation matrix is 0.000492; so it represents a valid correlation structure. Using (2.65) yields: $\gamma_1 = -0.097625$, $\gamma_2 = 0.100875$, $\gamma_3 = 0.129125$, $\gamma_4 = -0.111875$, $\gamma_5 = 0.109125$, $\gamma_6 = -0.101875$, $\gamma_7 = 0.811345$, and $\gamma_8 = 0.160875$. Then $\text{argmin}_\ell \{\gamma_\ell\} = 4$, $2 \in R$, and $4 \in L$. Since $|\gamma_4| > |\gamma_2|$, ADJUST does not return a valid set of probability weights. Therefore, $\boldsymbol{\rho} \notin P$, a fact easily verified using a LP. $\square$

A very similar process may be applied to the construction of non-Type L composite distributions, i.e., composite distributions with $\lambda_0 > 0$. Extending formula (2.42) to the quadrivariate case gives:

$$\gamma_\ell = \frac{1 - \gamma_0 + \sum_{i=1}^3 \sum_{j=i+1}^4 (2\delta_{ij}^\ell - 1) \frac{2(\rho_{ij} - (1-\gamma_0)\bar{\rho}_{ij})}{\rho_{ij}^+ - \rho_{ij}^-}}{8}, \quad \ell = 1, 2, \dots, 8, \quad (2.73)$$

for $0 \leq \gamma_0 \leq \gamma^*$. The formula for computing γ^* for trivariate random variables given in §2.5 may be applied to each of the trivariate marginal distributions of $\mathbf{Y}$ to determine γ^* when $k = 4$.

The trivariate marginal random variables for $\mathbf{Y}$ are (Y_1, Y_2, Y_3) , (Y_1, Y_2, Y_4) , (Y_1, Y_3, Y_4) , and (Y_2, Y_3, Y_4) . Each of the four trivariate marginal distributions for a quadrivariate random variable is a valid trivariate distribution for three of the four components of $\mathbf{Y}$. For a specified $\boldsymbol{\rho}$, let $\gamma_{(1)}^*$, $\gamma_{(2)}^*$, $\gamma_{(3)}^*$, and $\gamma_{(4)}^*$ be the respective γ^* values for each trivariate marginal random variable and

$$\gamma^* = \min\{\gamma_{(1)}^*, \gamma_{(2)}^*, \gamma_{(3)}^*, \gamma_{(4)}^*\}. \quad (2.74)$$

Let $0 \leq \gamma_0 \leq \gamma^*$. Whenever (2.73) returns a γ vector that violates (2.25) apply Procedure ADJUST to adjust the vector of composition weights leaving γ_0 unchanged.

Example 8. Let $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$, where each $Y_i, i = 1, 2, 3, 4$, is negative exponential with arbitrary means and let

$$\mathbf{R}_3 = \begin{pmatrix} 1 & -0.25 & -0.30 & -0.15 \\ -0.25 & 1 & 0.50 & 0.30 \\ -0.30 & 0.50 & 1 & 0.17 \\ -0.15 & 0.30 & 0.17 & 1 \end{pmatrix}, \quad (2.75)$$

be the desired correlation matrix. For the trivariate margin (Y_1, Y_2, Y_3) , $\gamma_{(1)}^* = 0.45$, for (Y_1, Y_2, Y_4) , $\gamma_{(2)}^* = 0.60$, for (Y_1, Y_3, Y_4) , $\gamma_{(3)}^* = 0.0338$, and for (Y_2, Y_3, Y_4) , $\gamma_{(4)}^* = 0.37$ so that $\gamma^* = 0.0338$. Using $\gamma_0 = 0.0338$, application of (2.73) yields: $\gamma_1 = 0.3746$, $\gamma_2 = 0.2383$, $\gamma_3 = 0.1319$, $\gamma_4 = -0.0054$, $\gamma_5 = 0.1076$, $\gamma_6 = 0.0493$, $\gamma_7 = 0.0645$ and $\gamma_8 = 0.0054$. Procedure ADJUST returns $\gamma_0 = 0.0338$, $\gamma_1 = 0.37996$, $\gamma_2 = 0.2329$, $\gamma_3 = 0.1265$, $\gamma_5 = 0.1022$, $\gamma_6 = 0.0547$, $\gamma_7 = 0.0699$ and $\gamma_4, \gamma_8 = 0.0$, based on $\epsilon = 0.0054$. These results agree with LP results. $\square$

As explained and demonstrated before, if γ violates (2.25) after ADJUST is executed, then we conclude that $\rho \notin P$.

2.6.2 Choosing feasible quadrivariate correlation points

It is useful to be able to choose feasible correlation structures $\rho \in P$ for a quadrivariate random variable. A distribution for a quadrivariate random variable has six correlation terms and each term is associated with two of the four trivariate marginal distributions. The value specified for any $\rho_{ij}, i < j \leq 4$, must be feasible in both trivariate marginal distributions involving ρ_{ij} . For example,

any value for ρ_{23} must be feasible with respect to the trivariate marginal distributions involving both (Y_1, Y_2, Y_3) and (Y_2, Y_3, Y_4) . Determining the feasible range for any ρ_{ij} involves applying (2.62)-(2.64) to the appropriate trivariate marginal distributions.

Example 9. Let $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$ where each Y_i is negative exponential. Suppose that $\rho_{12} = 0.6$, $\rho_{13} = 0.65$, and $\rho_{14} = -0.25$ have been specified. The remaining correlation terms must satisfy

$$0.25 \leq \rho_{23} \leq 0.95, \quad (2.76)$$

$$-0.6397 \leq \rho_{24} \leq 0.15, \quad (2.77)$$

$$-0.6 \leq \rho_{34} \leq 0.1 \quad (2.78)$$

Ranges (2.76)-(2.78) are necessary but not sufficient to completely specify $\boldsymbol{\rho}$. Once either of ρ_{23} , ρ_{24} , or ρ_{34} are specified the remaining two ranges must be recomputed. $\square$

2.7 Summary and Discussion

This research has produced a characterization of composite distributions for multivariate random variables with a specified Pearson product-moment correlation structure using the extreme-correlation distributions and the joint distribution under independence. Type L and Type U distributions represent special types of composite distributions with extreme levels of independent sampling and they define a range of possible composite distributions for a specified correlation structure.

Explicit correlation induction based on composite distributions opens up many new avenues of research, both theoretical and empirical, concerning multivariate sampling and the influence of correlation structure in many types of systems. In future theoretical work, closed-form composition probabilities for higher dimensional random variables could be developed. One could also develop composite distributions based on other measures of dependency, such as Spearman rank correlation or positive regression dependence. There are many empirical research opportunities too. For instance, in the next chapter, two-dimensional knapsack problems are generated based on multivariate composite distributions and the rank correlation induction method of Iman and Conover (1982) to examine the influence of different types of correlation structures on the performance of solution procedures. Computational experiments should also be conducted on other types of optimization problems. In addition, composite distributions could be used in the simulation of tandem queueing and manufacturing systems. More analytically based applications could address issues in the design of experiments or variance reduction for simulation applications involving multiple random number streams and multiple measures of performance.

Appendix - Type L and Type U distributions for bivariate random variables

Type L Distributions

Consider the following LP for determining mixing probabilities for a Type L distribution for a bivariate random variable:

Min

$$Z = \lambda_0 \quad (2.79)$$

subject to

$$\rho^- \lambda_1 + \rho^+ \lambda_2 = \rho^0, \quad (2.80)$$

$$\lambda_0 + \lambda_1 + \lambda_2 = 1, \quad (2.81)$$

$$\lambda_0, \lambda_1, \lambda_2 \geq 0. \quad (2.82)$$

A basic feasible solution to (2.80)-(2.82) with $\lambda_0 = 0$ is given by:

$$\begin{aligned} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} &= \begin{bmatrix} \rho^- & \rho^+ \\ 1 & 1 \end{bmatrix}^{-1} \begin{pmatrix} \rho^0 \\ 1 \end{pmatrix} \\ &= \frac{1}{\rho^- - \rho^+} \begin{bmatrix} 1 & -\rho^+ \\ -1 & \rho^- \end{bmatrix} \begin{pmatrix} \rho^0 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} (\rho^+ - \rho^0)/(\rho^+ - \rho^-) \\ (\rho^0 - \rho^-)/(\rho^+ - \rho^-) \end{pmatrix} \geq 0, \end{aligned} \quad (2.83)$$

which consists of the composition probabilities given in §2.2.1. The complementary dual solution for extreme mixtures is dual feasible since

$$(0, 0) \frac{1}{\rho^- - \rho^+} \begin{bmatrix} 1 & -\rho^+ \\ -1 & \rho^- \end{bmatrix} \begin{pmatrix} 0 & \rho^- & \rho^+ \\ 1 & 1 & 1 \end{pmatrix} = (0, 0, 0) \leq (1, 0, 0). \quad (2.84)$$

Therefore, the composition probabilities given in (2.83) and $\lambda_0 = 0$ are the mixing probabilities for a Type L composite distribution.

Type U Distributions

Now consider the following LP for determining composition probabilities for a Type U distribution for a bivariate random variable:

Max

$$Z = \lambda_0 \quad (2.85)$$

subject to (2.80)-(2.82). Suppose $0 \leq \rho^0 \leq \rho^+$. A basic feasible solution to (2.80)-(2.82) with $\lambda_1 = 0$ is

$$\begin{aligned} \begin{pmatrix} \lambda_0 \\ \lambda_2 \end{pmatrix} &= \begin{bmatrix} 0 & \rho^+ \\ 1 & 1 \end{bmatrix}^{-1} \begin{pmatrix} \rho^0 \\ 1 \end{pmatrix} \\ &= \begin{bmatrix} -1/\rho^+ & 1 \\ 1/\rho^+ & 0 \end{bmatrix} \begin{pmatrix} \rho^0 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 - \rho^0/\rho^+ \\ \rho^0/\rho^+ \end{pmatrix} \geq 0. \end{aligned} \quad (2.86)$$

The complementary dual solution is dual feasible because

$$(1, 0) \begin{pmatrix} -1/\rho^+ & 1 \\ 1/\rho^+ & 0 \end{pmatrix} \begin{pmatrix} 0 & \rho^- & \rho^+ \\ 1 & 1 & 1 \end{pmatrix} = (0, \frac{-\rho^-}{\rho^+} + 1, 0) \geq (1, 0, 0). \quad (2.87)$$

Therefore, the composition probabilities in (2.86) and $\lambda \geq 0$ are the mixing probabilities for a Type U distribution when $0 \leq \rho^0 \leq \rho^+$. A similar argument may be used to show that $\lambda_0 = 1 - \rho^0/\rho^-$, $\lambda_1 = \rho^0/\rho^-$, and $\lambda_2 = 0$ is an optimal solution to (2.85), (2.80)-(2.82) when $\rho^- \leq \rho^0 \leq 0$.

References

- Amini, M. M. and M. Racer. 1994. A Rigorous Computational Comparison of Alternative Solution Methods for the Generalized Assignment Problem. *Management Science*, **40**(7), 868-890.
- Balas, E. and C.H. Martin. 1980. Pivot and Complement - A Heuristic for 0-1 Programming. *Management Science* **26**(1): 86-96.
- Balas, E. and E. Zemel. 1980. An algorithm for large zero-one knapsack problems. *Operations Research* **28**(5): 1130-1154.
- Cario, M. C., J. J. Clifford, R. R. Hill, J. Yang, K. Yang, C. H. Reilly. 1995. Alternative Methods for Generating Synthetic Generalized Assignment Problems. *Working Paper Series Number 1995-006*. Department of Industrial, Welding and Systems Engineering, The Ohio State University, Columbus, Ohio.
- Devroye, L. 1986. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.
- Fisher, M., R. Jaikumar, and L. Van Wassenhove. 1986. A Multiplier Adjustment Method for the Generalized Assignment Problem. *Management Science*, **32**(9), 1095-1103.
- Fréchet, M. 1951. Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon, Section A*, **14**, 53-77.
- Guignard, M. and M.B. Rosenwein. 1989. An Improved Dual Based Algorithm for the Generalized Assignment Problem. *Operations Research*, **37**(4), 658-663.
- Hill, R. R. and C.H. Reilly. 1994. Composition for Multivariate Random Variables. *Proceedings of the 1994 Winter Simulation Conference*, eds. J.T. Tew, S. Manivannan, D.A. Sadowski, and A.F. Seila. 332-342. Institute of Electrical and Electronics Engineers, Orlando Florida.
- Iman, R.L. and W.J. Conover. 1982. A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables. *Communications in Statistics: Simulation and Computation*, **11**(3), 311-334.
- John, T. C. 1989. Tradeoff Solutions in Single Machine Production Scheduling for Minimizing Flow Time and Maximum Penalty. *Computers and Operations Research*, **16**(5), 471-479.
- Martello, S. and P. Toth. 1979. The 0-1 Knapsack Problem. *Combinatorial Optimization*, eds. N. Christofides, A. Mingozzi, C. Sandi. John Wiley and Sons, New York, New York, 237-279.

- Martello, S. and P. Toth. 1981. An algorithm for the generalized assignment problem. in J.P. Brans (eds.), *Operational Research '81*, North-Holland, Amsterdam, 589-603.
- Moore, B. A. and C. H. Reilly. 1993. Randomly Generating Synthetic Optimization Problems with Explicitly Induced Correlation. *OSU/ISE Working Paper Series Number 1993-002*. The Ohio State University, Columbus, Ohio.
- Mazzola, J.B. and A.W. Neebe. 1993. An Algorithm for the Bottleneck Generalized Assignment Problem. *Computers and Operations Research*, **20**(4), 355-362.
- Nelsen, R. B. 1987. Discrete Bivariate Distributions with Given Marginals and Correlation. *Communications in Statistics: Simulation and Computation*, **16**(1), 199-208.
- Olkin, I. 1981. Range Restrictions for Product-Moment Correlation Matrices. *Psychometrika*, **4**(4), 469-472.
- Page, E. S. 1965. On Monte Carlo Methods in Congestion Problems: I. Searching for an Optimum in Discrete Situations. *Operations Research*, **13**(2), 291-305.
- Peterson, J. A. 1990. A Parametric Analysis of a Bottleneck Transportation Problem Applied to the Characterization of Correlated Discrete Random Variables, M.S. Thesis, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, Ohio.
- Peterson, J. A. and C. H. Reilly. 1993. Joint Probability Mass Functions for Coefficients in Synthetic Optimization Problems. *Working Paper Series Number 1993-006*. The Ohio State University, Columbus, Ohio.
- Potts, C. N. and L. N. Van Wassenhove. 1988. Algorithms for Scheduling a Single Machine to Minimize the Weighted Number of Late Jobs. *Management Science*, **34**(7), 843-858.
- Potts, C. N. and L. N. Van Wassenhove. 1992. Single machine scheduling to minimize total late work. *Operations Research*, **40**(3), 586-595.
- Reilly, C. H. 1991. Optimization test problems with uniformly distributed coefficients. *Proceedings of the 1991 Winter Simulation Conference*, eds. B. L. Nelson, W. D. Kelton, G. M. Clark, 866-874. Institute of Electrical and Electronics Engineers, Phoenix, Arizona.
- Rousseeuw, P. J. and G. Molenberghs. 1994. The Shape of Correlation Matrices. *The American Statistician*, **48**(4), 276-279.
- Rushmeier, R. A. and G. L. Nemhauser. 1993. Experiments with Parallel Branch-and-Bound Algorithms for the Set Covering Problem. *Operations Research Letters*, **13**(5), 277-285.

- Schmeiser, B. W. and R. Lal. 1982. Bivariate Gamma Random Vectors. *Operations Research*, **30**(2), 355-374.
- Trick, M. 1982. A Linear Relaxation Heuristic for the Generalized Assignment Problem. *Naval Research Logistics*, **39**(2), 137-151.
- Yang, J. 1994. A Computational Study on 0-1 Knapsack Problems Generated Under Explicit Correlation Induction. MS Thesis, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, Ohio.

CHAPTER III

THE EFFECTS OF COEFFICIENT CORRELATION STRUCTURE IN TWO-DIMENSIONAL KNAPSACK PROBLEMS ON SOLUTION PROCEDURE PERFORMANCE

3.1 Introduction

This chapter presents an empirical study that examines the influence of correlation structure between the coefficients in synthetic optimization problems on solution procedure performance. One reason for empirical testing of solution procedures is to overcome the limitations inherent in deductive, analytical techniques like worst-case and average-case performance analyses, which often require very strong assumptions to ensure mathematical tractability of the results. Hooker (1994) sees the ability of deductive approaches in their current state “inadequate to its task,”

and he views computational testing as the only currently viable alternative. Understanding the influence of correlation structure between the types of coefficients on solution procedure performance is an example where current deductive analysis methods are inadequate.

In many empirical studies of optimization algorithms or heuristics, randomly generated, or synthetic, problems are assumed to be representative of real-world problem instances. However, defining a truly representative set of problems is difficult. The usual practice is to systematically vary the values of each factor across some range, and thereby include a variety of problem instances that may include instances that resemble real problem instances. Any inferences drawn for the entire set of test problems are assumed to apply to problem instances encountered in practice.

For certain classes of optimization problems, such as the multidimensional knapsack problem (MKP), and in particular the two-dimensional knapsack problem (2KP) studied here, the test problem coefficients should be generated by sampling from joint distributions of multivariate random variables. In this study, 2KP coefficients are generated based on a variety of correlation structures. In addition, the type of correlation measure (Pearson product-moment and Spearman rank correlation measures) used as the basis for generating the coefficients is varied. Solution procedure performance results are then examined to assess how correlation structure influences the performance of an algorithm and a heuristic.

Some computational studies have been conducted on test problems in which correlation is induced between the objective function and constraint coefficients. The absolute correlation level has been linked to performance differences for solu-

tion procedures. By generating test problems based on a multivariate distribution, the effect of the correlation between the coefficients in different constraints, as well as the correlations between the objective and constraint coefficients, may be assessed.

In §3.2, studies involving synthetic optimization test problems with correlation induced among the coefficient types are reviewed and past research involving the MKP is summarized. The test problem generation methodologies used in the present study are discussed in §3.3, and the design of the experiment and the analysis methods used in this study are presented in §3.4. Differences in sample distributions due to the correlation induction method are examined in §3.5. The influence of each type of correlation measure on solution procedure performance is discussed in §3.6. Computational results for CPLEX, a branch-and-bound procedure, are discussed in §3.7, while §3.8 provides a similar analysis of the results for the heuristic by Toyoda (1975). There is a brief discussion in §3.9 of how test problem generation parameters influence the size of the LP-IP gap in the synthetic test problems. Finally, §3.10 contains a discussion and concluding remarks.

3.2 Background

This section begins with an introduction to MKP, a class of problems in which 2KP is a special case. A review of previous computational studies involving synthetic optimization problems with correlation induced among the coefficient types follows in §3.2.2. Results of some previous MKP studies are summarized in §3.2.3.

3.2.1 The multidimensional knapsack problem

MKP is a 0-1 programming problem of the following form:

Maximize

$$Z = \sum_{j=1}^n c_j x_j \quad (3.1)$$

subject to

$$\sum_{j=1}^n a_{ij} x_j \leq b_i \quad i = 1, 2, \dots, m, \quad (3.2)$$

$$x_j = 0 \text{ or } 1 \quad j = 1, 2, \dots, n, \quad (3.3)$$

where all $c_j > 0$ and all $a_{ij} \geq 0$. Additionally, at least one $a_{ij} > 0$ for each j . This general form applies to a wide variety of optimization applications, including capital budgeting problems. A special case of MKP is 2KP, where $m = 2$.

MKP is known to be NP-hard (Frieze and Clarke, 1984), meaning that there is no known polynomial-time solution algorithm for MKP. As n increases, exact solution methods, such as branch-and-bound, may require large commitments of computing resources. Consequently, heuristics are often used to find solutions that are close to the optimum at a fraction of the computational cost of an exact algorithm. Much of the recent research on MKP investigates improved heuristics.

3.2.2 Empirical studies involving problems with correlated coefficients

Many studies have examined the effect of correlation between objective function and constraint coefficients in synthetic optimization problems on the performance of solution procedures. A common, or “legacy,” aspect of these studies is the test problem generation methods employed, which mimic the test problem generation

methods used in earlier studies, such as those by Martello and Toth (1979, 1981). Martello and Toth (1979, 1988) and Balas and Zemel (1980) study knapsack solution procedures, while Potts and Van Wassenhove (1988, 1992) and John (1989) study solution procedures for scheduling problems. Yet, all use nearly the same test problem generation method and all report significant performance degradation of solution methods which they attribute to higher positive population correlation between objective function and constraint coefficients. Martello and Toth (1981), Fisher, Jaikumar, and Van Wassenhove (1986), Guignard and Rosenwein (1989), Trick (1992), Mazzola and Neebe (1993) and Amini and Racer (1994) report worsening solution procedure performance due to stronger, negative population correlation levels between the objective function and capacity constraint coefficients in the generalized assignment problem (GAP).

Interestingly, the correlation levels induced are not quantified in the studies cited above. Common parameter settings for the generation methods, such as those in Martello and Toth (1979), induce "weak" population correlation above 0.97, while other settings, such as those in Martello and Toth (1981), induce population correlation below -0.97. These generation methods are called "implicit correlation induction" methods by Moore and Reilly (1993) because the correlation levels induced are determined, or implied, by the parameters specified for the problem generation method. Any desired variation in the population correlation requires changing either the parameter settings or the form of the univariate marginal distributions.

The correlation level between two types of coefficients may be explicitly induced and varied across the range of feasible correlation values. Moore and Reilly (1993) use composition to induce a specified population correlation level between objective function coefficients and constraint matrix column sums in weighted set covering problems. Reilly (1991) and Yang (1994) induce correlation in the 0-1 knapsack problem and Pollock (1992) induces correlation in the weighted set covering problem by generating coefficients based on various composite probability mass functions (pmfs). Each of these four studies shows that increasing positive correlation between the objective function and constraint coefficients degrades solution procedure performance. Cario *et al.* (1995) induce various correlation levels between objective function and capacity constraint coefficients in the GAP and find that solution performance degrades with decreasing correlation between the objective function and constraint coefficients. In addition, Cario *et al.* find that GAP instances generated under explicit correlation induction are more challenging than those generated under implicit correlation induction.

3.2.3 Some empirical studies involving MKP

Table 3.1 summarizes the design of various studies of the performance of heuristics for MKP. The current study is shown for comparison purposes. Past studies of MKP heuristics indicate that problem size, the distribution of the constraint coefficients, and the method used to determine the right-hand side coefficients (or constraint slackness) influence heuristic performance.

Table 3.1: Factors and Measures Used in Previous Empirical Studies of MKP Heuristics

Study Authors	Problems Generated	Factors					Measures			
		m	n	S	D	Σ	Tm	Err	OpS	Iter
Toyoda (1975)	904	o	o				o	o		
Loulou & Michaelides (1979)	2250	o	o		o		o	o		
Balas & Martin (1980)	41	o	o			o	o	o		
Pirkul (1987)	230	o	o		o	o	o	o		
Zanakis (1977)	135	o	o	o			o	o		
Fréville & Plateau(1993)	Lots	o	o	o	o		o			o
Fréville & Plateau(1994)	610	o	o	o	o	o	o	o		
Current study	2240			o		o		o	o	
m = number of constraints n = number of decision variables S = slackness of constraints D = distribution of constraint coefficients Σ = population correlation induced between problem coefficients Tm = CPU time required Err = measure of relative error between heuristic and optimal solution value OpS = number of problems solved to optimality Iter = number of iterations										

Pirkul (1987) and Balas and Martin (1980) implicitly induce population correlation between the objective function and constraint coefficients for each variable of approximately 0.66. In addition, they induce correlations of about 0.43 between the coefficients in every pair of constraints. Generally, the performance of the heuristics worsens as the objective function-constraint correlations increase. However, there is no discussion of how their results might be influenced by the interconstraint correlations. Fréville & Plateau(1994) generate objective function coefficients and right-hand side values as functions of independently generated constraint coefficients. They conclude that independent problems are easier to solve than the problems with correlation.

3.3 The Test Problem Generation Methods

Two multivariate correlation induction methods, each associated with a different correlation measure, are used in this study. Both methods require the user to specify the univariate marginal distributions and the correlation structure. The composition method discussed in Hill (Chapter 2) is used to induce specified Pearson product-moment population correlation structures, and the method presented in Iman and Conover (1982) is used to approximately induce specified sample Spearman rank correlation structures. Both methods are used to generate values of trivariate random variables to represent the coefficients (c_j, a_{1j}, a_{2j}) for each variable x_j in 2KP.

3.3.1 Pearson product-moment-based correlation induction method

The Pearson product-moment correlation coefficient is a measure of the linear dependence between two random variables. Hill (Chapter 2) shows how to construct a multivariate composite distribution based on a specified Pearson product-moment population correlation structure, using the 2^{k-1} extreme-correlation distributions and the joint distribution under independence. For a k -variate random variable, $\mathbf{Y}$, each extreme-correlation distribution is a joint distribution for which each correlation term is at either the extreme positive or extreme negative level. The 2^{k-1} extreme-correlation distributions are denoted $h_\ell(\mathbf{y})$, $\ell = 1, 2, \dots, 2^{k-1}$, and the joint distribution under independence is denoted $h_0(\mathbf{y})$.

Chapter 2 provides formulas for computing composition probabilities, $\lambda_\ell, \ell = 0, 1, 2, 3, 4$, based on a specified correlation structure for a trivariate random variable. A joint distribution with the specified correlation structure is the composite distribution

$$h(\mathbf{y}) = \sum_{\ell=0}^4 \lambda_\ell h_\ell(\mathbf{y}). \quad (3.4)$$

The value of $\lambda_\ell, \ell = 0, 1, 2, 3, 4$, represents the relative frequency of sampling based on $h_\ell(\mathbf{y})$. When λ_0 is at its minimum value, the composite distribution is called a Type L distribution. When λ_0 is at its maximum value, the composite distribution is called a Type U distribution. The Type L and Type U distributions define a range of composite distributions with a specified population correlation structure.

3.3.2 Spearman rank correlation-based correlation induction method

The Spearman rank correlation coefficient is a measure of the monotonic dependency between two random variables. Let $\mathbf{M}$ be a specified correlation matrix. The method of Iman and Conover (1982) may be used to induce a Spearman rank correlation structure, given by $\mathbf{M}$, among a set of random variables.

Suppose n observations of k random variables with correlation structure $\mathbf{M}$ is required. First generate two matrices, $\mathbf{R}$ and $\mathbf{V}$ such that $\mathbf{R}$ is an $(n \times k)$ matrix of van der Waerden scores, randomized within each of the k columns, and $\mathbf{V}$ is an $(n \times k)$ matrix of n independent observations of each of the k random variables. Consider each column of $\mathbf{R}$ as n observations of k random variables and compute $\mathbf{T}$, the corresponding sample rank correlation matrix. Compute the

Choleski factorizations $\mathbf{A}$ and $\mathbf{Q}$ such that $\mathbf{T} = \mathbf{A}\mathbf{A}'$ and $\mathbf{M} = \mathbf{Q}\mathbf{Q}'$. Compute

$$\mathbf{S} = \mathbf{R}(\mathbf{A}\mathbf{Q}^{-1})', \quad (3.5)$$

which is a transformed matrix of scores. The k columns of n values in $\mathbf{S}$ have a sample rank correlation structure that approximates $\mathbf{M}$. The entries in each column of $\mathbf{V}$ are reordered so that their rankings are the same as the rankings in the corresponding columns of $\mathbf{S}$. The sample Spearman rank correlation structure of the shuffled matrix of observations, $\mathbf{V}$, approximates the specified correlation structure, $\mathbf{M}$.

This method is applicable for any marginal distributions. However, since this method involves computing Choleski factorizations, matrix inverses, and ranking of the data, the method becomes more computationally intensive as k or n increases.

3.4 The Experiment Design and Analysis Methods

The goal of this study is to gain a deeper understanding of how test problem generation methods influence the performance of solution procedures. This is not an advocacy study for a particular solution procedure or an experiment on state-of-the-art solution methods for MKP (2KP). Rather, this is an investigation into how the performance of representative techniques, the branch-and-bound code CPLEX and the heuristic by Toyoda (1975), are affected by the correlation structure between the coefficient types, and by other test problem characteristics.

Let $A^1 \sim U\{1, 2, \dots, 40\}$ be the random variable representing the values of the coefficients in the first constraint, $A^2 \sim U\{1, 2, \dots, 15\}$ be the random variable representing the values of the coefficients in the second constraint, and $C \sim$

$U\{1, 2, \dots, 100\}$ be the random variable representing the values of the objective function coefficients in 2KP. Different distributions for A^1 and A^2 virtually guarantee that both constraints in each 2KP instance are different. Suppose the distributions of A^1 and A^2 are identical. Then $\rho_{A^1 A^2} \in [-1, 1]$, and for each 2KP instance generated with $\rho_{A^1 A^2} = 1$, the coefficients for each variable would be identical in both constraints.

The three correlation terms in the correlation structure of 2KP are ρ_{CA^1} , ρ_{CA^2} , and $\rho_{A^1 A^2}$ with $\boldsymbol{\rho} = (\rho_{CA^1}, \rho_{CA^2}, \rho_{A^1 A^2})$. When referring to a particular correlation measure, $\mathbf{P} = (\rho_{CA^1}^P, \rho_{CA^2}^P, \rho_{A^1 A^2}^P)$ denotes the Pearson product-moment correlation structure while a Spearman rank correlation structure is denoted $\mathbf{S} = (\rho_{CA^1}^S, \rho_{CA^2}^S, \rho_{A^1 A^2}^S)$.

3.4.1 Definition of the experiment design settings

Three problem generation parameters are varied in this experiment: the correlation structure between the sets of problem coefficients, the constraint slackness, and the correlation measure (Pearson or Spearman). It is well established that problem size influences solution procedure performance, so problem size is held constant with two constraints (i.e., the 2KP) and 100 variables.

For the marginal distributions previously defined for C , A^1 , and A^2 , the ranges of feasible Pearson correlation levels for each correlation term are:

$$\rho_{CA^1}^P \in [-0.99997, 0.99997] \quad (3.6)$$

$$\rho_{CA^2}^P \in [-0.99773, 0.99773] \quad (3.7)$$

$$\text{and } \rho_{A^1 A^2}^P \in [-0.99752, 0.99752]. \quad (3.8)$$

The population correlation structures are varied by systematically varying each correlation term. Five equally-spaced correlation values across the feasible range for each correlation term yields 125 potential correlation structures. However, 80 of these combinations yield would-be correlation matrices that are not positive semi-definite. Table 3.2 lists the 45 feasible correlation structures; Figure 3.1 is a 3-dimensional plot of the feasible correlation structures. For each feasible correlation structure, there is a composite joint distribution (3.4). For the Pearson correlation induction method, the joint distribution for 34 of these 45 feasible correlation structures is expressible only as a Type L composite distribution with $\lambda_0 = 0$. For each correlation structure in Table 3.2 marked with a $\bullet$, there are composite distributions with $\lambda_0 > 0$ in addition to a Type L distribution. For these correlation structures, both the Type L and Type U forms of the composite distribution are used in the experiment.

A “slackness” measure for constraint i , S_i , is defined as the ratio of the right-hand side coefficient in constraint i to the sum of the coefficients in that constraint. Low slackness values give “tight” constraints and high slackness values give “loose” constraints. Constraints are “mixed” if both low and high slackness values are specified for the same test problem. Table 3.3 summarizes the slackness levels used in the studies cited previously. Two levels of slackness are examined in this study: $S_i = 0.30, 0.70$, $i = 1, 2$. Each of the four possible settings of S_1 and S_2 is referred to as a constraint slackness setting. Since the marginal distribution of A^1 differs from the marginal distribution of A^2 , then $(S_1, S_2) = (0.30, 0.70)$ is considered to be a different slackness setting than $(S_1, S_2) = (0.70, 0.30)$.

Table 3.2: Experiment Design Correlation Structures

Number	Correlation Values			Number	Correlation Values		
	ρ_{CA^1}	ρ_{CA^2}	$\rho_{A^1A^2}$		ρ_{CA^1}	ρ_{CA^2}	$\rho_{A^1A^2}$
1	0.99997	0.99773	0.99752	24 •	-0.49999	0.00000	0.00000
2	0.49999	0.49887	0.99752	25	-0.99997	0.00000	0.00000
3	0.00000	0.00000	0.99752	26	0.49999	-0.49887	0.00000
4	-0.49999	-0.49887	0.99752	27 •	0.00000	-0.49887	0.00000
5	-0.99997	-0.99773	0.99752	28	-0.49999	-0.49887	0.00000
6	0.49999	0.99773	0.49876	29	0.00000	-0.99773	0.00000
7	0.99997	0.49887	0.49876	30	-0.49999	0.99773	-0.49876
8 •	0.49999	0.49887	0.49876	31	0.00000	0.49887	-0.49876
9	0.00000	0.49887	0.49876	32 •	-0.49999	0.49887	-0.49876
10	0.49999	0.00000	0.49876	33	-0.99997	0.49887	-0.49876
11 •	0.00000	0.00000	0.49876	34	0.49999	0.00000	-0.49876
12	-0.49999	0.00000	0.49876	35 •	0.00000	0.00000	-0.49876
13	0.00000	-0.49887	0.49876	36	-0.49999	0.00000	-0.49876
14 •	-0.49999	-0.49887	0.49876	37	0.99997	-0.49887	-0.49876
15	-0.99997	-0.49887	0.49876	38 •	0.49999	-0.49887	-0.49876
16	-0.49999	-0.99773	0.49876	39	0.00000	-0.49887	-0.49876
17	0.00000	0.99773	0.00000	40	0.49999	-0.99773	-0.49876
18	0.49999	0.49887	0.00000	41	-0.99997	0.99773	-0.99752
19 •	0.00000	0.49887	0.00000	42	-0.49999	0.49887	-0.99752
20	-0.49999	0.49887	0.00000	43	0.00000	0.00000	-0.99752
21	0.99997	0.00000	0.00000	44	0.49999	-0.49887	-0.99752
22 •	0.49999	0.00000	0.00000	45	0.99997	-0.99773	-0.99752
23 •	0.00000	0.00000	0.00000				

Table 3.3: Slackness Settings From Previous MKP studies

Study Authors	Slackness Setting S_i
Toyoda	$S_i = 0.67$
Loulou & Michaelides	$b_i = 1 \forall i = 1, 2, \dots, m$
Balas & Martin	$S_i \sim U(0.5, 0.9)$
Pirkul	$S_i = 0.50$
Zanakis	$S_i = 0.30, 0.50, 0.90$
Fréville & Plateau(1993)	$S_i = 0.25, 0.50, 0.75$
Fréville & Plateau(1994)	$S_i = 0.25, 0.50, 0.75$
Current study	$S_i = 0.30, 0.70$

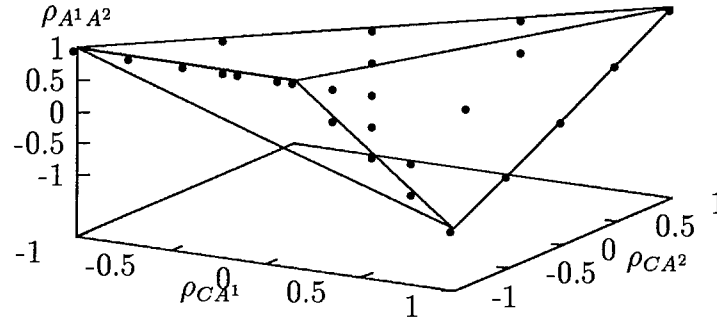


Figure 3.1: Three Dimensional Plot of Experiment Design Correlation Structures

One purpose for this study is to examine how the correlation measure (Pearson or Spearman) used affects solution procedure performance. For each specified correlation structure and constraint slackness setting combination, five test problems are generated using both the Pearson correlation induction method (i.e., composition) and the Spearman correlation induction method (i.e., Iman and Conover's method). Random numbers are not synchronized since the shuffling involved in Iman and Conover's method would undermine any synchronization.

Each combination of correlation structure, constraint slackness setting, and correlation measure forms an experiment design point. With 45 correlation structures, four constraint slackness settings, two correlation measures, and five replications

for each design point, 1800 optimization test problems are generated. Additionally, the 11 correlation structures in Table 3.2 that may be represented by a Type U composite distribution provide an additional 440 test problems, for a total of 2240 2KP test problems generated for this study.

A representative algorithm and heuristic were chosen based on availability and general acceptance of the procedures. CPLEX from CPLEX Optimization, Inc., is contained in many commercially available packages and is available as a standalone product (see review by Saltzman, 1994). The mixed-integer optimizer in CPLEX, version 2.1, was selected and utilized in a depth-first search, branch-and-bound mode. Many of the studies involving heuristics for MKP benchmark to Toyoda's (1975) heuristic. Hence, Toyoda's heuristic was chosen as a representative heuristic. Hereafter, these procedures are referred to as CPLEX and TOYODA, respectively.

There are a variety of performance measures for assessing solution procedure effectiveness and efficiency. Typical performance measures for branch-and-bound procedures include CPU time, iteration count, or the number of nodes enumerated in the branch-and-bound tree. The three measures are clearly related to one another. The number of nodes is used in this study and is referred to as NODES. Typical measures for heuristics include CPU time, iteration count, or relative error. This study uses the relative error denoted as REL, where

$$\text{REL} = 100 \times \frac{(Z_{IP} - Z_H)}{Z_{IP}}, \quad (3.9)$$

where Z_H is the heuristic solution value and Z_{IP} is the optimal (or best known) integer solution value for the 2KP.

The size of the LP-IP gap in an optimization problem is often viewed as a factor influencing the performance of solution procedures (Chang and Shepardson, 1982). In this study, the influence of the factors in the experiment on the size of the LP-IP gap is briefly examined.

3.4.2 Methods for analyzing results

Two non-parametric statistical tests, the sign test and the Kruskal-Wallis (KW) test, are used to analyze the data from the experiment. These tests are summarized below; additional details are found in Conover (1980).

The sign test is useful for establishing whether observations from one population tend to differ in magnitude when compared to observations from another population. Let $X_i^{(1)}$ and $X_i^{(2)}$, $i = 1, 2, \dots, n$, be n observations from two populations paired in some logical fashion, and let $d_i = X_i^{(1)} - X_i^{(2)}$, $i = 1, 2, \dots, n$, be the differences between the observations. If there is no difference in magnitude between the populations, the probability of a positive sign on each d_i follows the binomial distribution with $p = 0.5$. Therefore, the null and alternative hypotheses are:

$$H_0 : \Pr(+) = \Pr(-)$$

$$H_1 : \Pr(+) \neq \Pr(-)$$

where $\Pr(+)$ and $\Pr(-)$ are, respectively, the probabilities of positive and negative signs on each d_i . The test statistic, T_1 , is the total number of positive d_i s, ignoring ties. The decision rule is to reject H_0 at the α level of significance if the probability

of observing T_1 positive d_i s under a true null hypothesis is less than α . The primary use of the sign test in this study is to test whether there is a difference in solution procedure performance due to the correlation measure.

The KW test is a rank test for differences among the means in m populations. Let $X_i^{(j)}, i = 1, 2, \dots, n_j, j = 1, 2, \dots, m$, be the i^{th} observation from the j^{th} population and $R(X_{(i)}^j), i = 1, 2, \dots, n_j, j = 1, 2, \dots, m$, be the overall rank of each observation among all $N = \sum_{j=1}^m n_j$ observations. The null and alternative hypotheses are:

- H_0 : All m population distribution functions have identical means,
 H_1 : The m population distribution functions do not have identical means.

Define $R_j = \sum_{i=1}^{n_j} R(X_{(i)}^j)$, $j = 1, 2, \dots, m$. The test statistic T_2 for the KW test is

$$T_2 = \frac{1}{S^2} \left(\sum_{j=1}^m \frac{R_j^2}{n_j} - \frac{N(N+1)^2}{4} \right) \quad (3.10)$$

where

$$S^2 = \frac{1}{N-1} \left(\sum_{j=1}^m \sum_{i=1}^{n_j} R(X_{(i)}^j)^2 - \frac{N(N+1)^2}{4} \right). \quad (3.11)$$

The decision rule is to reject H_0 at the α level of significance if T_2 exceeds the $1 - \alpha$ quantile of the chi-square distribution with $m - 1$ degrees of freedom. The KW test is used in this study to test whether or not there are solution procedure performance differences due to either the correlation structure or the constraint slackness setting.

In addition to the sign and KW tests, regression models are constructed to quantify the relationships between the experiment design parameters and each performance measure. The models constructed based on the 2KP experiment were developed using a stepwise regression procedure to obtain the regression model that maximizes the value of the coefficient of determination, R^2 .

3.5 Comparing Samples From The Correlation Induction Methods

One motivation for this study is to examine whether solution procedure performance is affected by differences in the form of the underlying multivariate distribution of coefficient values. A simple experiment with a bivariate random variable provides some insight into the differences in the form of the underlying distribution associated with each correlation measure. (The correlation structure also affects the form of a joint distribution, even for a fixed correlation measure.) Assume that Y_1 and Y_2 are each discrete uniform random variables where $Y_1 \sim U(1, 2, \dots, 20)$, $Y_2 \sim U(1, 2, \dots, 10)$, and $\text{Corr}(Y_1, Y_2) = 0.49876$. The sample joint distributions that result from 100,000 observations from the generation method for each correlation measure are shown in Figures 3.2 and 3.3. To simplify the generation of data based on the Spearman measure, the 100,000 observations came from 1000 replications of 100 observations each.

Figure 3.2 contains a sample pmf (multiplied by 10000) for Pearson product-moment correlation induction using a Type U distribution. There is a minimal probability in each cell and a concentration of probability along the upper left to

lower right “diagonal” of cells. This concentration of probability is characteristic of compositions involving extreme-correlation distributions (Devroye, 1986). The concentration of probability along the diagonal is minimized (maximized) and the minimum probability in each cell is maximized (minimized) with Type U (Type L) distributions (Hill and Reilly, 1994). Suppose 100,000 observations of (Y_1, Y_2) were generated based on a Type L distribution. Then, all of the observations would be concentrated on the upper left to lower right and lower left to upper right diagonals. The cells off these diagonals would have probability zero.

Figure 3.3 contains a sample pmf (multiplied by 10000) for Spearman rank correlation induction. There is a wide band of probability along the main “diagonal” discussed for Figure 3.2; several cells have very small probability. Generally, the probability in the cells drops off as one looks at cells further and further away from the main diagonal.

For each of the 2240 problems generated for this study sample Pearson product-moment and Spearman rank correlation values were computed. Table 3.4 summarizes these sample correlations by correlation induction method and by the value specified for each correlation term. The data in Table 3.4 indicates that both correlation induction methods effectively generate data with the specified correlation structure. The standard errors of the average sample correlation values under the Spearman induction method are generally smaller than the corresponding standard errors under the Pearson induction method. This phenomenon may be explained by recognizing the different approach to inducing correlation that these two variate-

Y_1	Y_2									
	1	2	3	4	5	6	7	8	9	10
1	276	23	26	25	26	23	24	26	25	24
2	279	22	24	25	24	26	25	24	25	23
3	25	279	23	25	24	24	29	24	24	28
4	25	285	23	23	26	26	26	24	27	27
5	25	27	275	27	25	28	23	24	23	25
6	23	27	264	26	26	24	25	22	27	25
7	25	27	25	268	27	26	24	25	25	23
8	26	24	24	268	26	25	25	23	23	23
9	26	25	23	26	275	24	27	27	27	27
10	25	24	26	22	278	23	27	23	26	23
11	23	23	26	24	25	273	27	26	22	24
12	26	26	23	26	26	277	25	28	23	24
13	22	27	26	26	26	25	272	24	26	26
14	26	27	23	26	25	26	277	27	24	26
15	28	25	26	22	23	27	27	276	24	25
16	26	26	23	25	27	23	22	277	23	25
17	24	25	26	24	21	26	25	26	274	26
18	27	24	27	28	27	23	23	27	274	23
19	28	24	27	25	26	26	22	25	24	282
20	25	26	28	27	24	23	26	26	24	271
Note: Proportions multiplied by 10000										

Figure 3.2: Sample Distributional Form From Pearson Induction Method,
 $\rho = 0.49876$

Y_1	Y_2									
	1	2	3	4	5	6	7	8	9	10
1	239	105	24	17	24	38	48	2	1	1
2	78	102	72	46	61	61	55	12	7	1
3	93	88	94	31	24	22	18	54	68	4
4	77	26	42	68	44	28	17	92	89	16
5	50	12	25	114	109	60	18	53	49	16
6	72	29	33	123	89	58	15	38	19	7
7	110	47	49	96	53	56	36	42	7	7
8	98	33	54	66	32	75	68	52	7	25
9	56	37	57	61	29	57	84	73	9	46
10	42	55	38	45	44	38	84	81	18	53
11	28	77	25	32	61	35	84	73	30	53
12	11	80	28	44	71	40	60	54	45	60
13	8	68	67	77	86	37	39	32	46	46
14	13	61	114	84	67	27	26	35	49	21
15	11	52	125	36	39	38	45	58	85	15
16	3	26	77	15	35	78	50	55	118	33
17	3	21	18	4	52	86	55	54	155	54
18	8	48	10	4	29	74	96	80	129	38
19	2	17	48	19	32	84	74	47	59	118
20	1	1	15	6	5	14	39	13	10	391
Note: Proportions multiplied by 10000										

Figure 3.3: Sample Distributional Form From Spearman Induction Method,
 $\rho = 0.49876$

Table 3.4: Sample Correlations by Target and Induction Type

	Target Correlation Value	Number of Problems	Induction Method			
			Pearson		Spearman	
			Mean	Std Error	Mean	Std Error
ρ_{CA^1}	0.99997	100	0.98968	0.00007	0.99914	0.00002
	0.49999	280	0.49524	0.00645	0.47747	0.00102
	0.00000	360	-0.00328	0.00668	0.00068	0.00078
	-0.49999	280	-0.49764	0.00656	-0.47494	0.00107
	-0.99997	100	-0.98967	0.00005	-0.99737	0.00004
ρ_{CA^2}	0.99773	100	0.98780	0.00003	0.99436	0.00009
	0.49887	280	0.48964	0.00658	0.47724	0.00123
	0.00000	360	-0.01640	0.00679	0.00350	0.00092
	-0.49887	280	-0.49967	0.00625	-0.47052	0.00122
	-0.99773	100	-0.98775	0.00003	-0.98762	0.00014
$\rho_{A^1A^2}$	0.99752	100	0.98756	0.00003	0.99318	0.00010
	0.49876	280	0.48105	0.00614	0.47468	0.00140
	0.00000	360	-0.00254	0.00693	0.00515	0.00110
	-0.49876	280	-0.49975	0.00626	-0.46844	0.00138
	-0.99752	100	-0.98762	0.00003	-0.98564	0.00017

generation methods use. The Pearson induction method samples from a composite distribution with a specified Pearson product-moment *population* correlation structure, while the Spearman induction method (Iman and Conover, 1982) targets a specified *sample* rank correlation structure.

Table 3.5 summarizes Spearman sample correlations for the 2KP coefficients generated with the Pearson method, and vice versa. Consider the mean values reported in Table 3.5. The Pearson method effectively generates data with the specified Spearman rank correlation structure. However, the Spearman method is less effective at generating data with a specified Pearson product-moment correlation structure. In both cases, standard errors are small, and better for the problems generated with the Spearman method.

Table 3.5: Sample Correlations by Target, Method, and Alternate Measure

	Target Correlation Value	Number of Problems	Pearson Method		Spearman Method	
			Spearman Value		Pearson Value	
			Mean	Std Error	Mean	Std Error
ρ_{CA^1}	0.99997	100	0.99943	0.00010	0.98218	0.00046
	0.49999	280	0.50093	0.00651	0.46857	0.00120
	0.00000	360	-0.00311	0.00669	-0.00047	0.00100
	-0.49999	280	-0.50085	0.00656	-0.46815	0.00130
	-0.99997	100	-0.99773	0.00008	-0.98219	0.00035
ρ_{CA^2}	0.99773	100	0.99698	0.00005	0.97821	0.00047
	0.49887	280	0.49608	0.00667	0.46831	0.00140
	0.00000	360	-0.01349	0.00680	-0.00002	0.00108
	-0.49887	280	-0.49948	0.00624	-0.46805	0.00138
	-0.99773	100	-0.99039	0.00011	-0.97792	0.00053
$\rho_{A^1A^2}$	0.99752	100	0.99692	0.00003	0.97691	0.00050
	0.49876	280	0.48805	0.00618	0.46546	0.00146
	0.00000	360	0.00172	0.00690	0.00135	0.00119
	-0.49876	280	-0.49782	0.00629	-0.46719	0.00149
	-0.99752	100	-0.98949	0.00011	-0.97687	0.00047

Table 3.6: Performance Measure Averages by Correlation Measure

Performance Measure	Mean Performance Measure	
	Pearson	Spearman
NODES	2337.50	3748.83
REL	0.77	1.50

Table 3.7: Results of Sign Test on Performance Measures

Performance Measure	Total $d_i \neq 0$	Total $d_i > 0$	Acceptance Region	p -value
NODES	1111	354	(522,588)	< 0.0001
REL	1095	759	(515,580)	< 0.0001

3.6 Influence of Population Correlation Measure

Table 3.6 summarizes the results for each performance measure by population correlation measure. The test problems generated based on Spearman rank correlation require more branch-and-bound nodes with CPLEX and have larger relative errors with the TOYODA heuristic than the problems generated based on the Pearson correlation induction method.

Suppose the data are separated by correlation measure (i.e., by induction method) and then paired by design point and replication number to develop the vector of differences used in a sign test. Table 3.7 provides the sign test results for each performance measure and the $\alpha = 0.05$ acceptance regions. These results indicate that test problems based on the Spearman correlation measure require more NODES by CPLEX and have a larger REL for TOYODA.

A sign test was applied to the data associated with each of the 45 population correlation structures listed in Table 3.2. Table 3.8 provides the p -values associated with each of these 45 tests. An asterisk (*) indicates those tests with p -values below 0.05, 28 tests for NODES data and 21 for REL data. For all tests significant at the $\alpha = 0.05$ level, problems based on the Spearman induction method required either more NODES or REL was higher. Within the REL column of Table 3.8 the asterisks tend to be associated with ρ_{CA1} and ρ_{CA2} values above 0.4, while there is no such obvious pattern to the asterisks within the NODES column.

Table 3.9 provides sign test results by Type L and Type U distributions for the eleven correlation structures permitting both types of composite distributions. An asterisk (*) indicates those cases where there is a significant difference in performance between problems generated based on the Pearson measure and problems based on the Spearman measure for an $\alpha = 0.05$ significance level. Test problems based on the Pearson measure with a Type L distribution are more likely to produce results different from their Spearman measure counterparts (i.e., less NODES, larger REL value) than those test problems with a Type U distribution. This phenomenon is observed because a Type U distribution more closely resembles the underlying distribution for the Spearman correlation induction method than does a Type L distribution.

asterisk (*) highlights those tests with p -values

Table 3.8: Sign Test Results On Each Correlation Structure

Correlation Values			NODES	REL
ρ_{CA^1}	ρ_{CA^2}	$\rho_{A^1A^2}$	$p\text{-value}^\dagger$	$p\text{-value}^\dagger$
0.99997	0.99773	0.99752	0.0835	0.0059 *
0.49999	0.49887	0.99752	0.0059 *	0.0577
0.00000	0.00000	0.99752	0.2403	0.6762
-0.49999	-0.49887	0.99752	0.0013 *	0.4073
-0.99997	-0.99773	0.99752	0.0577	0.1250
0.49999	0.99773	0.49876	0.1796	0.0013 *
0.99997	0.49887	0.49876	< 0.0001 *	< 0.0001 *
0.49999	0.49887	0.49876	0.0011 *	0.2148
0.00000	0.49887	0.49876	0.0207 *	0.1316
0.49999	0.00000	0.49876	0.0059 *	0.0013 *
0.00000	0.00000	0.49876	0.0011 *	0.9459
-0.49999	0.00000	0.49876	0.4119	0.4119
0.00000	-0.49887	0.49876	0.0002 *	0.5881
-0.49999	-0.49887	0.49876	0.6821	0.0541
-0.99997	-0.49887	0.49876	0.4119	0.0245 *
-0.49999	-0.99773	0.49876	0.1316	0.0577
0.00000	0.99773	0.00000	0.0835	0.0002 *
0.49999	0.49887	0.00000	0.0013 *	0.1316
0.00000	0.49887	0.00000	< 0.0001 *	0.0003 *
-0.49999	0.49887	0.00000	0.0207 *	0.0013 *
0.99997	0.00000	0.00000	0.1316	0.0002 *
0.49999	0.00000	0.00000	< 0.0001 *	0.0083 *
0.00000	0.00000	0.00000	0.0192 *	0.4373
-0.49999	0.00000	0.00000	0.3746	0.0769
-0.99997	0.00000	0.00000	0.0022 *	0.1316
0.49999	-0.49887	0.00000	0.1316	0.0059 *
0.00000	-0.49887	0.00000	0.0192 *	0.5627
-0.49999	-0.49887	0.00000	0.9423	0.2517
0.00000	-0.99773	0.00000	< 0.0001 *	0.2517
-0.49999	0.99773	-0.49876	0.0096 *	0.0577
0.00000	0.49887	-0.49876	0.0013 *	0.0002 *
-0.49999	0.49887	-0.49876	0.0403 *	0.0001 *
-0.99997	0.49887	-0.49876	0.0577	0.0059 *
0.49999	0.00000	-0.49876	0.0002 *	0.0002 *
0.00000	0.00000	-0.49876	0.0192 *	0.5627
-0.49999	0.00000	-0.49876	0.0577 *	0.1316
0.99997	-0.49887	-0.49876	0.0013 *	0.0013 *
0.49999	-0.49887	-0.49876	0.0032 *	< 0.0001 *
0.00000	-0.49887	-0.49876	0.8684	0.5881
0.49999	-0.99773	-0.49876	0.0002 *	0.0002 *
-0.99997	0.99773	-0.99752	0.1316	0.0059 *
-0.49999	0.49887	-0.99752	0.0002 *	< 0.0001 *
0.00000	0.00000	-0.99752	0.1316	0.4119
0.49999	-0.49887	-0.99752	0.0002 *	0.0013 *
0.99997	-0.99773	-0.99752	0.0004 *	0.1316

† Null hypothesis on no difference

Table 3.9: Sign Test Results For Type L Versus Type U Distributions

Correlation Values			<i>p</i> -values					
			NODES			REL		
			Type L	Type U		Type L	Type U	
ρ_{CA^1}	ρ_{CA^2}	$\rho_{A^1A^2}$						
0.49999	0.49887	0.49876	0.9940	*	0.9793	*	0.0577	0.7483
0.49999	0	0	0.9987	*	0.9998	*	0.0002	* 0.0588
0.49999	-0.49887	-0.49876	0.9940	*	0.9423		0.0000	* 0.0577
0	0.49887	0	0.9987	*	0.9940	*	0.0013	* 0.0577
0	0	0.49876	0.9999	*	0.7483		0.6762	0.9793 *
0	0	0	0.9940	*	0.7483		0.0577	0.9423
0	0	-0.49876	0.9423		0.9423		0.4783	0.4119
0	-0.49887	0	0.9940	*	0.7483		0.4119	0.7483
-0.49999	0.49887	-0.49876	0.9987	*	0.4119		0.0000	* 0.1316
-0.49999	0	0	0.6762		0.5881		0.0207	* 0.5881
-0.49999	-0.49887	0.49876	0.8684		0.0835		0.5000	0.0207 *

Based on the results in this study, it appears that the choice of correlation measure influences the performance of solution procedures on synthetic test problems. The next issue is whether the population correlation structure and constraint slackness settings, for test problems generated based on each type of correlation measure, influence solution procedure performance. CPLEX performance is examined first, followed by similar analyses of TOYODA performance.

3.7 Analysis of CPLEX performance

In this section, the influence of the population correlation structure and constraint slackness settings on CPLEX performance is examined. Two regression models are constructed to summarize the effects of correlation structure and constraint slackness on CPLEX performance.

3.7.1 Correlation structure influence

Tables 3.10 and 3.11 summarize CPLEX results for Pearson and Spearman problems, respectively. An artificial limit of 250,000 NODES was imposed on CPLEX processing. The rightmost columns in Tables 3.10 and 3.11 indicate how many, if any, problems were not solved to optimality in this study due to this limit. The number of CPLEX NODES varies greatly as the population correlation structure changes.

A KW test was conducted on the data grouped by population correlation structure to test for a difference in average NODES due to correlation structure. The KW test statistics of 180.95 for Pearson correlation problems and 135.59 for Spearman correlation problems equate to p -values of less than 1.0×10^{-10} for each test. Clearly, there is a CPLEX performance difference due to correlation structure.

Independent sampling is represented by $\rho = (0, 0, 0)$, Type U in Table 3.10. Independent sampling is a generally accepted method of generating test problems, however, these results suggest that generating test problems with independent sampling only provides little information about the full range of test problem difficulty that is observed with a more systematic problem generation scheme involving correlation induction. In fact, independent sampling seems to provide information only about median performance. To appreciate the CPLEX performance variation possible with different correlation structures, one need only scan down the columns of Tables 3.10 and 3.11 and notice the range in average NODES, the corresponding standard errors, and the drastic difference in average NODES between the first

Table 3.10: CPLEX Results using Pearson Correlation Induction

Type	Correlation Values			Mean NODES	Standard Error	Not Solved
	ρ_{CA1}^P	ρ_{CA2}^P	ρ_{A1A2}^P			
L	0	-0.49887	-0.49876	30284.1	12347.01	2
L	0.99997	-0.99773	-0.99752	27947.5	11937.36	1
L	-0.99997	0.99773	-0.99752	25250.4	12153.85	1
L	-0.99997	-0.99773	0.99752	21507.3	8969.49	1
U	0.49999	-0.49887	-0.49876	3749.9	2822.78	
L	0	0	-0.99752	2206.5	954.87	
L	-0.99997	-0.49887	0.49876	2051.4	1166.00	
L	-0.99997	0.49887	-0.49876	1849.1	703.87	
L	-0.49999	0	-0.49876	1752.7	770.87	
L	0	0	-0.49876	1534.2	923.33	
U	0.49999	0.49887	0.49876	1204.2	629.53	
L	0	-0.99773	0	1105.6	622.93	
U	0	-0.49887	0	659.7	210.35	
U	-0.49999	0.49887	-0.49876	641.2	197.34	
L	-0.49999	-0.99773	0.49876	624.8	165.17	
U	0	0	-0.49876	622.3	161.56	
L	-0.49999	-0.49887	0	589.1	127.79	
U	0	0	0	551.9	128.34	
L	-0.49999	-0.49887	0.49876	541.1	210.68	
L	0.49999	-0.99773	-0.49876	535.7	168.92	
U	-0.49999	-0.49887	0.49876	522.9	142.40	
L	0	0.49887	-0.49876	365.6	115.00	
L	-0.99997	0	0	354.8	111.54	
L	0.49999	0	-0.49876	325.9	69.87	
U	-0.49999	0	0	292.2	85.33	
U	0.49999	0	0	275.1	85.93	
L	0.49999	-0.49887	-0.99752	257.9	94.70	
L	0.49999	-0.49887	-0.49876	222.8	146.88	
L	-0.49999	0	0	220.9	83.20	
L	0	-0.49887	0	210.3	60.69	
U	0	0	0.49876	206.6	51.84	
L	0.49999	0.49887	0.49876	158.3	43.76	
L	0	0	0.49876	157.7	36.82	
L	0.49999	0.49887	0	137.3	26.89	
L	0	0	0	131.5	36.91	
U	0	0.49887	0	129.8	36.86	
L	0.49999	-0.49887	0	129.1	23.82	
L	-0.49999	0.99773	-0.49876	125.7	28.57	
L	0.49999	0	0	122.4	44.82	
L	0.99997	0	0	114.2	26.20	
L	-0.49999	0.49887	-0.99752	113.2	24.71	
L	-0.49999	0	0.49876	110.6	15.08	
L	0	0	0.99752	109.1	19.74	
L	0	0.49887	0	94.2	27.90	
L	-0.49999	0.49887	-0.49876	83.5	36.19	
L	-0.49999	-0.49887	0.99752	82.6	15.22	
L	0.99997	0.99773	0.99752	75.9	18.20	
L	-0.49999	0.49887	0	72.9	13.58	
L	0.49999	0.99773	0.49876	72.8	17.72	
L	0	0.99773	0	66.8	17.62	
L	0.99997	0.49887	0.49876	65.3	18.33	
L	0	-0.49887	0.49876	65.3	13.57	
L	0.49999	0.49887	0.99752	62.3	14.30	
L	0	0.49887	0.49876	58.9	14.36	
L	0.49999	0	0.49876	56.9	12.69	
L	0.99997	-0.49887	-0.49876	40.7	15.65	

Table 3.11: CPLEX Results using Spearman Correlation Induction

Correlation Values			Mean	Standard	Not
$\rho_{CA^1}^S$	$\rho_{CA^2}^S$	$\rho_{A^1A^2}^S$	NODES	Error	Solved
0.99997	-0.99773	-0.99752	42945.3	14333.93	3
-0.99997	0.99773	-0.99752	38482.7	14423.24	3
-0.99997	0	0	33519.9	12393.77	2
0	0	-0.99752	25838.9	10735.01	1
-0.99997	-0.99773	0.99752	17667.7	3934.72	
-0.99997	-0.49887	0.49876	17310.8	8920.90	1
-0.49999	-0.99773	0.49876	4032.9	2320.58	
0.49999	-0.49887	-0.99752	2754.1	1501.03	
-0.49999	0.49887	-0.99752	2368.1	781.46	
0.49999	-0.99773	-0.49876	1977.2	648.52	
0	-0.99773	0	1755.6	728.72	
-0.99997	0.49887	-0.49876	1441.0	676.44	
0.49999	0	-0.49876	1203.3	296.88	
0.99997	-0.49887	-0.49876	989.5	214.62	
0	0.49887	-0.49876	871.2	134.87	
0	0	-0.49876	797.0	110.24	
0.49999	-0.49887	-0.49876	660.7	117.69	
-0.49999	0	-0.49876	591.6	123.42	
0.49999	0.49887	0	545.3	88.25	
0.49999	0	0	544.9	59.64	
0	0	0	510.2	58.84	
0.49999	-0.49887	0	481.8	83.92	
0.99997	0	0	477.1	95.03	
-0.49999	0.99773	-0.49876	474.9	87.02	
0	-0.49887	0	473.3	48.62	
0	0.49887	0	446.8	53.61	
-0.49999	0.49887	-0.49876	428.0	56.26	
0	-0.49887	-0.49876	416.0	69.71	
0.49999	0.49887	0.49876	407.7	64.65	
-0.49999	0	0	370.9	52.23	
0	0	0.49876	367.7	47.56	
0.99997	0.49887	0.49876	355.9	57.60	
0	-0.49887	0.49876	352.3	49.64	
-0.49999	-0.49887	0.49876	345.9	34.33	
-0.49999	-0.49887	0.99752	342.6	40.79	
0	0.99773	0	297.5	63.24	
-0.49999	-0.49887	0	262.3	41.00	
0	0	0.99752	237.2	57.48	
0.49999	0.49887	0.99752	213.9	68.99	
0	0.49887	0.49876	212.2	43.64	
-0.49999	0.49887	0	204.5	37.19	
0.49999	0	0.49876	189.8	33.10	
0.99997	0.99773	0.99752	182.8	31.32	
-0.49999	0	0.49876	145.4	29.24	
0.49999	0.99773	0.49876	87.9	18.58	

several correlation structures listed in each table and those of the remaining correlation structures. Not surprisingly, the correlation structures with the largest average NODES include the unsolved problem instances (250,000 NODES for each problem included in the average).

The ranked results in Tables 3.10 and 3.11 highlight patterns among the correlation structures. Though the correlation structures listed at the top of Table 3.10 are Type L distributions, when both Type L and Type U distributions were available for a particular correlation structure, the Type U distribution yielded the more difficult problems. There are more negative values of $\rho_{A^1A^2}$ in the correlation structures listed in the top portion of Tables 3.10 and 3.11 than in the bottom portion. The potential influence of the individual correlation terms is examined in the next subsection. Finally, challenging problems seem to have larger differences between the values of each correlation terms within a correlation structure.

3.7.2 Individual correlation term influence

This experiment contains three subsets of design points in which each correlation term is specified at all five design settings while the two remaining correlation terms are zero. These subsets allow one to determine whether changes in that particular correlation term influence CPLEX performance and which correlation terms have the greatest relative influence. For each correlation term, a null hypothesis of no influence was tested using the KW test. The test statistics and corresponding p -values are provided in Table 3.12. For both correlation measures, the ρ_{CA^2} and $\rho_{A^1A^2}$ terms have a significant influence on CPLEX performance, while the ρ_{CA^1}

Table 3.12: Results of Kruskal-Wallis Tests on Each Correlation Term

Correlation Term	Test Statistic	p-value	Correlation Term	Test Statistic	p-value
ρ_{C,A^1}^P	1.2304	0.873	ρ_{C,A^1}^S	4.5899	0.332
ρ_{C,A^2}^P	18.2345	0.001	ρ_{C,A^2}^S	11.6578	0.020
ρ_{A^1,A^2}^P	14.1310	0.007	ρ_{A^1,A^2}^S	13.6610	0.008

(a) Pearson Measure

(b) Spearman Measure

term does not. A possible explanation is that the distribution for A^1 has a larger mean and variance than the distribution for A^2 . The larger variance of A^1 then produces a larger variance in the right-hand side coefficient for the first constraint, b_1 , which in turn produces more problems with loose constraints. As will be shown in the next subsection, loose constraints make for easier problems thereby reducing the influence of the ρ_{CA^1} term on CPLEX performance.

The lack of significance of $\rho_{CA^1}^S$ seems to conflict with Table 3.11 which lists $\mathbf{S} = (-0.99997, 0, 0)$ as a challenging correlation structure. The large average NODES associated with this structure is due to the extremely difficult problems that result when this structure is coupled with slackness setting $(S_1, S_2) = (0.30, 0.70)$; average NODES is 133,206 for these 5 problems. Interestingly, for the same correlation structure when $(S_1, S_2) = (0.70, 0.30)$ average NODES is just 45. A KW test is a ranks test that is not necessarily affected by such extreme data points. However, this example demonstrates that constraint slackness settings likely represent a significant influence on CPLEX.

Table 3.13: Mean NODES by Constraint Slackness Setting

S_1	S_2	Mean NODES	S_1	S_2	Mean NODES
0.30	0.30	3524.17	0.30	0.30	4290.87
0.30	0.70	3191.36	0.30	0.70	6307.78
0.70	0.30	2299.12	0.70	0.30	3110.35
0.70	0.70	335.34	0.70	0.70	1286.33

(a) Pearson Measure (b) Spearman Measure

3.7.3 Constraint slackness influences

Table 3.13 summarizes CPLEX performance by constraint slackness settings. Tight constraints, and in particular a tight first constraint (i.e., $S_1 = 0.3$), seem to produce the more challenging test problems. The data are grouped by slackness settings for a KW test for a difference in CPLEX performance due to constraint slackness setting. The KW test statistics of 76.76 for Pearson correlation test problems and 235.07 for Spearman correlation test problems equate to p -values of less than 1.0×10^{-16} in each test. So, constraint slackness settings represent a significant influence on CPLEX performance.

Notice the mean NODES in Table 3.13(a) and 3.13(b) for $(S_1, S_2) = (0.30, 0.70)$ and $(S_1, S_2) = (0.70, 0.30)$ are quite different. The standard errors listed do not suggest the means are different. However, by pairing observations with $(S_1, S_2) = (0.30, 0.70)$ with the corresponding observations having $(S_1, S_2) = (0.70, 0.30)$, a sign test may be conducted for a null hypothesis of no difference in CPLEX performance for mixed slackness settings. The sign test results are provided in Table 3.14 for each correlation measure along with the $\alpha = 0.05$ acceptance regions.

Table 3.14: Sign Test Results for Performance Differences Between Mixed Constraint Slackness Levels

Correlation Measure	Total $d_i \neq 0$	Total $d_i > 0$	Acceptance Region	p -value
Pearson Measure	273	177	(120,153)	<0.0001
Spearman Measure	278	201	(123,155)	<0.0001

The p -values indicate that, with a mixed constraint slackness setting, a tight first constraint produces a more challenging problem. CPLEX performance appears more sensitive to low S_1 values than to low S_2 values given mixed slackness settings and the distributions specified for A^1 and A^2 .

3.7.4 The interaction between correlation structure and constraint slackness

The information contained in Tables 3.15 through 3.18 provide some insight into the interaction between correlation and constraint slackness for each correlation measure. Table 3.15 lists average NODES for each setting of ρ_{CA^1} and S_1 and for each setting of ρ_{CA^2} and S_2 . Regardless of slackness value, extreme negative correlation between objective function and constraint coefficients yields problems that challenge CPLEX. For the experiment design points selected, $\rho_{A^1A^2} < 0$ whenever $\rho_{CA^1} < 0$ or $\rho_{CA^2} < 0$. So the effect of negative ρ_{CA^2} or ρ_{CA^1} , which conflicts with other studies on similar types of problems, may be better attributed to $\rho_{A^1A^2} < 0$. There is also a tendency for the combination of extreme positive correlation and a loose constraint to yield problems that challenge CPLEX.

Table 3.15: Interaction of Constraint Slackness and Correlation Type on Average NODES

First Constraint Effects		
$\rho_{CA^1}^P$	$S_1 = 0.30$	$S_1 = 0.70$
-0.99997	19668.6	17803.9
-0.49999	575.2	208.0
0.0	3833.8	208.9
0.49999	872.7	141.6
0.99997	82.4	11211.6
Second Constraint Effects		
$\rho_{CA^2}^P$	$S_2 = 0.30$	$S_2 = 0.70$
-0.99773	14994.5	5693.8
-0.49887	5115.6	513.9
0.0	625.6	390.5
0.49887	352.1	367.2
0.99773	54.1	10182.5

(a) Pearson Measure

First Constraint Effects		
$\rho_{CA^1}^S$	$S_1 = 0.30$	$S_1 = 0.70$
-0.99997	40558.7	35594.5
-0.49999	815.0	804.3
0.0	3427.7	377.9
0.49999	1057.8	451.1
0.99997	1210.0	12271.4
Second Constraint Effects		
$\rho_{CA^2}^S$	$S_2 = 0.30$	$S_2 = 0.70$
-0.99773	19611.6	7739.9
-0.49887	875.8	2819.8
0.0	3375.9	4111.3
0.49887	694.0	559.8
0.99773	5286.8	10523.5

(b) Spearman Measure

Table 3.16 lists average NODES by slackness setting for each level of interconstraint correlation. Extreme negative correlation between constraint coefficients yields challenging problems. The average NODES values are relatively high for extreme positive values of $\rho_{A^1A^2}$ when $S_1 = 0.30$. With a mixed slackness setting and extreme negative interconstraint correlation, the average NODES is also very high. The “bump” in Table 3.16(a) for $\rho_{A^1A^2}^P = -0.49876$ and tight constraints is caused by a particularly challenging design point, $\mathbf{P} = (0, -0.49887, -0.49876)$ and $(S_1, S_2) = (0.30, 0.30)$, for which two test problems were not solved in the full 250,000 NODES limit.

Space prohibits listing the performance averages for all 224 design points, so Tables 3.17 and 3.18 list just the three design points at each extreme level of performance. Notice the disparity of CPLEX performance between the design points

Table 3.16: Average NODES by Inter-Constraint Slackness and Correlation

$\rho_{A^1 A^2}^P$	$S_1 = 0.30$ $S_2 = 0.30$	$S_1 = 0.70$ $S_2 = 0.70$
-0.99752	655.9	1337.1
-0.49876	10888.3	289.9
0.0	497.1	177.4
0.49876	565.4	282.3
0.99752	4954.9	177.7
$\rho_{A^1 A^2}^P$	$S_1 = 0.30$ $S_2 = 0.70$	$S_1 = 0.70$ $S_2 = 0.30$
-0.99752	20260.3	22367.2
-0.49876	655.5	204.4
0.0	191.2	302.6
0.49876	648.3	188.67
0.99752	11144.0	1193.1

(a) Pearson Measure

$\rho_{A^1 A^2}^S$	$S_1 = 0.30$ $S_2 = 0.30$	$S_1 = 0.70$ $S_2 = 0.70$
-0.997752	33393.4	11110.9
-0.49876	1564.7	483.4
0.0	975.2	307.7
0.49876	520.4	266.7
0.99752	5315.4	200.0
$\rho_{A^1 A^2}^S$	$S_1 = 0.30$ $S_2 = 0.70$	$S_1 = 0.70$ $S_2 = 0.30$
-0.99752	22411.3	22995.7
-0.49876	690.3	614.7
0.0	7699.1	403.8
0.49876	5225.6	1149.9
0.99752	3954.4	5445.6

(b) Spearman Measure

listed in Tables 3.17(a) and 3.18(a) versus those in Tables 3.17(b) and 3.18(b). Problems requiring more NODES involve negative correlation and tight constraints while problems requiring less NODES involve primarily positive correlation values and loose constraints. Further, it appears that problems with large differences between the values of correlation terms require more NODES.

One way to generate a difficult 2KP instance is to induce extreme negative correlation between the objective function and a tight constraint. Another approach to creating difficult problems is to induce negative correlation between constraint coefficients, such as specifying either of the extreme correlation structures $\rho = (0.99997, -0.99773, -0.99752)$ or $\rho = (-0.99997, 0.99773, -0.99752)$. Conversely, one can create easier problems by avoiding any negative correlation in the population correlation structure and making the constraints loose.

Table 3.17: Design Points Requiring Most and Least Average NODES for Pearson Correlation Problems

(a) Design Points Averaging Most NODES						
$\rho_{CA^1}^P$	$\rho_{CA^2}^P$	$\rho_{A^1A^2}^P$	S_1	S_2	Mean	Std Error
0.0	-0.49887	-0.49876	0.30	0.30	119724.4	55912.53
0.99997	-0.99773	-0.99752	0.70	0.30	111466.4	55586.62
-0.99997	0.99773	-0.99752	0.30	0.70	100801.2	60914.42
(b) Design Points Averaging Least NODES						
$\rho_{CA^1}^P$	$\rho_{CA^2}^P$	$\rho_{A^1A^2}^P$	S_1	S_2	Mean	Std Error
0.99997	0.49987	0.49876	0.70	0.30	3.8	1.96
0.49999	0.0	0.0	0.70	0.30	3.8	2.35
0.99997	-0.49987	-0.49876	0.70	0.30	4.4	2.16

Table 3.18: Design Points Requiring Most and Least Average NODES for Spearman Correlation Problems

(a) Design Points Averaging Most NODES						
$\rho_{CA^1}^S$	$\rho_{CA^2}^S$	$\rho_{A^1A^2}^S$	S_1	S_2	Mean	Std Error
0.99997	-0.99773	-0.99752	0.70	0.30	113852.0	56803.73
-0.99997	0.99773	-0.99752	0.30	0.70	103109.6	60021.89
0.0	0.0	-0.99752	0.30	0.30	99214.6	50753.77
(b) Design Points Averaging Least NODES						
$\rho_{CA^1}^S$	$\rho_{CA^2}^S$	$\rho_{A^1A^2}^S$	S_1	S_2	Mean	Std Error
0.0	0.49987	0.49876	0.70	0.70	22.8	9.19
0.49999	0.99773	0.49876	0.70	0.70	23.0	18.03
-0.99997	0.0	0.0	0.70	0.70	27.6	20.08

3.7.5 Regression models for NODES

This section describes the regression models fit to the experiment data to describe the relationship between the experiment design factors and the performance measure NODES. These models were developed using a stepwise regression procedure that maximizes the coefficient of determination, R^2 . The transformed response in each model is the natural logarithm of NODES.

Define disparity, D , as the largest absolute deviation between any two of the correlation terms. Table 3.19 lists the best regression model for each correlation measure. For each term significant at the $\alpha = 0.05$ level, the regression model coefficient is provided, while p -values are provided for the remaining terms in the model. Despite the transformation on NODES, both regression models have low values for R^2 .

The model for the Pearson measure is similar to that for the Spearman measure. There are ten significant factors in common, each significant term having the same sign. The factor D is significant, supporting earlier observations regarding its influence. The coefficient for the constraint slackness factor indicates that loose constraints tend to reduce NODES. Also supporting earlier findings are the significant interaction terms on constraint slackness and correlation setting. Neither of the ρ_{CA^1} and ρ_{CA^2} factors are significant, although the sign test previously found ρ_{CA^2} significant. This may be due to these terms being correlated with D , and the interaction terms involving ρ_{CA^1} and ρ_{CA^2} being significant factors in the model.

Table 3.19: Regression Model of CPLEX Results

Source	Pearson Measure LN(NODES)		Spearman Measure LN(NODES)	
	Coefficient	<i>p</i> -value	Coefficient	<i>p</i> -value
Intercept	9.20		10.09	
S_1	-9.02		-8.36	
S_2	-5.25		-5.67	
ρ_{CA^1}		0.725		0.465
ρ_{CA^2}		0.637		0.323
$\rho_{A^1A^2}$	-1.56		-1.25	
D	-1.89		-0.99	
$S_1 \times S_2$	10.61		10.24	
$S_1 \times \rho_{CA^1}$	2.49		2.51	
$S_1 \times \rho_{CA^2}$	-2.19		-1.15	
$S_1 \times \rho_{A^1A^2}$		0.583		0.149
$S_2 \times \rho_{CA^1}$	-1.54			0.243
$S_2 \times \rho_{CA^2}$	1.83		1.63	
$S_2 \times \rho_{A^2A^2}$	1.48		1.25	
$\rho_{CA^1} \times \rho_{CA^2}$		0.152		
$\rho_{CA^1} \times \rho_{A^1A^2}$	-0.82			0.181
$\rho_{CA^2} \times \rho_{A^1A^2}$		0.179	-1.95	
$S_1 \times D$	1.98			
$S_2 \times D$				0.394
$\rho_{CA^1} \times D$	-0.51			0.105
$\rho_{CA^2} \times D$		0.117	-0.77	
$\rho_{A^1A^2} \times D$		0.323	-0.41	
	$R^2 = 0.216$		$R^2 = 0.255$	

3.8 Analysis of heuristic performance

In this section, the influence of population correlation structure and constraint slackness settings on TOYODA performance is examined. The measure of performance for the analysis is REL. Two regression models are constructed to summarize the influence of constraint slackness and correlation on TOYODA performance.

3.8.1 Correlation structure influence

Tables 3.20 and 3.21 summarize TOYODA performance by population correlation structure for Pearson and Spearman problems, respectively. Past research has shown that TOYODA generally provides very good solutions, and the results in this study support this point. Although there is no drastic change between any two consecutive correlation structures listed in the tables, there is a noticeable difference in REL averages and the standard errors between those correlation structures at the top and those at the bottom of Tables 3.20 and 3.21. The $\rho_{A^1A^2}$ term appears to be a particularly important factor since the correlation structures with the larger average REL values generally include negative values of $\rho_{A^1A^2}$ while the opposite holds for those correlation structures with smaller average REL values. Independent sampling ($\rho = (0, 0, 0)$, Type U in Table 3.20), does not appear to provide particularly difficult problems. In fact the average REL with independent sampling is at the mean and just above the median value of average REL over all design points. Table 3.20 specifies REL by distribution type, Type L or Type U.

When both distributions were available for a particular correlation structure, the Type U distribution yielded the more difficult problems. This means Type U distributions might be preferable for generating difficult 2KP problem instances.

Of the 2240 test problems generated, TOYODA found an optimal solution in 156. Tables 3.20 and 3.21 list the number of optimal solutions found for each population correlation structure. A common feature among most of these correlation structures is non-positive values for ρ_{CA^1} and ρ_{CA^2} and non-negative values for $\rho_{A^1A^2}$. Consider the population correlation structure $\rho = (-0.99997, -0.99773, 0.99752)$ listed at the bottom of Table 3.20 and near the top of Tables 3.10 and 3.11. This structure produces challenging test problems for the CPLEX procedure, but for 37 of the 40 test problems generated with this correlation structure, TOYODA found an optimal solution. This is interesting because the problems that are challenging for one procedure may not be challenging for the other.

A formal test of no performance differences due to correlation structure is conducted using the KW test on the data grouped by correlation structure. The KW test statistics of 328.32 for Pearson correlation problems and 402.46 for Spearman correlation problems have p -values near zero. So, the correlation structure is a significant influence on TOYODA performance.

3.8.2 Individual correlation term influence

This experiment includes three subsets of design points in which each correlation term is specified at all design settings while the two remaining correlation terms are specified as zero. These subsets may be used to determine whether each particular

Table 3.20: TOYODA Results using Pearson Correlation Induction

Type	Correlation Values			Mean REL	Standard Error	Solved to CPLEX Value
	ρ_{CA1}^P	ρ_{CA2}^P	ρ_{A1A2}^P			
U	-0.49999	0.49887	-0.49876	2.04	0.407	
U	0.49999	-0.49887	-0.49876	1.89	0.362	
U	0.49999	0	0	1.74	0.314	
U	0	0.49887	0	1.61	0.259	
L	0	-0.49887	-0.49876	1.56	0.272	
L	0	0	-0.99752	1.50	0.311	1
U	-0.49999	0	0	1.45	0.331	1
L	-0.49999	0.49887	-0.99752	1.31	0.196	
U	0	-0.49887	0	1.28	0.281	
L	0	0.49887	-0.49876	1.24	0.224	
L	0	0.99773	0	1.22	0.171	
L	0	0	-0.49876	1.21	0.348	
L	-0.49999	0.49887	-0.49876	1.20	0.242	
U	0	0	-0.49876	1.03	0.205	
L	0.49999	-0.49887	-0.99752	0.99	0.115	
L	-0.49999	0.99773	-0.49876	0.97	0.115	1
L	0	0.49887	0	0.91	0.140	
L	0.49999	0	-0.49876	0.87	0.108	
L	0	-0.49887	0	0.84	0.215	
U	0.49999	0.49887	0.49876	0.82	0.121	
U	0	0	0	0.77	0.132	
L	0.49999	-0.49887	-0.49876	0.73	0.153	
L	-0.49999	0	-0.49876	0.73	0.129	2
L	-0.49999	0.49887	0	0.72	0.083	
L	0.49999	0.99773	0.49876	0.72	0.081	
L	-0.49999	0	0	0.70	0.095	
L	0.49999	0.49887	0	0.70	0.082	
L	-0.99997	0.99773	-0.99752	0.70	0.051	1
L	0	0.49887	0.49876	0.66	0.089	
U	0	0	0.49876	0.66	0.124	1
L	0.99997	-0.49887	-0.49876	0.64	0.061	1
L	0.99997	0	0	0.60	0.076	4
L	0.49999	0	0	0.57	0.045	
L	0	0	0	0.56	0.053	
L	-0.49999	0	0.49876	0.54	0.088	
L	-0.99997	0.49887	-0.49876	0.53	0.067	3
L	0.99997	-0.99773	-0.99752	0.51	0.039	
L	0	-0.99773	0	0.51	0.072	3
U	-0.49999	-0.49887	0.49876	0.51	0.219	4
L	0.99997	0.49887	0.49876	0.50	0.049	2
L	0.49999	-0.99773	-0.49876	0.42	0.066	3
L	0.49999	-0.49887	0	0.41	0.042	2
L	0.49999	0.49887	0.49876	0.40	0.055	
L	-0.49999	-0.49887	0	0.39	0.073	2
L	-0.99997	0	0	0.39	0.050	4
L	-0.49999	-0.99773	0.49876	0.38	0.080	8
L	0	0	0.49876	0.36	0.051	
L	-0.49999	-0.49887	0.49876	0.36	0.064	3
L	0.49999	0	0.49876	0.36	0.038	1
L	0.99997	0.99773	0.99752	0.36	0.044	1
L	0	-0.49887	0.49876	0.35	0.034	1
L	0.49999	0.49887	0.99752	0.22	0.025	2
L	-0.99997	-0.49887	0.49876	0.20	0.043	7
L	-0.49999	-0.49887	0.99752	0.18	0.026	4
L	0	0	0.99752	0.16	0.022	2
L	-0.99997	-0.99773	0.99752	0.00	0.000	20

Table 3.21: TOYODA Results using Spearman Correlation Induction

Correlation Values			Mean	Standard	Solved to
$\rho_{CA^1}^S$	$\rho_{CA^2}^S$	$\rho_{A^1A^2}^S$	REL	Error	CPLEX Value
-0.49999	0.49887	-0.99752	3.29	0.325	
0.99997	-0.49887	-0.49876	3.24	0.471	2
0.49999	-0.49887	-0.99752	3.15	0.342	
-0.49999	0.99773	-0.49876	2.99	0.406	1
0	0.49887	-0.49876	2.77	0.359	
-0.49999	0.49887	-0.49876	2.73	0.233	
0.99997	0	0	2.64	0.451	1
0.49999	0.99773	0.49876	2.53	0.305	
-0.99997	0.99773	-0.99752	2.48	0.425	
0.49999	-0.49887	-0.49876	2.37	0.183	
0	0.99773	0	2.35	0.337	
0.49999	0	-0.49876	2.28	0.245	
0.99997	-0.99773	-0.99752	2.24	0.419	
0.99997	0.49887	0.49876	2.24	0.289	
0	0.49887	0	2.14	0.213	
-0.49999	0.49887	0	1.86	0.311	
0.49999	-0.49887	0	1.75	0.314	1
0.49999	0	0	1.65	0.153	
-0.99997	0.49887	-0.49876	1.64	0.281	
0	0	-0.99752	1.63	0.226	1
0	0.49887	0.49876	1.60	0.312	2
0	-0.49887	-0.49876	1.53	0.191	
0.49999	-0.99773	-0.49876	1.53	0.242	
0.49999	0.49887	0	1.32	0.140	1
-0.49999	0	0	1.29	0.140	
0.49999	0	0.49876	1.18	0.172	
0	-0.99773	0	1.16	0.221	1
0	-0.49887	0	1.16	0.135	
-0.49999	0	-0.49876	1.15	0.172	
0	0	-0.49876	1.14	0.105	2
-0.99997	0	0	1.01	0.217	
-0.49999	0	0.49876	0.75	0.129	
0	-0.49887	0.49876	0.71	0.152	5
0.49999	0.49887	0.49876	0.69	0.053	1
0	0	0	0.69	0.048	1
0.99997	0.99773	0.99752	0.69	0.101	1
0	0	0.49876	0.61	0.087	2
-0.49999	-0.49887	0.49876	0.58	0.087	7
-0.49999	-0.49887	0	0.54	0.069	2
-0.49999	-0.99773	0.49876	0.46	0.085	1
0.49999	0.49887	0.99752	0.36	0.037	
-0.99997	-0.49887	0.49876	0.36	0.074	6
0	0	0.99752	0.18	0.036	8
-0.49999	-0.49887	0.99752	0.16	0.023	5
-0.99997	-0.99773	0.99752	0.07	0.044	17

Table 3.22: Results of Kruskal-Wallis Tests on Each Type of Correlation for REL

Type of Correlation	Test Statistic	p-value	Type of Correlation	Test Statistic	p-value
$\rho_{CA^1}^P$	17.205	0.0018	$\rho_{CA^1}^S$	26.787	< 0.0001
$\rho_{CA^2}^P$	17.193	0.0018	$\rho_{CA^2}^S$	30.930	< 0.0001
$\rho_{A^1A^2}^P$	43.79	< 0.0001	$\rho_{A^1A^2}^S$	36.058	< 0.0001

(a) Pearson Measure

(b) Spearman Measure

correlation term influences TOYODA performance and which correlation terms have the greatest influence. For each correlation term, a null hypothesis of no influence was tested using the KW test. The test statistics and corresponding p -values are provided in Table 3.22. In each case the null hypothesis is rejected, and the conclusion is that each correlation term influences TOYODA performance.

The test statistics for $\rho_{A^1A^2}^P$ and $\rho_{A^1A^2}^S$ imply that each term is a particularly significant factor influencing TOYODA procedure performance. As noted earlier, many of the previous studies on MKP heuristics include TOYODA as a benchmark procedure. Past research with induced correlation involves high positive values for each of ρ_{CA^1} and ρ_{CA^2} , which then implies a high positive value of $\rho_{A^1A^2}$, such as was the case in Balas and Martin (1980). However, this study finds that $\rho_{A^1A^2} < 0$ actually produces the more challenging test problems for TOYODA.

CPLEX and TOYODA results differ regarding the influence of the $\rho_{CA^1}^P$ and $\rho_{CA^1}^S$ terms. While neither term significantly influenced CPLEX, both strongly influence TOYODA performance. This is reconciled by noting that TOYODA transforms each 2KP so that $b_i = 1$, $i = 1, 2$. Thus, the influence of the mean and variance of A^1 in countering the influence of each correlation term is mitigated.

3.8.3 Constraint slackness influence

Table 3.23 lists the average REL by constraint slackness settings. Clearly, $(S_1, S_2) = (0.30, 0.30)$ represents the most difficult type of problem for TOYODA in terms of overall average REL. The differences among the other average REL values are quite small, as are their standard errors. A KW test is used to test for a difference in REL due to slackness setting. The KW test statistics of 192.09 for Pearson correlation problems and 178.49 for Spearman correlation problems have p -values near zero. So constraint slackness settings have a statistically significant influence on TOYODA performance, which agrees with past research.

Unlike the results from CPLEX presented in Table 3.13 there does appear to be a significant difference in REL values when $(S_1, S_2) = (0.30, 0.70)$ as compared to REL values when $(S_1, S_2) = (0.70, 0.30)$. Table 3.24 presents the sign test results for each correlation measure when slackness settings are mixed, with the $\alpha = 0.05$ acceptance regions for the test provided. These results indicate that when mixed slackness settings are involved, and the problems are based on the Pearson measure, a tight first constraint tends to be associated with more challenging problems. For the problems based on the Spearman measure both mixed slackness settings yield problems of about equal REL. However, because TOYODA normalizes each constraint, the interpretation of these results is not as straightforward as for CPLEX results.

Past heuristic research has not addressed the influence of mixed slackness levels. Balas and Martin (1980) use randomly generated slackness levels but do not examine the effects of mixed levels on heuristic procedure performance.

Table 3.23: Mean REL by Constraint Slackness Setting

S_1	S_2	Mean REL	S_1	S_2	Mean REL
0.30	0.30	1.53	0.30	0.30	2.95
0.30	0.70	0.42	0.30	0.70	0.95
0.70	0.30	0.58	0.70	0.30	1.00
0.70	0.70	0.54	0.70	0.70	1.10

(a) Pearson Measure

(b) Spearman Measure

Table 3.24: Sign Test Results for Performance Differences Between Mixed Constraint Slackness Levels

Correlation Measure	Total $d_i \neq 0$	Total $d_i > 0$	Acceptance Region	p -value
Pearson Measure	273	162	(120,153)	0.0008
Spearman Measure	276	151	(122,154)	0.0520

3.8.4 The interaction between correlation structure and constraint slackness

While previous studies have examined the influence of constraint slackness settings, none have been able to examine the interaction between constraint slackness settings and population correlation structure.

Figures 3.4 through Figure 3.7 plot average REL values for various slackness-correlation combinations. In each plot, the correlation values on the X-axis are rounded for ease of presentation. Figures 3.4 and 3.5 plot results for Pearson problems; Figures 3.6 and 3.7 plot results for Spearman problems. In the two plots shown in both Figures 3.4 and 3.6, average REL values are plotted versus the correlation value between coefficients of the objective function and each constraint, for both tight and loose constraints. REL averages tend to vary directly with the correlation values meaning that the more challenging problems have higher corre-

lation values. At each correlation value, REL averages were lower for the larger slackness values, which means that better solutions are found for problems with looser constraints. Finally, the increasing differences between the REL values plotted for increasing correlation values indicate there is an interaction effect between the correlation and constraint slackness factors on TOYODA performance.

Figures 3.5 and 3.7 plot REL against the interconstraint correlation for tight and loose slackness settings for Pearson and Spearman problems, respectively. The trend is for decreasing values of $\rho_{A^1A^2}$ to result in increasing average REL values. At negative values of $\rho_{A^1A^2}$ and $(S_1, S_2) = (0.30, 0.30)$, there is quite a large difference in average REL as compared to when $(S_1, S_2) = (0.70, 0.70)$. Finally, these plots provide strong evidence of an interaction between correlation and the constraint slackness setting.

Space prohibits listing the performance averages for all 224 design points, so the three design points with the best and worst levels of performance are listed in Tables 3.25 and 3.26. These results are what would be expected after examining Figures 3.4 through 3.7. Both constraints being tight yields harder problems, particularly when combined with negative values of $\rho_{A^1A^2}$. Test problems with the worst REL averages have $\rho_{CA^1} \geq 0$, $\rho_{CA^2} \geq 0$, and $\rho_{A^1A^2} \leq 0$. Within the correlation structures for the easier problems, the largest absolute difference between any correlation terms seems larger than in the correlation structure of the harder problems. The influence of this phenomenon is examined later using a regression model. Tables 3.25 and 3.26, though purposely not all inclusive, illustrate the range of TOYODA performance, which is not so apparent in the data presented in Tables 3.20 and 3.21.

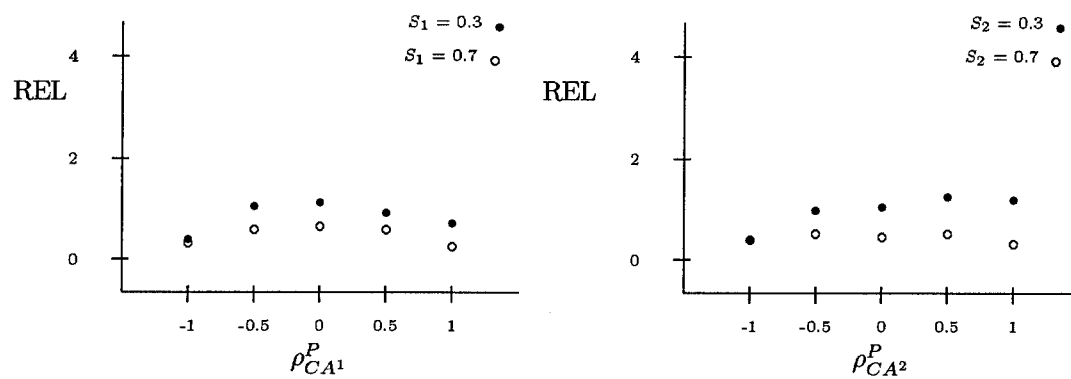


Figure 3.4: Pearson Correlation Measure - REL $\times$ Correlation Setting

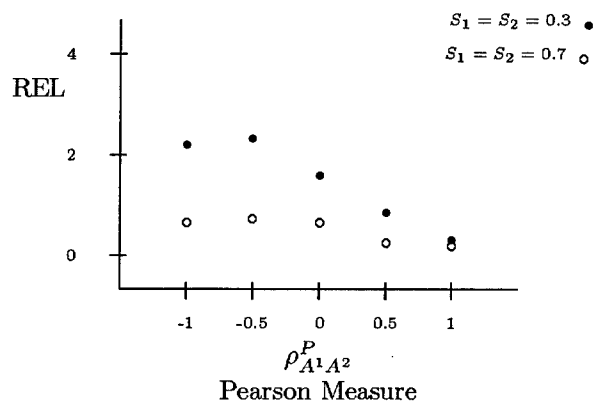


Figure 3.5: Inter-Constraint Correlation - REL $\times$ Correlation Setting

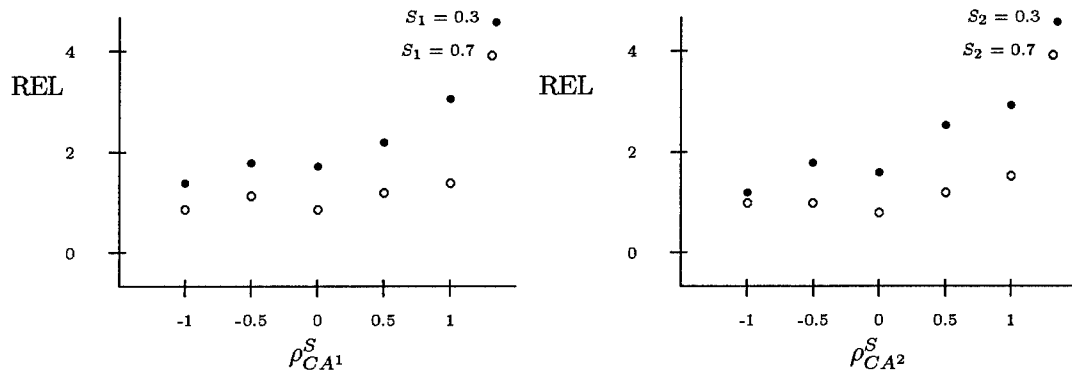
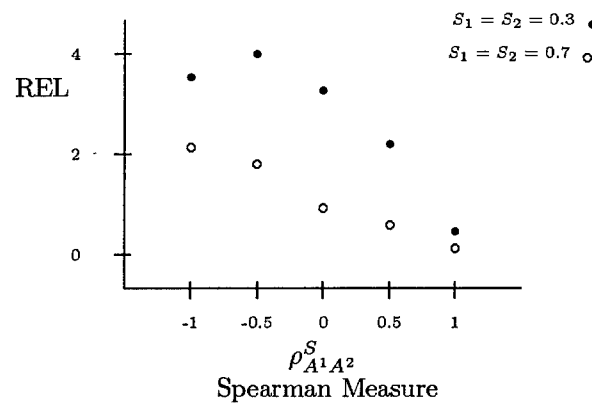
Figure 3.6: Spearman Correlation Measure - REL $\times$ Correlation SettingFigure 3.7: Inter-Constraint Correlation - REL $\times$ Correlation Setting

Table 3.25: Design Points with Extreme REL Averages for Pearson Correlation Problems

(a) Design Points Averaging Worst REL						
$\rho_{CA^1}^P$	$\rho_{CA^2}^P$	$\rho_{A^1A^2}^P$	S_1	S_2	Mean	Std Error
-0.49999	0.49887	-0.49876	0.30	0.30	4.6	0.73
0.0	0.0	-0.99752	0.30	0.30	4.2	1.06
0.0	-0.49887	-0.49876	0.30	0.30	4.1	0.71
(b) Design Points Averaging Best REL						
$\rho_{CA^1}^P$	$\rho_{CA^2}^P$	$\rho_{A^1A^2}^P$	S_1	S_2	Mean	Std Error
-0.99997	-0.99773	0.99752	Any	Any	0.0	0.0
-0.99997	-0.49887	0.49876	0.30	0.70	0.0	0.0
-0.99997	0.0	0.0	0.30	0.30	0.01	0.005

Table 3.26: Design Points with Extreme REL Averages for Spearman Correlation Problems

(a) Design Points Averaging Worst REL						
$\rho_{CA^1}^S$	$\rho_{CA^2}^S$	$\rho_{A^1A^2}^S$	S_1	S_2	Mean	Std Error
-0.49999	0.49887	-0.49876	0.30	0.30	6.7	0.66
0.99997	0.0	0.0	0.30	0.30	6.4	1.32
0.0	0.49887	-0.49876	0.30	0.30	6.1	0.94
(b) Design Points Averaging Best REL						
$\rho_{CA^1}^S$	$\rho_{CA^2}^S$	$\rho_{A^1A^2}^S$	S_1	S_2	Mean	Std Error
-0.99997	-0.99773	0.99752	0.70	0.70	0.0	0.0
-0.99997	-0.99773	0.99752	0.30	0.70	0.0	0.0
-0.99997	-0.99773	0.99752	0.30	0.30	0.01	0.006

3.8.5 Regression models for REL

Table 3.27 contains the regression models developed to describe REL in terms of the experiment design factors. For these models, a “best” regression is defined as the model maximizing the value of R^2 . The disparity term, D, represents the largest absolute deviation between any two correlation terms within the correlation structure. The coefficients of the significant terms in the regression model are indicated; p -values are provided for the insignificant terms included in the model.

Each model contains significant terms for both constraint slackness factors, which agrees with the KW test results previously presented. Also significant are the constraint slackness and objective function-constraint correlation interactions, which is the trend seen in Figures 3.4-3.7. The D term is significant, as was the case with the regression model for the CPLEX results. Comparing these models side-by-side there are clear differences attributable to the type of correlation measure. Although the regression models have nine common significant terms, and most of these common terms agree in sign (8 of the 9), the magnitudes of most of these terms differ.

Although the TOYODA heuristic provides generally good solutions across a wide range of problems, the procedure is sensitive to variations in correlation structure and constraint slackness settings. With certain types of correlation structures TOYODA provides excellent solutions. Of particular interest is the influence of negative values of $\rho_{A^1A^2}$ and the interaction between constraint slackness settings and values specified for the correlation terms in the correlation structure.

Table 3.27: Regression Model of TOYODA Results

Source	Pearson Measure LN(REL)		Spearman Measure REL	
	Coefficient	p-value	Coefficient	p-value
Intercept	3.46		6.61	
S_1	-5.23		-8.81	
S_2	-6.21		-9.27	
ρ_{CA^1}		0.124		0.520
ρ_{CA^2}		0.608	1.05	
$\rho_{A^1A^2}$		0.165	-2.00	
Disparity	-0.90		0.69	
$S_1 \times S_2$	8.83		13.11	
$S_1 \times \rho_{CA^1}$	-2.49		-1.13	
$S_1 \times \rho_{CA^2}$	1.83			0.113
$S_1 \times \rho_{A^1A^2}$		0.180	0.77	
$S_2 \times \rho_{CA^1}$	3.26		1.12	
$S_2 \times \rho_{CA^2}$	-1.40		-1.32	
$S_2 \times \rho_{A^1A^2}$		0.761	0.80	
$\rho_{CA^1} \times \rho_{CA^2}$		0.124		0.058
$\rho_{CA^1} \times \rho_{A^1A^2}$		0.908		0.361
$\rho_{CA^2} \times \rho_{A^1A^2}$				
$S_1 \times \text{Disparity}$		0.229		
$S_2 \times \text{Disparity}$		0.787		0.398
$\rho_{CA^1} \times \text{Disparity}$	0.31		0.32	
$\rho_{CA^2} \times \text{Disparity}$	0.38			0.134
$\rho_{A^1A^2} \times \text{Disparity}$		0.205		0.149
	$R^2 = 0.477$		$R^2 = 0.466$	

3.9 Analysis of LP-IP Gap

The size of the LP-IP gap in an optimization problem is sometimes viewed as a factor influencing the performance of solution procedures (Chang and Shepardson, 1982). Though not known in advance, by solving the test problems, one can examine how the experiment design factors influence the size of the LP-IP gap.

The sign test is used to test the null hypothesis of no difference in LP-IP gap value in synthetic test problems based on the type of correlation measure. This is done by separating the problems by the correlation measure and then pairing the problems by design point and replication number. The LP-IP gap was larger for Spearman correlation-based problems in 875 of the 1120 pairings (16 pairings were equal). The sign test results indicate that problems generated based on the Spearman correlation structure have larger LP-IP gap values. Among the 201 test problems for which the LP solution equaled the IP solution, only 51 problems were generated using the Spearman correlation measure.

KW tests were used to test whether there were LP-IP gap differences due to the individual correlation values within each correlation structure and whether the constraint slackness settings matter. The results of these tests are presented in Table 3.28. The constraint slackness setting is a very significant factor influencing the size of the LP-IP gap for the test problems generated in this study. Problems

Table 3.28: KW Test results for LP-IP Gap

Parameters	Pearson Measure		Spearman Measure	
	Test		Test	
	Statistic	<i>p</i> -value	Statistic	<i>p</i> -value
ρ_{CA^1}	28.45	< 0.0001	17.18	0.0017
ρ_{CA^2}	18.03	0.0012	14.89	0.0049
$\rho_{A^1A^2}$	15.22	0.0043	9.11	0.0584
Slackness	60.67	< 0.0001	78.60	< 0.0001

with tighter constraints tend to have larger LP-IP gap values. All the correlation terms, with the exception of $\rho_{A^1A^2}$ for Spearman problems, significantly influenced the LP-IP gap. Among the three correlation terms, the ρ_{CA^1} term seems to be the most significant.

Although the LP-IP gap for synthetic optimization problems is unknown beforehand, the insight gained in this section could be useful when defining problem generation parameters for examining solution procedure performance. This could be particularly useful when proposed procedures, either exact or heuristic, rely on LP relaxations during the solution process.

3.10 Discussion and Conclusions

This paper examined 2KP solution procedures using synthetic test problems generated based on a variety of correlation structures and constraint slackness settings using both Pearson product-moment and Spearman rank correlation generation

methods. Hooker (1994) states that an alternative to empirical studies with unrepresentative optimization problem sets, is to investigate "how algorithmic performance depends on problem characteristics." This study shows that the correlation structure among test problem coefficients and the type of correlation induced influence solution procedure performance on 2KP instances.

Not only does the correlation structure matter, but the correlation measure affects solution procedure performance. Systematically varying the problem correlation structure yields a more complete range of problems than independent sampling does. Interconstraint correlation is shown to be a significant factor influencing performance of solution methods. For a specified correlation structure, a Type U composite distribution tends to produce a more difficult problem than a Type L distribution. So, the level of independent sampling with a composition-based sampling method affects solution procedure performance. Constraint slackness is an established problem generation parameter but this study highlights the interaction between constraint slackness and correlation structure. Finally, for some design points, CPLEX performed poorly while TOYODA found the optimal solutions. So, one must always be cautious about generalizing the results observed with one solution method to other methods. Furthermore, this result indicates that different test problems may be appropriate for evaluating different types of solution procedures.

There are several areas of further investigation. For instance, one could examine other correlation induction methods, more constraint slackness settings, larger test problems or problems involving more than two constraints. Another avenue could examine various types of heuristics, such as was done by Zanakakis (1977), or various

optimization methods, to compare how the procedures react to particular test problem parameter settings. In terms of optimization methods, one could easily examine how problem generation parameters settings influence the effectiveness of pre-processing routines such as valid cut generators, bounding procedures, or problem reduction algorithms.

References

- Amini, M. M. and M. Racer. 1994. A Rigorous Computational Comparison of Alternative Solution Methods for the Generalized Assignment Problem. *Management Science*, **40**(7), 868-890.
- Balas, E. and C.H. Martin. 1980. Pivot and Complement - A Heuristic for 0-1 Programming. *Management Science*, **26**(1), 86-96.
- Balas, E. and E. Zemel. 1980. An algorithm for large zero-one knapsack problems. *Operations Research*, **28**(5), 1130-1154.
- Cario, M. C., J. J. Clifford, R. R. Hill, J. Yang, K. Yang, C. H. Reilly. 1995. Alternative Methods for Generating Synthetic Generalized Assignment Problems. *Working Paper Series Number 1995-006*. Department of Industrial, Welding and Systems Engineering, The Ohio State University, Columbus, Ohio.
- Chang, M. G. and F. Shepardson. 1982. An Integer Programming Test Problem Generator, in *Lecture Notes in Economics and Mathematical Systems: Evaluating Mathematical Programming Techniques*, J. M. Mulvey (ed.). Springer-Verlag, Berlin, 146-159.
- CPLEX Optimization Inc. 1993. *Using the CPLEX Callable Library and CPLEX Mixed Integer Library*. Incline Village, NV.
- Conover, W. J. 1980. *Practical Nonparametric Statistics*, 2ed. John Wiley & Sons, New York.
- Devroye, L. 1986. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.
- Fisher, M., R. Jaikumar, and L. Van Wassenhove. 1986. A Multiplier Adjustment Method for the Generalized Assignment Problem. *Management Science*, **32**(9), 1095-1103.

- Fréville, A. and G. Plateau. 1993. An exact search for the solution of the surrogate dual of the 0-1 bidimensional knapsack problem. *European Journal of Operational Research*, **68**(3), 413-421.
- Fréville, A. and G. Plateau. 1994. An efficient preprocessing procedure for the multidimensional 0-1 knapsack problem. *Discrete Applied Mathematics*, **49**, 189-212.
- Frieze, A. M. and M. R. B. Clarke. 1984. Approximation algorithms for the m -dimensional 0-1 knapsack problem: Worst-case and probabilistic analyses. *European Journal of Operational Research*, **15**, 100-109.
- Greenberg, H. J. 1990. Computational Testing: Why, How and How Much. *ORSA Journal on Computing*, **2**(1), 94-97.
- Guignard, M. and M.B. Rosenwein. 1989. An Improved Dual Based Algorithm for the Generalized Assignment Problem. *Operations Research*, **37**(4), 658-663.
- Hill, R. R. and C.H. Reilly. 1994. Composition for Multivariate Random Variables. *Proceedings of the 1994 Winter Simulation Conference*, eds. J.T. Tew, S. Manivannan, D.A. Sadowski, and A.F. Seila. 332-342. Institute of Electrical and Electronics Engineers, Orlando Florida.
- Hooker, J. N. 1994. Needed: An Empirical Science of Algorithms. *Operations Research*, **42**(2), 201-212.
- Iman, R.L. and W.J. Conover. 1982. A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables. *Communications in Statistics: Simulation and Computation*, **11**(3), 311-334.
- John, T. C. 1989. Tradeoff Solutions in Single Machine Production Scheduling for Minimizing Flow Time and Maximum Penalty. *Computers and Operations Research*, **16**(5), 471-479.
- Loulou, R. and E. Michaelides. 1979. New Greedy Heuristics for the Multidimensional 0-1 Knapsack Problem. *Operations Research*, **27**(6), 1101-1114.
- Martello, S. and P. Toth. 1979. The 0-1 Knapsack Problem. *Combinatorial Optimization*, N. Christofides, A. Mingozzi, C. Sandi, (eds.), John Wiley and Sons, New York, New York, 237-279.
- Martello, S. and P. Toth. 1981. An algorithm for the generalized assignment problem. in J.P. Brans (ed.), *Operational Research '81*, North-Holland, Amsterdam, 589-603.
- Mazzola, J.B. and A.W. Neebe. 1993. An Algorithm for the Bottleneck Generalized Assignment Problem. *Computers and Operations Research*, **20**(4), 355-362.

- Moore, B. A. and C. H. Reilly. 1993. Randomly Generating Synthetic Optimization Problems with Explicitly Induced Correlation. *OSU/ISE Working Paper Series Number 1993-002*. The Ohio State University, Columbus, Ohio.
- Pirkul, H. 1987. A Heuristic Solution Procedure for the Multiconstraint Zero-One Knapsack Problem. *Naval Research Logistics*, **34**(2), 161-172.
- Pollock, G.A. 1992. Evaluation of Solution Methods for Weighted Set Covering Problems Generated with Correlated Uniform Random Variables. Undergraduate Honors Thesis, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, OH.
- Potts, C. N. and L. N. Von Wassenhove. 1988. Algorithms for Scheduling a Single Machine to Minimize the Weighted Number of Late Jobs. *Management Science*, **34**(7), 843-858.
- Potts, C. N. and L. N. Van Wassenhove. 1992. Single machine scheduling to minimize total late work. *Operations Research*, **40**(3), 586-595.
- Reilly, C. H. 1991. Optimization test problems with uniformly distributed coefficients. *Proceedings of the 1991 Winter Simulation Conference*, eds. B. L. Nelson, W. D. Kelton, G. M. Clark, 866-874. Institute of Electrical and Electronics Engineers, Phoenix, Arizona.
- Saltzman, M. J. 1994. Survey: Mixed Integer Programming. *OR/MS Today*, **21**(2), 42-51.
- Toyoda, Y. 1975. A Simplified Algorithm for Obtaining Approximate Solutions to Zero-One Programming Problems. *Management Science*, **21**(12), 1417-1427.
- Trick, M. 1982. A Linear Relaxation Heuristic for the Generalized Assignment Problem. *Naval Research Logistics*, **39**(2), 137-151.
- Yang, J. 1994. A Computational Study on 0-1 Knapsack Problems Generated Under Explicit Correlation Induction. MS Thesis, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, Ohio.
- Zanakis, S. H. 1977. Heuristic 0-1 Linear Programming: An Experimental Comparison of Three Methods. *Management Science*, **24**, 91-104.

CHAPTER IV

CONCLUSIONS AND DISCUSSION

Chapter 2 presents a new composition method for generating values of multivariate random variables with explicit correlation induction. Several new concepts are introduced during this development: correlation points, extreme-correlation distributions, Type L and Type U composite distributions.

Using composite distributions for explicit correlation induction has several benefits. Sampling is easy to implement since the constituent components of the composite distribution, the extreme-correlation distributions, and the joint distribution under independence are easy to sample from. Many feasible correlation points have an associated composite distribution, which combined with the applicability of composite distributions to both continuous and discrete distributions, yields a method with wide applicability. Finally, additional modeling flexibility is available since for nearly all correlation points expressible as a composite distribution, there is an entire range of joint distributions available.

Chapter 3 applied composite distributions for trivariate random variables in an empirical study of the 2KP. This study examined three problem generation factors: the type of correlation, the correlation structure, and constraint slackness. A heuristic and a branch-and-bound procedure were examined to determine how performance is influenced by the problem generation factors.

The empirical study of the 2KP produced some exciting findings. The type of correlation, Pearson product-moment or Spearman rank, leads to differences in solution procedure performance. Each correlation term in the correlation structure and in particular the inter-constraint correlation term are significant problem generation parameters. Previous studies have not isolated the effect of the inter-constraint correlation. Mixed levels of constraint slackness are found to influence solution procedure performance. Moreover, this empirical study highlighted the synergistic effect between correlation structure and constraint slackness levels.

Many research opportunities could follow this effort. Some are mentioned in Chapters 2 and 3, but there are others. For example, the influence of factors such as the inter-constraint correlation on CPLEX performance might suggest new types of algorithms that would be more effective on instances where current branch-and-bound approaches fare poorly.

Bibliography

- [1] Amini, M. M. and M. Racer. 1994. A Rigorous Computational Comparison of Alternative Solution Methods for the Generalized Assignment Problem. *Management Science*, **40**(7), 868-890.
- [2] Balas, E. and C.H. Martin. 1980. Pivot and Complement - A Heuristic for 0-1 Programming. *Management Science*, **26**(1), 86-96.
- [3] Balas, E. and E. Zemel. 1980. An algorithm for large zero-one knapsack problems. *Operations Research*, **28**(5), 1130-1154.
- [4] Cario, M. C., J. J. Clifford, R. R. Hill, J. Yang, K. Yang, C. H. Reilly. 1995. Alternative Methods for Generating Synthetic Generalized Assignment Problems. *Working Paper Series Number 1995-006*. Department of Industrial, Welding and Systems Engineering, The Ohio State University, Columbus, Ohio.
- [5] Chang, M. G. and F. Shepardson. 1982. An Integer Programming Test Problem Generator, in *Lecture Notes in Economics and Mathematical Systems: Evaluating Mathematical Programming Techniques*, J. M. Mulvey (ed.). Springer-Verlag, Berlin, 146-159.
- [6] CPLEX Optimization Inc. 1993. *Using the CPLEX Callable Library and CPLEX Mixed Integer Library*. Incline Village, NV.
- [7] Conover, W. J. 1980. *Practical Nonparametric Statistics*, 2ed. John Wiley & Sons, New York.
- [8] Devroye, L. 1986. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.
- [9] Evans, J. R. 1984. The factored transportation problem. *Management Science*, **30**(8), 1021-1024.
- [10] Fisher, M., R. Jaikumar, and L. Van Wassenhove. 1986. A Multiplier Adjustment Method for the Generalized Assignment Problem. *Management Science*, **32**(9), 1095-1103.
- [11] Fréchet, M. 1951. Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon, Section A*, **14**, 53-77.

- [12] Fréville, A. and G. Plateau. 1993. An exact search for the solution of the surrogate dual of the 0-1 bidimensional knapsack problem. *European Journal of Operational Research*, **68**(3), 413-421.
- [13] Fréville, A. and G. Plateau. 1994. An efficient preprocessing procedure for the multidimensional 0-1 knapsack problem. *Discrete Applied Mathematics*, **49**, 189-212.
- [14] Frieze, A. M. and M. R. B. Clarke. 1984. Approximation algorithms for the m -dimensional 0-1 knapsack problem: Worst-case and probabilistic analyses. *European Journal of Operational Research*, **15**, 100-109.
- [15] Greenberg, H. J. 1990. Computational Testing: Why, How and How Much. *ORSA Journal on Computing*, **2**(1), 94-97.
- [16] Guignard, M. and M.B. Rosenwein. 1989. An Improved Dual Based Algorithm for the Generalized Assignment Problem. *Operations Research*, **37**(4), 658-663.
- [17] Hill, R. R. and C.H. Reilly. 1994. Composition for Multivariate Random Variables. *Proceedings of the 1994 Winter Simulation Conference*, eds. J.T. Tew, S. Manivannan, D.A. Sadowski, and A.F. Seila. 332-342. Institute of Electrical and Electronics Engineers, Orlando Florida.
- [18] Hooker, J. N. 1994. Needed: An Empirical Science of Algorithms. *Operations Research*, **42**(2), 201-212.
- [19] Iman, R.L. and W.J. Conover. 1982. A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables. *Communications in Statistics: Simulation and Computation*, **11**(3), 311-334.
- [20] John, T. C. 1989. Tradeoff Solutions in Single Machine Production Scheduling for Minimizing Flow Time and Maximum Penalty. *Computers and Operations Research*, **16**(5), 471-479.
- [21] Johnson, M. E., C. Wang, J. S. Ramberg. 1984. Generation of Continuous Multivariate Distributions for Statistical Applications. *American Journal of Mathematical and Management Sciences*, **4**(3 & 4), 225-248.
- [22] Johnson, M. E. and A. Tenenbein. 1981. A Bivariate Distribution Family with Specified Marginals. *Journal of the American Statistical Association*, **76**(3), 198-201.
- [23] Lewis, P. A. and E. J. Orav. 1989. *Simulation Methodology for Statisticians, Operations Analysts, and Engineers: Volume 1*. Wadsworth & Brooks/Cole: California.
- [24] Loulou, R. and E. Michaelides. 1979. New Greedy Heuristics for the Multidimensional 0-1 Knapsack Problem. *Operations Research*, **27**(6), 1101-1114.
- [25] Martello, S. and P. Toth. 1979. The 0-1 Knapsack Problem. *Combinatorial Optimization*, eds. N. Christofides, A. Mingozzi, C. Sandi. John Wiley and Sons, New York, New York, 237-279.

- [26] Martello, S. and P. Toth. 1981. An algorithm for the generalized assignment problem. in J.P. Brans (eds.), *Operational Research '81*, North-Holland, Amsterdam, 589-603.
- [27] Mazzola, J.B. and A.W. Neebe. 1993. An Algorithm for the Bottleneck Generalized Assignment Problem. *Computers and Operations Research*, **20**(4), 355-362.
- [28] Moore, B. A. and C. H. Reilly. 1993. Randomly Generating Synthetic Optimization Problems with Explicitly Induced Correlation. *OSU/ISE Working Paper Series Number 1993-002*. The Ohio State University, Columbus, Ohio.
- [29] Nelsen, R. B. 1987. Discrete Bivariate Distributions with Given Marginals and Correlation. *Communications in Statistics: Simulation and Computation*, **16**(1), 199-208.
- [30] Olkin, I. 1981. Range Restrictions for Product-Moment Correlation Matrices. *Psychometrika*, **4**(4), 469-472.
- [31] Page, E. S. 1965. On Monte Carlo Methods in Congestion Problems: I. Searching for an Optimum in Discrete Situations. *Operations Research*, **13**(2), 291-305.
- [32] Peterson, A.V. and R. A. Kronmal. 1982. On Mixture Methods for the Computer Generation of Random Variables. *The American Statistician*, **36**(3), 184-191.
- [33] Peterson, J. A. 1990. A Parametric Analysis of a Bottleneck Transportation Problem Applied to the Characterization of Correlated Discrete Random Variables, M.S. Thesis, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, Ohio.
- [34] Peterson, J. A. and C. H. Reilly. 1993. Joint Probability Mass Functions for Coefficients in Synthetic Optimization Problems. *Working Paper Series Number 1993-006*. The Ohio State University, Columbus, Ohio.
- [35] Pirkul, H. 1987. A Heuristic Solution Procedure for the Multiconstraint Zero-One Knapsack Problem. *Naval Research Logistics*, **34**(2), 161-172.
- [36] Pollock, G.A. 1992. Evaluation of Solution Methods for Weighted Set Covering Problems Generated with Correlated Uniform Random Variables. Undergraduate Honors Thesis, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, OH.
- [37] Potts, C. N. and L. N. Von Wassenhove. 1988. Algorithms for Scheduling a Single Machine to Minimize the Weighted Number of Late Jobs. *Management Science*, **34**(7), 843-858.
- [38] Potts, C. N. and L. N. Van Wassenhove. 1992. Single machine scheduling to minimize total late work. *Operations Research*, **40**(3), 586-595.
- [39] Reilly, C. H. 1991. Optimization test problems with uniformly distributed coefficients. *Proceedings of the 1991 Winter Simulation Conference*, eds. B. L. Nelson, W. D. Kelton, G. M. Clark, 866-874. Institute of Electrical and Electronics Engineers, Phoenix, Arizona.

- [40] Reilly, C. H. 1994. Alternative Input Models for Generating Synthetic Optimization Problems: Analysis and Implications. Working Paper 1994-001, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, OH.
- [41] Rousseeuw, P. J. and G. Molenberghs. 1994. The Shape of Correlation Matrices. *The American Statistician*, **48**(4), 276-279.
- [42] Rushmeier, R. A. and G. L. Nemhauser. 1993. Experiments with Parallel Branch-and-Bound Algorithms for the Set Covering Problem. *Operations Research Letters*, **13**(5), 277-285.
- [43] Saltzman, M. J. 1994. Survey: Mixed Integer Programming. *OR/MS Today*, **21**(2), 42-51.
- [44] Schmeiser, B. W. and R. Lal. 1982. Bivariate Gamma Random Vectors. *Operations Research*, **30**(2), 355-374.
- [45] Toyoda, Y. 1975. A Simplified Algorithm for Obtaining Approximate Solutions to Zero-One Programming Problems. *Management Science*, **21**(12), 1417-1427.
- [46] Trick, M. 1982. A Linear Relaxation Heuristic for the Generalized Assignment Problem. *Naval Research Logistics*, **39**(2), 137-151.
- [47] Whitt, W. 1976. Bivariate Distributions with Given Marginals. *The Annals of Statistics*, **4**(6), 1280-1289.
- [48] Yang, J. 1994. A Computational Study on 0-1 Knapsack Problems Generated Under Explicit Correlation Induction. MS Thesis, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, Ohio.
- [49] Zanakakis, S. H. 1977. Heuristic 0-1 Linear Programming: An Experimental Comparison of Three Methods. *Management Science*, **24**, 91-104.