Comparisons of Correlation Methods in Risk Analysis

by

Julie Carolyn Moore

Master's Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master's of Science

in

Mathematics

APPROVED:

Dr. Terry Herdman, Committee Chairman

Dr. Max Gunzburger

Dr. Kenneth Hannsgen

Dr. David Hudak

Dr. Eugene Cliff

April 21, 1994
Blacksburg, Virginia

Comparisons of Correlation Methods in Risk Analysis

by

Julie Carolyn Moore

Committee Chairman: Terry L. Herdman
Mathematics

Abstract

This thesis presents a comparison of correlation methods in risk analysis. A theoretical solution is given to the correlation problem along with a discussion of each method.

Each method is compared to a developed test case and two other cost projects. Restrictions on correlation coefficients are also given followed by the advantages and disadvantages of each method.

With All My Love

This Thesis is Dedicated to my Parents

In Honor of my Mother

Retta Cameron

And in Honor of my Father

Louis Cameron

Acknowledgments

I would like to thank Dr. Herdman, my advisor, for all of his help, guidance and time. It was greatly appreciated. I would also like to thank Dr. Gunzburger, Dr. Hannsgen, and Dr. Cliff for serving on my committee.

I would also like to acknowledge Dave Hudak of The Analytic Science Corporation for his special assistance. He devoted a tremendous amount of personal time and interest in this thesis.

I would like to especially thank James K. Lynch and Jeff Borggaard for all of their time and help.

Contents

Cover i

Abstract ii

Dedication iii

Acknowledgment iv

1 Introduction 1

2 Background Definitions and Theorems 4

2.1 Definitions and Theorems 4

2.2 Cost Risk Simulation Process 9

2.3 Theoretical Solution to Correlation 10

2.4 Properties of Correlation Matrix 11

2.5 Approximating Solutions to Correlation 13

2.5.1 Beta Fit 14

2.5.2 Choleski Method 15

2.5.3 Eigenvalue Method 18

2.5.4 Rank Correlation 21

3 Results 23

3.1 Test Case 23
3.2 30 Component Project 34
3.3 Three Component Case 39
3.4 Limitations on Correlation Coefficients 41

4 Conclusion **44**

Vita **48**

List of Figures

3.1 Direct Order 37

3.2 Reverse Order 38

List of Tables

3.1 Table 1 Risk Methods with Direct Order 31

3.2 Table 2 Choleski Method with Different Orders 32

3.3 Table 3 Eigenvalue Method with Different Orders 33

3.4 Table 4 Accuracy Based on Maximum Norms 33

3.5 Table 5 Risk Methods and their Accuracy 40

3.6 Table 6: Operation Counts for Different Correlation Methods 42

3.7 Table 7 Advantages and Disadvantages 43

Chapter 1

Introduction

Cost estimates are important for buying new military systems, protecting our environment, and building homes for communities. Cost analysts recognize that uncertainty is inherent in cost estimates, cost estimating models, development schedules, and the pace of technological maturation. Quantifying these uncertainties into a range of costs is known as cost risk analysis.

The importance of accurately portraying the range of possible costs for a project stems from the likelihood that actual costs will differ from estimated costs, usually in the direction of cost growth. Knowledge of the amount of additional funds required to successfully complete a project at a specified confidence level may influence a decision-maker's decision to proceed or halt a project.

When necessary data is available, a complete cost risk assessment identifies the cost uncertainty for each component of the total cost estimate. This is done by estimating a probability distribution for each component. From these individual distributions, an overall probability distribution is computed for the overall cost to the project. Computing a closed form overall distribution, however, is extremely complicated. Consequently, simulations are usually used to compute total cost probability distributions.

The distributions for each component are determined either from statistical data or from expert judgment on a range of possible costs for components. The types of distributions to be used can be chosen from a wide variety of distributions available in standard literature. Three of the most commonly used distributions in risk analysis are normals, triangulars, and betas. Normal distributions require only two parameters, the mean and variance. Triangular distributions require three, a high, low, and most likely estimate. Betas require four, a low and high estimate as well as a mean and variance. Although more complicated distributions are sometimes used, a lack of data often makes their use impractical.

Once the cumulative distribution has been established the cost analyst can pick off the cost at a desired confidence level such as the 50th, 70th, and 90th percentiles. With this information, the amount of funding required to obtain a desired confidence level can be estimated.

A major assumption of this process concerns the correlation between the cost components. That is, does the probability of the cost of one component affect the probability of the cost of another component and vice-versa? In many cases, every component is assumed to be independent. This, however, is usually not the case.

When correlations among cost components exist, several methods are available to handle these correlations. Some of these methods have been compared, but only against each other. Thus, no validation against a test case has been done. Further, these methods have only been compared for a small sample of components. Specifically, these correlation methods have typically been applied to four components or less. Finally, discussions of these methods have not considered possible restrictions on correlations when probability distributions for components are chosen.

This research considers the above issues. As background, a theoretical solution

will be given to the correlation problem. Once this is done, a test case will be developed. Four standard correlation methods will then be implemented with the input data used in the test case. Their results will be compared to the test's case results. Next, the four methods will be compared for a cost project decomposed into 30 cost components and last is a three cost component project where the cost components are different orders of magnitude. This will be followed by a theoretical discussion with a few examples of restrictions to correlations. Finally, the advantages and disadvantages of each method will be discussed.

Chapter 2

Background Definitions and Theorems

2.1 Definitions and Theorems

The following section gives a theoretical foundation for this research. This foundation includes definition, derivations, theorems and proofs of statistical and probabilistic concepts used in this work.

The following definitions are generally quoted from *Mathematical Statistics with Applications* by Mendenhall, Wackerly, and Scheaffer [1]. The first definition, however, is from *Probability and Statistics* by Degroot [2].

Definition 2.1 *For a sample space S, a random variable is a real valued function that is defined on the space S. In other words, in a particular experiment a random variable Y would be some function that assigns a real number $Y(s)$ to each possible outcome $s \in S$.*

The most basic component of probability theory is a random variable. A random variable Y is said to be discrete if it assumes only a finite or countably infinite number of distinct values. The expression $(Y = y)$ can be read as the set of all points in S assigned the value y by the random variable Y. The probability that Y takes on

the value y, $P(Y = y)$, is defined to be the sum of all the probabilities of all sample points in S that are assigned the value of y. $P(Y = y)$ will sometimes be denoted by $p(y)$. $p(y)$ is called the probability function for Y. For any discrete probability distribution, the following are true: $0 \leq y \leq 1$ for all y and $\sum_y p(y) = 1$, where the summation is over all values of y with nonzero probability. The probability distribution for a random variable is a theoretical model for the empirical distribution of data associated with a real population.

The type of random variable that takes on any value in an interval is called continuous. This leads to probability distributions for continuous random variables. Throughout this study only continuous random variables will be considered.

First a cumulative distribution function will be defined.

Definition 2.2 *Let y denote any random variable. The cumulative distribution function of y, denoted by $F(y)$, is given by $F(y) = P(Y \leq y), -\infty < y < \infty$.*

If $F(y)$ is a cumulative distribution function, then

1. $\lim_{y \to -\infty} F(y) = 0$
2. $\lim_{y \to \infty} F(y) = 1$
3. $F(y_b) \geq F(y_a)$ *if* $y_b > y_a$.

Thus $F(y)$ is always a nondecreasing function and since Y is continuous then $F(Y)$ must also be a smooth continuous curve. Next, let $F(y)$ be the distribution function for a continuous random variable Y. If the derivative of $F(y)$ exists, then $f(y)$ given by

$$f(y) = \frac{dF(y)}{dy} = F'(y)$$

is called the probability density function for the random variable Y . $F(y)$ can now be expressed as $F(y) = \int_{-\infty}^{y} f(t)\,dt$ where $f(y)$ is the probability density function. Because $F(y)$ is a nondecreasing function, it follows that the derivative $f(y)$ is never negative. Also, $\lim_{y\to\infty} F(y) = 1$ implies that $\int_{-\infty}^{\infty} f(t)\,dt = 1$. As a result of the previous discussion, the theorem follows:

Theorem 2.3 *If the random variable Y has a density function $f(y)$ and $a \leq b$, then the probability that Y falls in the interval $[a, b]$ is*

$$P(a \leq Y \leq b) = \int_{a}^{b} f(y)\,dy$$

where $f(y)$ is the probability density function for Y.

Numerical descriptives of probability functions of continuous random variables such as their means, variances, and standard deviations are also defined. The next definition describes the mean (average) for a continuous random variable Y.

Definition 2.4 *The expected value of a continuous random variable Y is*

$$E(Y) = \int_{-\infty}^{\infty} y\,f(y)\,dy$$

provided the integral exists.

The symbol μ is also used to express the expected value of a continuous random variable, $E(Y)$ as the mean of a population. The variance and standard deviation of a continuous random variable are defined as follows:

Definition 2.5 *The variance of a random variable Y is defined to be the expected value of $(Y - \mu^2)$. That is, $VAR(Y) = E[(Y - \mu)^2] = E(Y^2) - \mu^2$. The standard deviation of Y is the positive square root of $VAR(Y)$.*

Another statistical relationship used in this study is the notion of a normalized random variable. A normalized random variable is computed from a random variable with an expected value, μ, and a variance, σ^2.

Definition 2.6 *The normalized random variable Z is computed from the random variable Y using the relationship*

$$Z = \frac{Y - \mu}{\sigma}.$$

Thus the mean value of Z is 0 and its standard deviation is equal to 1.

Suppose that $Y_1, Y_2, Y_3, ..., Y_n$ denote the outcomes of n successive trials of an experiment. A set of specific outcomes, or sample costs can be expressed in terms of the intersection of n events $(Y_1 = y_1), (Y_2 = y_2), ..., (Y_n = y_n)$ denoted $(y_1, y_2, ..., y_n)$. Then to make inferences about the total system cost from which the element samples were drawn, the probability of the intersection $(y_1, y_2, ..., y_n)$ is calculated. Knowledge of their probability is fundamental to making inferences about the system from which the sample was drawn.

Definition 2.7 *Let Y_1 and Y_2 be discrete random variables. The joint probability distribution for Y_1 and Y_2 is given by*

$$h(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2)$$

defined for all real numbers y_1 and y_2.

The function $h(y_1, y_2)$ is referred to as the joint probability density function. Properties of the joint probability function $h(y_1, y_2)$ are $h(y_1, y_2) \geq 0$ for all y_1,y_2 and $\sum_{y_1,y_2} h(y_1, y_2) = 1$, where the sum is over all values (y_1, y_2) that are assigned nonzero probabilities.

Definition 2.8 *For any random variables, Y_1 and Y_2, the joint distribution function $H(a,b)$ is given by*

$$H(a,b) = P(Y_1 \leq a, Y_2 \leq b).$$

Up to this point, the calculation of the probability of the intersection of two events has been discusssed. There is also a relationship, however, between the probability distributions for Y_1 (and Y_2) and their joint density function, i.e. $p_1(y_1)$ (and $p_2(y_2)$) and $h(y_1, y_2)$. Finding $p_1(y_1)$ implies summing $p(y_1, y_2)$ over all values of y_2 and hence accumulating the probabilities on the y_1-axis (or margin). This leads to the following definition for continuous random variables.

Definition 2.9 *Let Y_1 and Y_2 be jointly continuous random variables with joint density function $h(y_1, y_2)$. Then the marginal density functions of Y_1 and Y_2, respectively, are given by*

$$f_1(y_1) = \int_{-\infty}^{\infty} h(y_1, y_2)\, dy_2 \text{ and } f_2(y_2) = \int_{-\infty}^{\infty} h(y_1, y_2)\, dy_1.$$

Finally, a definition is given quantifying the degree that Y_2 increases (or decreases) as Y_1 is increased.

Definition 2.10 *The covariance of Y_1 and Y_2 is defined to be the expected value of $(Y_1 - \mu_1)(Y_2 - \mu_2)$. In notation of expectation, the covariance will equal*

$$COV(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)]$$

where $\mu_1 = E(Y_1)$ and $\mu_2 = E(Y_2)$.

The larger the absolute value of the covariance of Y_1 and Y_2, the greater the dependence between Y_1 and Y_2. It is difficult to employ the covariance as an absolute

measure of dependence because its value depends upon the scale of measurement and so it is hard to determine whether a particular covariance is large at first glance. In order to avoid this problem, the measurement of dependence between Y_1 and Y_2 with a coefficient of linear correlation is standardized. The correlation coefficient, ρ, is related to the covariance and is defined as

$$\rho = \frac{COV(Y_1, Y_2)}{\sigma_1 \sigma_2}$$

where σ_1 and σ_2 are the standard deviations of Y_1 and Y_2, respectively. The correlation coefficient, ρ, satisfies the inequality $-1 \leq \rho \leq 1$. If $\rho = \pm 1$ then Y_1 and Y_2 are completely dependent and if $\rho = 0$, Y_1 and Y_2 have no dependency. The correlation between any two random variables Y_i and Y_j will be denoted $corr(Y_i, Y_j) = \rho_{ij}$. The following theorem is a convenient computational formula for the covariance.

Theorem 2.11 *Let Y_1 and Y_2 be random variables with a joint density function of $f(y_1, y_2)$. Then*

$$COV(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)] = E(Y_1 Y_2) - E(Y_1)E(Y_2).$$

At this point, all pertinent terms used in this research have been defined. Next, a description of a risk method is given.

2.2 Cost Risk Simulation Process

This risk method is based on a simulation process known as a Monte Carlo simulation. Monte Carlo simulation refers to the technique for using random numbers to sample from a probability distribution. Consider a system with n items. Each item

has a marginal distribution for its cost.

Item	Cost Random Variable	Marginal Distribution
1	x_1	$f_1(x_1)$
2	x_2	$f_2(x_2)$
3	x_3	$f_3(x_3)$
$\vdots$	$\vdots$	$\vdots$
n	x_n	$f_n(x_n)$

The simulation is iterated to obtain a large enough sample size to allow the range of possible outcomes to be realized. In other words, the overall cost distribution is approximated. For each iteration, a random sample cost is determined for each component. The total cost for that iteration is obtained by summing those values together. The result, when performed over numerous iterations, is an estimate of the probabilistic distribution of the cost of the system. Statistics such as percentiles and cumulative distributions are collected to provide information to decision makers. For a more detailed discussion concerning cost risk process, refer to [3].

2.3 Theoretical Solution to Correlation

The theoretical solution of implementing correlations is described next. As will be seen, attempting to implement this theoretical solution is extremely complicated and generally impractical. It, however, will be used for a specific test case in the next section.

Suppose that the joint distribution of the continuous non-negative random variables, $Y_1, Y_2, ..., Y_n$ is known. Further, this distribution is called $h(y_1, y_2, ..., y_n)$. Then, the probability that the sum of the $y_i's$ is less than some number S can be

expressed as

$$P(y_1+y_2+...+y_n \leq S) = \int_0^S \int_0^{S-y_n} \cdots \int_0^{S-y_n-y_{n-1}-...-y_2} h(s_1, s_2, ..., s_n)\, ds_1\, ds_2 \cdots ds_n$$

As can be seen, this computation requires calculating n integrals. Further, note that this specification required only knowledge of the joint distribution. In many cases, however, the joint distribution is unknown. Consequently, approximating methods which use marginal distributions and a correlation matrix are typically used. Four of these methods will be considered next.

2.4 Properties of Correlation Matrix

In order to incorporate the correlation coefficients they are introduced in matrix form. The correlation matrix consists of the correlation coefficients between every pair of components.

$$C = correlation\ matrix = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \rho_{nn} \end{bmatrix}$$

If $i = j$, then $\rho_{ij} = 1$ and if $i \neq j$ then $\rho_{ij} = \rho_{ji}$. That is, this matrix is symmetric and has ones on the main diagonal. Further, the correlation matrix is positive definite if all random variables have non-zero variance and no x_i has an exact relationship with another x_j, i.e. $\rho_{ij} \neq 1$, for $i \neq j$. This can easily be proved. First, consider a

vector $\vec{a}$ of real numbers. Then

$$\begin{aligned}
\vec{a}^T C \vec{a} &= \vec{a}^T E[(\vec{x} - E(\vec{x}))(\vec{x} - E(\vec{x}))^T]\vec{a} \\
&= E[[\vec{a}^T(\vec{x} - E(\vec{x}))][(\vec{x} - E(\vec{x}))^T\vec{a}]] \\
&= E[[\vec{a}^T(\vec{x} - E(\vec{x}))]^2] \\
&= E[[\vec{a}^T\vec{x} - E(\vec{a}^T\vec{x})]^2] \\
&= VAR[\vec{a}^T\vec{x}] \geq 0.
\end{aligned}$$

Further $VAR[\vec{a}^T\vec{x}] = 0$ implies

$$\sum_{i=1}^{n} a_i^2\sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=1}^{n} a_i a_j \sigma_i \sigma_j \rho_{ij} = \sum_{i=1}^{n} \alpha_i a_i^2 \sigma_i^2 + (\sum_{i=1}^{n} \beta_i a_i \sigma_i)^2 = 0$$

where $\alpha_i \geq 0$ and $\alpha_i + \beta_i^2 = 1$. Thus,

$$\sum_{i=1}^{n} \alpha_i^2 \sigma_i^2 = 0.$$

The non-zero variance implies $\sigma_i > 0 \quad i = 1, \ldots, n$. The condition $\rho_{ij} \neq 1$ implies $\alpha_i > 0 \quad i = 1, ..., n$. Thus, $\vec{a} = 0$, i.e. $\vec{a}^T C \vec{a} = 0$ implies $\vec{a} = 0$. One important consideration is that a correlation matrix and a computed set of marginal distributions cannot be randomly specified. If the marginal distributions are specified then restrictions exist on the correlation coefficients. The restrictions are defined as follows:

Theorem 2.12 *Let Y_1 and Y_2 be any two random variables with a joint probability distribution $H(y_1, y_2)$, then the correlation between Y_1 and Y_2 is defined by*

$$corr(Y_1, Y_2) = \frac{\int_0^\infty \int_0^\infty H(s,t)\, dt\, ds - \int_0^\infty F_1(s)\, ds \int_0^\infty F_2(t)\, dt}{\sigma_1 \sigma_2}$$

where $F_1(y_1)$ and $F_2(y_2)$ are cumulative probability distributions for Y_1 and Y_2 and σ_1 and σ_2 are the standard deviations of Y_1 and Y_2, respectively.

The maximum correlation between Y_1 and Y_2 is found by using the cumulative joint probability, $H^*(y_1, y_2) = min(F_1(y_1), F_2(y_2))$ and the minimum correlation found by using $H_*(y_1, y_2) = max(0, F_1(y_1) + F_2(y_2) - 1)$. Thus for a set of marginal distributions and a given set of correlation coefficients (under certain restrictions), a unique cumulative joint distribution exists. This uniquely specifies the conditions required to add random variables. Next, a method of incorporating the correlation coefficients will be discussed.

2.5 Approximating Solutions to Correlation

All four approximation methods of this study specify marginal distributions for each item and correlation coefficients between every item, i.e. the following problem is specified:

1.

random variable	marginal distribution
x_1	$f_1(x_1)$
x_2	$f_2(x_2)$
x_3	$f_3(x_3)$
$\vdots$	$\vdots$
x_n	$f_n(x_n)$

2. $corr(x_i, x_j) = \rho_{ij}$

Four methods of approximating correlation in risk analysis are described. The first method fits a beta distribution to data aggregated at the system level. The next two methods decompose the correlation matrix into LL^T. One of these methods uses a Cholesky decomposition and the other uses an eigenvalue decomposition. The final method is known as rank correlation and is available in a commercial risk software package known as Crystal Ball.

2.5.1 Beta Fit

The first cost estimating method considered uses a beta fit. The beta density function is a four parameter density function defined over the closed interval $a \leq y \leq b$. It requires a mean, variance, minimum, and maximum value. The following is a definition of the beta density function of the random variable Y.

Definition 2.13 *A random variable Y is said to have a beta probability distribution with parameters α and β if it has the density function*

$$f(y) = \begin{cases} \dfrac{(y-a)^{\alpha-1}(b-y)^{\beta-1}}{G(\alpha,\beta)(a+b)^{\alpha+\beta-1}} & 0 \leq y \leq 1 \\ 0 & elsewhere \end{cases}$$

where $G(\alpha,\beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,dy = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

The cumulative distribution function for the beta random variable is

$$F(y) = \int_0^y \frac{(t-a)^{\alpha-1}(b-t)^{\beta-1}}{[G(\alpha,\beta)](a+b)^{\alpha+\beta-1}}\,dt.$$

The parameters α and β can be found through the computations of

$$\alpha = \frac{-1}{\sigma^2(b-a)}(\mu-b)(\mu-a)^2 - \frac{(\mu-b)}{(b-a)},$$

$$\beta = -\alpha\left(\frac{\mu-b}{\mu-a}\right)$$

where a is the minimum, b is the maximum, μ is the mean (expected value) and σ^2 is the variance of the random variable. a, b, μ, σ^2 are the sum of the individual elements minimums, maximums, expected values, and variances, respectively. The

correlation comes into play since

$$VAR(\sum_{i=1}^{n} y_i) = \sum_{i=1}^{n} VAR(y_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \sigma_i \sigma_j Cov(y_i, y_j).$$

$G(\alpha, \beta)$ and $\int_0^y (t-a)^{\alpha-1}(b-t)^{\beta-1}\,dt$ are approximated using a numerical scheme. In this study, the numerical scheme used is contained in a mathematical software package, Matlab.

2.5.2 Choleski Method

The Choleski method factors the correlation matrix $C = LL^T$ where L is a lower triangular matrix. This factorization is possible if the original matrix is positive definite which is true for correlation matrices, if properly specified.

The following algorithm generates the factor L in the factorization $C = LL^T$. Set $l_{1,1} = \sqrt{a_{1,1}}$ and for $i = 2, \ldots, n$ set $l_{i,1} = a_{i,1}$. For $k = 2, \ldots, n$ set

$$l_{k,k} = (a_{k,k} - \sum_{l=1}^{k-1} |l_{k,l}|^2)^{\frac{1}{2}}$$

and for $i = k+1, \ldots, n$ set

$$l_{i,k} = \frac{1}{l_{k,k}}(a_{i,k} - \sum_{l=1}^{k-1} l_{i,l} l_{k,l}).$$

For further information concerning this algorithm, refer to [4].

Once the correlation matrix is decomposed, independent random draws,

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

are taken from the marginal distributions. These draws are then normalized such that each random variable has a mean equal to zero and a variance equal to one. This set of random draws are then transformed as follows

$$\vec{z} = L\vec{x}.$$

This process was presented by Book and Young at the 24th annual Department of Defense Cost Symposium [5].

Theorem 2.14 *The random vector $\vec{z}$ has the following properties:*

$$\begin{array}{ll} E[z_i] = 0 & \\ Var[z_i] = 1 & i, j = 1, \ldots, n. \\ Corr(z_i, z_j) = \rho_{ij} & \end{array}$$

where ρ_{ij} is the correlation between i and j in the original correlation matrix, C.

Proof:

Since $\vec{x}$ is normalized, $E[x_i] = 0$, for $i = 1, \ldots, n$. Thus,

$$E[z_i] = E[\sum L_{ij} x_j] = \sum L_{ij} E[x_j] = 0.$$

Let $C_{\vec{z}}$ = covariance matrix of vector $\vec{z}$ and let $C_{\vec{x}}$ = covariance matrix of vector $\vec{x}$ which is equal to the identity matrix. (Since $\vec{x}$ are independent draws and $Var(x_i) =$

$1, i = 1, \ldots, n.$) Thus,

$$\begin{aligned} C_{\vec{z}} &= E[\vec{z}\vec{z}^T] - E[\vec{z}]E[\vec{z}^T] \\ &= E[\vec{z}\vec{z}^T] \\ &= E[L\vec{x}\vec{x}^T\, L^T] \\ &= LE[\vec{x}\vec{x}^T]L^T. \end{aligned}$$

We have

$$E[\vec{x}\vec{x}^T] = C_{\vec{x}} - E[\vec{x}]E[\vec{x}^T] = I - 0 = I.$$

Thus,

$$C_{\vec{z}} = LL^T = C.$$

This implies

$$(C_{\vec{z}})_{ii} = 1$$

or,

$$Var(z_i) = 1,\ i = 1, \ldots, n$$

and

$$\sigma_{z_i} = 1,\ i = 1, \ldots, n.$$

Consequently,

$$Corr(z_i, z_j) = \frac{(C_{\vec{z}})_{ij}}{\sigma_{z_i}\sigma_{z_j}} = (C_{\vec{z}})_{ij} = \rho_{ij}.$$

Thus, the correlation structure of the system is maintained in the transformation $L\vec{x}$. Note that the above proof holds not only for a Choleski factorization, but for any square root factorization of the form LL^T.

2.5.3 Eigenvalue Method

The eigenvalue method factors the correlation matrix into three matrices, $C = U\Omega U^T$ where Ω is a diagonal $n \times n$ matrix that contains the eigenvalues of the correlation matrix. U is an $n \times n$ matrix whose columns contain the eigenvectors of the correlation matrix. The eigenvectors in U correspond to the eigenvalues of in Ω, i.e,

$$\Omega = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

and

$$U = \begin{bmatrix} v_1 & v_2 & \dots & v_n \\ & & & \end{bmatrix}$$

where the v_i's are the eigenvectors that corresponds to the eigenvalues λ_i for $i = 1, \dots, n$.

Many methods exist for computing the eigenvalues and eigenvectors of a matrix. Most of these methods are based on a numerical method called a QR factorization. The QR method along with its algorithm can be found in *Matrix Computations* by Golub and Van Loan [6].

As with the Choleski method, the eigenvalue method uses a transformation $\vec{z} = L\vec{x}$ where

$$L = U\Omega^{\frac{1}{2}}$$

The correlation structure is also maintained using this method as the previous proof

showed.

Although both methods preserve the original correlation matrix, the Choleski method does not preserve the order of the components while the eigenvalue method does. That is, the order of the random variables $x_1, x_2, \ldots, x_n$ does not matter when using the eigenvalue method. To prove this, a permutation matrix and its properties are defined.

Definition 2.15 *An $n \times n$ permutation matrix is a matrix whose columns consist of a rearrangement of the n unit vectors $e^{(j)}$, $j = 1, \ldots, n$ in $\mathcal{R}^n$, i.e. a rearrangement of the columns of the $n \times n$ identity matrix.*

The effect of premultiplying (postmultiplying) a matrix A by a permutation matrix P is to rearrange the rows (columns) of A into the same order as the rows (columns) of the identity matrix are ordered in P. Properties of permutations are $P(j,k) = P^T(j,k) = P^{-1}(j,k)$. If the system's elements are arranged in any order then the correlation matrix must change to fit the corresponding order of the elements, i.e. for any change we will do the following:

1. interchange k, l
2. $C_{k,j} \rightarrow C_{l,j}$ and $C_{i,k} \rightarrow C_{i,l}$
3. $C_{l,j} \rightarrow C_{k,j}$ and $C_{i,l} \rightarrow C_{i,k}$

Interchanging the rows and columns k and l of the orignal matrix C_{ij} produces a new correlation matrix $\hat{C}$.

Theorem 2.16 *The eigenvalue method is order independent.*

Proof: Recall that a correlation matrix has the following decompostion

$$C = U\Omega U^T$$

where Ω = eigenvalue matrix and U = corresponding eigenvector matrix. Further, the vector $\vec{z}$,

$$\vec{z} = U\Omega^{\frac{1}{2}}\vec{x},$$

contains the correlated random variables and $\vec{x}$ contains independent random variables. Suppse $\hat{C}$ is a matrix derived from interchanging a row and column of C. Then,

$$\begin{aligned} \hat{C} &= PCP \\ &= PU\Omega U^T P \\ &= PUPP\Omega PPU^T P \\ &= (PUP)(P\Omega P)(PU^T P). \end{aligned}$$

Let $\hat{U}= PUP$ and $\hat{\Omega}= P\Omega P$. It will be shown that $\hat{\Omega}$ is the eigenvalue matrix of $\hat{C}$ with $\hat{U}$ as its corresponding eigenvector matrix. Since $CU = U\Omega$, the following equalities hold,

$$\begin{aligned} \hat{C}\hat{U} &= PCPPUP \\ &= PCUP \\ &= PU\Omega P \\ &= PUPP\Omega P \\ &= \hat{U}\hat{\Omega}\,. \end{aligned}$$

Thus $\hat{C}$ has the eigenvalue decomposition

$$\hat{C}=\hat{U}\hat{\Omega}\hat{U}^T\,.$$

The result now follows. Let $\hat{x} = P\vec{x}$. This implies

$$\begin{aligned}
\hat{z} &= \hat{U}\hat{\Omega}^{\frac{1}{2}}\hat{x} \\
&= PUPP\Omega^{\frac{1}{2}}P\,\hat{x} \\
&= PU\Omega^{\frac{1}{2}}P\,\hat{x} \\
&= PU\Omega^{\frac{1}{2}}\vec{x} \\
&= P\vec{z}.
\end{aligned}$$

For the Choleski method, the proof fails at the following place,

$$C = LL^T$$

implies

$$\begin{aligned}
\hat{C} &= PCP \\
&= PLL^TP \\
&= PLPPL^TP \\
&= \hat{L}\hat{L}^T .
\end{aligned}$$

But $\hat{L} = PLP$ is not lower triangular. Thus $\hat{C}$ has a Choleski decomposition different than the decomposition above.

2.5.4 Rank Correlation

The final method discussed is known as rank correlation. This method considers correlation differently than discussed previously.

As an example of rank correlation, consider k observations of random variables X and Y. These observations form pairs $(x_i, y_i), i = 1, \ldots, k$. Further, suppose the observations are ranked according to magnitude. For example, if the third observation of $\vec{x}$ is the lowest of all observations of $\vec{x}$, then the rank of x_3 is 1. Thus,

the pairs (x_i, y_i) are transformed into the pairs (r_i, s_i) where $r_i = rank(x_i)$ and $s_i = rank(y_i)$.

The ordinary sample correlation coefficients can then be computed for the k pairs of ranks (r_i, s_i), $i = 1, \ldots, k$. The following statistic is known as the Spearman's Rank correlation coefficient:

$$r_3 = \frac{\sum_{i=1}^{k}(r_i - \overline{r})(s_i - \overline{s})}{\sqrt{\sum_{i=1}^{k}(r_i - \overline{r})^2 \sum_{i=1}^{k}(s_i - \overline{s})^2}}.$$

After numerous calculations r_3 becomes

$$r_3 = 1 - \frac{6\sum_{i=1}^{k} d_i^2}{\mu(\mu^2 - 1)}$$

where $d_i = r_i - s_i$, $i = 1, \ldots, k$.

This method is implemented by reodering the rank draws of individual items using a Choleski decomposition to obtain Spearman's Rank correlations, which are the same as the problem's original correlation matrix. This method is also available in the commercial risk software package Crystal Ball and is discussed in [7]. The next section compares the results of these four approximating methods for a few examples. Also, results from an algorithm which computes maximum and minimum correlations between elements is given.

Chapter 3

Results

This section gives results which compare the four different correlation methods using a twelve element test case, a 30 element example, and a three cost component example. Examples of maximum and minimum correlations are also included along with the advantages and disadvantages of each methodology.

3.1 Test Case

A test case with a theoretical solution was constructed as a means for measuring the accuracy of the different approximating methods. The theoretical problem was constructed for twelve components as follows. First, it is assumed that the system has the following marginal distributions

$$
\begin{aligned}
f_1(x_1) &= 2x_1 \quad 0 \le x_1 \le 1 \\
f_2(x_2) &= \tfrac{1}{2}x_2 \quad 0 \le x_2 \le 2 \\
f_3(x_3) &= \tfrac{2}{9}x_3 \quad 0 \le x_3 \le 3 \\
f_{3i+1} &= f_1(x) \quad i = 1,2,3 \\
f_{3i+2} &= f_2(x) \quad i = 1,2,3 \\
f_{3i+3} &= f_3(x) \quad i = 1,2,3
\end{aligned}
$$

and the correlations

$$\begin{aligned} Corr(x_i, x_j) &= 1 \quad i = j \\ Corr(x_i, x_j) &= .9 \quad i \neq j. \end{aligned}$$

The corresponding cumulative distributions are

$$\begin{aligned} F_1(x_1) &= x_1^2 \\ F_2(x_2) &= \frac{x_2^2}{4} \\ F_3(x_3) &= \frac{x_3^2}{9} \\ F_{3i+1}(x) &= F_1(x) \quad i = 1, 2, 3 \\ F_{3i+2}(x) &= F_2(x) \quad i = 1, 2, 3 \\ F_{3i+3}(x) &= F_3(x) \quad i = 1, 2, 3. \end{aligned}$$

The following joint distribution has the above properties.

$$H(x_1, \ldots, x_{12}) = .1H_0(x_1, \ldots, x_{12}) + .9H^*(x_1, \ldots, x_{12})$$

where

$$H_0(x_1, \ldots, x_{12}) = \Pi_{i=1}^{12} F_i(x_i)$$

and

$$H^*(x_1, \ldots, x_{12}) = min\{F_1(x_1), \ldots, F_{12}(x_{12})\}.$$

Note that both $H_0(x_1, \ldots, x_{12})$ and $H^*(x_1, \ldots, x_{12})$ have the above marginal distributions. Thus, $H(x_1, \ldots, x_{12})$ will also have the above marginals. Further, for the joint distributions $H_0(x_1, \ldots, x_{12})$,

$$Corr(x_i, x_j) = 0 \ \ if \ \ i \neq j$$

and for $H^*(x_1, \ldots, x_{12})$,

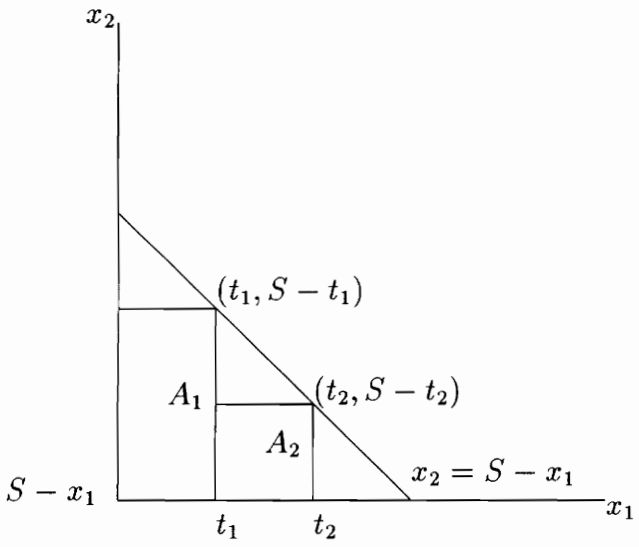
$$Corr(x_1, \ldots, x_{12}) = 1 \quad if \quad i \neq j.$$

Consequently, for the joint distribution $H(x_1, \ldots, x_{12})$

$$\begin{aligned}
Corr(x_i, x_j) &= \frac{\int \cdots \int H \, dx_1 \cdots dx_{12} - \int F_i \, dx_i \int G \, dx_i}{\sigma_i \sigma_j} \\
&= \frac{\int \cdots \int (.1 H_0 + .9 H^*) \, dx_1 \cdots dx_{12}}{\sigma_i \sigma_j} \\
&\quad \frac{-.1 \int F_i \, dx_i \int G_j \, dx_j - .9 \int F_i \, dx_i \int G_j \, dx_j}{\sigma_i \sigma_j} \\
&= .1 \frac{\int \cdots \int H_0 \, dx_1 \cdots dx_{12} - \int F_i \, dx_i \int G_j \, dx_j}{\sigma_i \sigma_j} \\
&\quad + .9 \frac{\int \cdots \int H^* \, dx_1 \cdots dx_{12} - \int F_i \, dx_i \int G_j dx_j}{\sigma_i \sigma_j} \\
&= .1(0) + .9(1) = .9
\end{aligned}$$

The above expression for correlation can be found in [8].

Now to determine $P(x_1 + \ldots + x_{12} \leq S)$, the simple case with only x_1 and x_2 is considered. Graphically, the solution to this problem is the area under the curve



where $A_1 = H(t_1, S - t_1)$ and $A_2 = H(t_2, S - t_2) - H(t_1, S - t_2)$.

Total area under the curve $\approx H(t_1, S - t_1) + H(t_2, S - t_2) - H(t_1, S - t_2)$. Thus the limiting case gives

$$AREA = \int_0^S \frac{\partial}{\partial u_1} H(u_1, u_2) \Bigg|_{\substack{u_1 = t \\ u_2 = S - t}} dt.$$

Next, the problem is expanded to three variables. First, let

$$A(S, x_3) = \int_0^S \frac{\partial}{\partial u_1} H(u_1, u_2, u_3) \Bigg|_{\substack{u_1 = t \\ u_2 = S - t \\ u_3 = x_3}} dt.$$

Then,

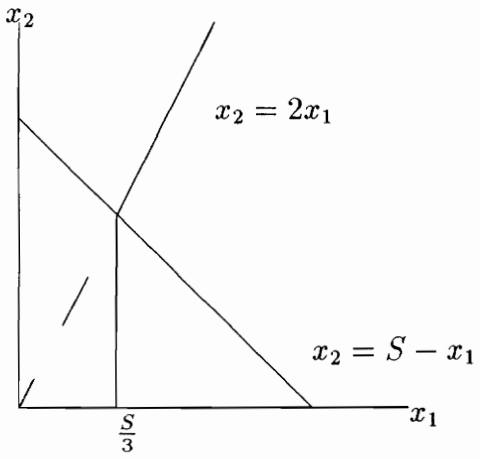
$$P(x_1 + x_2 + x_3 \leq S) = \int_0^S \frac{\partial}{\partial u_2} A(u_1, u_2) \Big|_{\substack{u_1 = S - x_3 \\ u_2 = x_3}} dx_3.$$

This derivation is used for

$$H^*(x_1, x_2, x_3) = min\{F_1(x_1), F_2(x_2), F_3(x_3)\}.$$

First consider

$$\begin{aligned} H^*(x_1, x_2) &= min\{F_1(x_1), F_2(x_2)\} \\ &= \begin{cases} x_1^2 & if\ x_1 \leq \frac{x_2}{2} \\ \frac{x_2^2}{4} & if\ x_1 > \frac{x_2}{2}. \end{cases} \end{aligned}$$



This gives

$$\begin{aligned} A(S) &= \int_0^3 \frac{\partial}{\partial u_1} H^*(u_1, u_2) \Big|_{\substack{u_1 = t \\ u_2 = S - t}} dt \\ &= \int_0^{\frac{S}{3}} 2\, u_1 \Big|_{\substack{u_1 = t \\ u_2 = S - t}} dt \\ &= \int_0^{\frac{S}{3}} 2t\, dt \\ &= t^2 \Big|_0^{\frac{S}{3}} \\ &= \frac{S^2}{9}. \end{aligned}$$

Thus, for three variables,

$$A(S, x_3) = \int_0^S \frac{\partial}{\partial u_1} H^*(u_1, u_2, u_3) \Big|_{\substack{u_1 = t \\ u_2 = S - t \\ u_3 = x_3}} dt.$$

Note that

$$x_1^2 \leq \frac{x_3^2}{9} \implies x_1 \leq \frac{x_3}{3}.$$

This gives

$$
\begin{aligned}
A(S,x_3) &= \int_0^{\frac{S}{3}} 2\,u_1 \Big|_{\substack{u_1=t\\ u_2=S-t\\ u_3=x_3}} dt \quad if\ x_3>3\\
&= \int_0^{\frac{x_3}{3}} 2u_1 \Big|_{\substack{u_1=t\\ u_2=S-t\\ u_3=x_3}} dt \quad if\ x_3\le 3\\
&= \begin{cases} \frac{S^2}{9} & if\ x_3>3\\ \\ \frac{x_3^2}{9} & if\ x_3\le 3. \end{cases}
\end{aligned}
$$

Therefore,

$$
P(x_1+x_2+x_3\le S)=\int_0^S \frac{\partial}{\partial u_2}A(u_1,u_2)\Big|_{\substack{u_1=S-x_3\\ u_2=x_3}} dx_3
$$

But $x_3 \le S-x_3$ implies $x_3 \le \frac{S}{2}$ consequently,

$$
\begin{aligned}
P(x_1+x_2+x_3\le S) &= \int_0^{\frac{S}{2}} \frac{2u_2}{9}\Big|_{u_2=t}\, dt\\
&= \tfrac{2}{9}\int_0^{\frac{S}{3}} t\,dt\\
&= \tfrac{2}{9}\frac{t^2}{2}\Big|_0^{\frac{S}{2}}\\
&= \frac{S^2}{36}.
\end{aligned}
$$

This process can be continued for all twelve components to give

$$P(x_1 + x_2 + \ldots + x_{12} \leq S) = \frac{S^2}{(24)^2} = \frac{S^2}{576}.$$

The 24 in the denominator is the sum of the twelve right endpoints for the distributions.

The closed form solution for $H_0(x_1, \ldots, x_{12})$ is very complicated to compute but is very easily simulated since this joint distribution assumes independence among the components. Consequently, using 10,000 interations, a distribution for $H_0(x_1, x_2, x_3)$ was simulated. A linear combination of this distribution with the closed form solution for $H^*(x_1, \ldots, x_{12})$ was then computed.

The results of the test cases are contained in the following four tables. The first table compares all four correlation methods to the actual output computed for the test case. The next two tables compare the Choleski method and eigenvalue method using four different orders of the twelve components.

The last table uses a max norm to compare the results. These norms are computed as follows. Let $T(x) = P(S_1 + \cdots + S_n \leq x)$ in the test case. Let $A(x) = P(S_1 + \cdots + S_n \leq x)$ in the approximating case. Then the $5 - 95\%$ norm is defined as

$$|x - \overline{x}|_{5/95} = max_{x,\overline{x}} \frac{|x - \overline{x}|}{|\overline{x}|}$$

where $x, \overline{x}$ are such that $.5 \leq T(x) = A(\overline{x}) \leq .95$ and the $20 - 80\%$ norm is defined as

$$|x - \overline{x}|_{20/80} = max_{x,\overline{x}} \frac{|x - \overline{x}|}{|\overline{x}|}$$

where $x, \overline{x}$ are such that $.20 \leq T(x) = A(\overline{x}) \leq .80$

Table 3.1: Table 1 Risk Methods with Direct Order

Percentile	Test case actual	Beta	Rank Correlation	Choleski method	Eigenvalue method
5	5.7	5.9	6.0	6.1	5.7
10	8.0	8.1	8.1	8.2	7.9
15	9.8	9.7	9.7	9.8	9.5
20	11.3	11.0	11.0	11.1	10.8
25	12.6	12.2	12.3	12.3	12.0
30	13.6	13.3	13.3	13.3	13.2
35	14.5	14.3	14.3	14.2	14.3
40	15.3	15.2	15.2	15.1	15.3
45	16.0	16.0	16.0	16.0	16.1
50	16.7	16.9	16.8	16.8	17.0
55	17.4	17.6	17.6	17.6	17.6
60	18.1	18.4	18.4	18.3	18.4
65	18.8	19.1	19.1	18.9	19.2
70	19.6	19.8	19.7	19.6	19.9
75	20.4	20.5	20.4	20.2	20.6
80	21.2	21.2	21.1	20.9	21.3
85	21.9	21.9	21.8	21.6	21.9
90	22.6	22.5	22.5	22.4	22.4
95	23.3	23.2	23.1	23.4	23.1

Table 3.2: Table 2 Choleski Method with Different Orders

Percentile	Direct	Reverse	Order #1	Order #2
5	6.1	5.9	6.0	6.0
10	8.2	8.0	8.1	8.0
15	9.8	9.6	9.7	9.7
20	11.1	11.0	11.1	11.0
25	12.3	12.1	12.3	12.1
30	13.3	13.3	13.4	13.2
35	14.2	14.2	14.3	14.2
40	15.1	15.2	15.3	15.1
45	16.0	16.0	16.2	15.9
50	16.8	16.8	17.0	16.8
55	17.6	17.5	17.7	17.5
60	18.3	18.3	18.4	18.2
65	18.9	19.0	19.1	18.8
70	19.6	19.7	19.8	19.5
75	20.2	20.4	20.5	20.2
80	20.9	21.0	21.1	20.9
85	21.6	21.7	21.7	21.7
90	22.4	22.5	22.5	22.4
95	23.4	23.4	23.4	23.5

Table 3.3: Table 3 Eigenvalue Method with Different Orders

Percentile	Direct	Reverse	Order #1	Order #2
5	5.7	5.8	5.6	5.7
10	7.9	8.0	7.9	7.8
15	9.5	9.6	9.6	9.5
20	10.8	10.9	10.9	10.9
25	12.0	12.2	12.1	12.1
30	13.2	13.3	13.2	13.2
35	14.3	14.3	14.2	14.2
40	15.3	15.2	15.1	15.2
45	16.1	16.2	16.0	16.1
50	17.0	16.9	16.9	17.0
55	17.6	17.7	17.7	17.7
60	18.4	18.4	18.4	18.5
65	19.2	19.2	19.1	19.2
70	19.9	19.8	19.8	19.9
75	20.6	20.5	20.5	20.6
80	21.3	21.2	21.1	21.2
85	21.9	21.8	21.8	21.8
90	22.4	22.5	22.4	22.4
95	23.1	23.1	23.0	23.1

Table 3.4: Table 4 Accuracy Based on Maximum Norms

Risk Method	$\triangle(5\% - 95\%)$	$\triangle(20\% - 80\%)$
Rank Correlation	5.3	2.7
Beta	3.7	3.0
Choleski (direct)	4.0	4.0
Eigenvalue (direct)	4.5	4.5
Choleski (reverse)	6.9	2.4
Eigenvalue (reverse)	3.4	3.4
Choleski (order #1)	5.1	2.8
Eigenvalue (order #1)	4.0	4.0
Choleski (order #2)	5.3	3.6
Eigenvalue (order #2)	4.0	4.0

where the orders are of the form

Order	Items
direct	1,2,...,12
reverse	12,11,...,1
#1	3,6,9,12,2,5,8,11,1,4,7,10
#2	1,4,7,10,2,5,8,11,3,6,9,12

The results in Table 4 depict the accuracy of each method. Overall the rank correlation method produced the best results. All four methods, however, provided reasonably good results especially in the 20-80 percentile range where all methods gave results accurate to five percent. In reverse order, the Choleski method produce the best output. Otherwise, the Choleski produced worse results for other orders. As proved earlier, the order of the components does not matter when using the eigenvalue method. The minor differences in the eigenvalue results in Table 3 are due to round off errors. In all methods, the results increased in accuracy towards the middle of the distributions.

3.2 30 Component Project

The larger system consisting of thirty components with the following triangular distributions and correlation matrix was also considered:

item	1	2	3	4	5	6	7	8	9	10
min	0	1	10	15	23	50	60	75	88	97
mode	.3	4	11	18	33	51	65	80	90	99
max	1	10	15	25	50	59	74	85	100	110

Items 11-20 and 21-30 use the same distributions as items 1-10, i.e.,

$$\begin{array}{ccccc} \text{item 1 distribution} & = & \text{item 11 distribution} & = & \text{item 21 distribution} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{item 10 distribution} & = & \text{item 20 distribution} & = & \text{item 30 distribution} \end{array}$$

The correlation matrix is symmetric with ones on the diagonal entries. All other entries are determined in this manner:

1. First row of 30x30 matrix

   $c_{1,1} = 1.00 \Longrightarrow c_{1,j} = c_{1,j-i} - .02$ for $j = 2, 3, \ldots, 30$

   Thus

   $$c_{1,2} = .98$$
   $$c_{1,3} = .92$$
   $$\vdots$$
   $$c_{1,30} = .42$$

2. Second row of 30x30 matrix

   $c_{2,2} = 1.00 \Longrightarrow c_{2,j} = c_{2,j-i} - .02$ for $j = 1, 2, \ldots, 30$

   Thus

   $$c_{2,2} = .98$$
   $$\vdots$$
   $$c_{2,30} = .44$$

3. Third row

   $\vdots$

4. 29th row

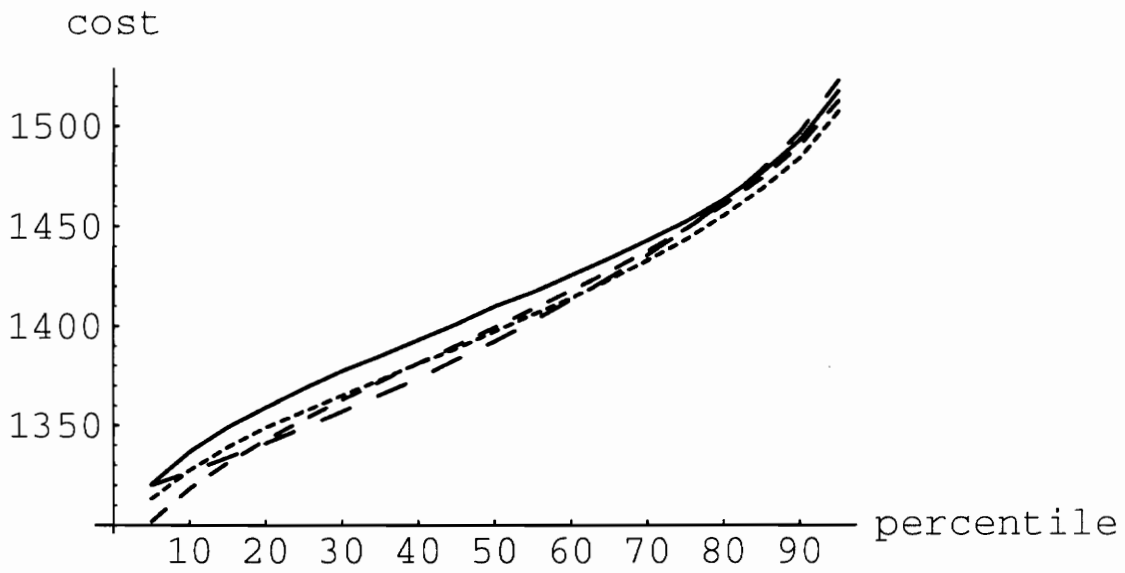
   $c_{29,29} = 1.00 \Longrightarrow c_{29,30} = c_{29,29} - .02 = .98$

5. 30th row

   $c_{30,30} = 1.00$

The results of the 30 component case are given in two graphs. Each graph contains all four correlation methods, the rank correlation, beta, Choleski, and eigenvalue. One graph contains results for the Choleski method in direct order and the

other is for reverse order.



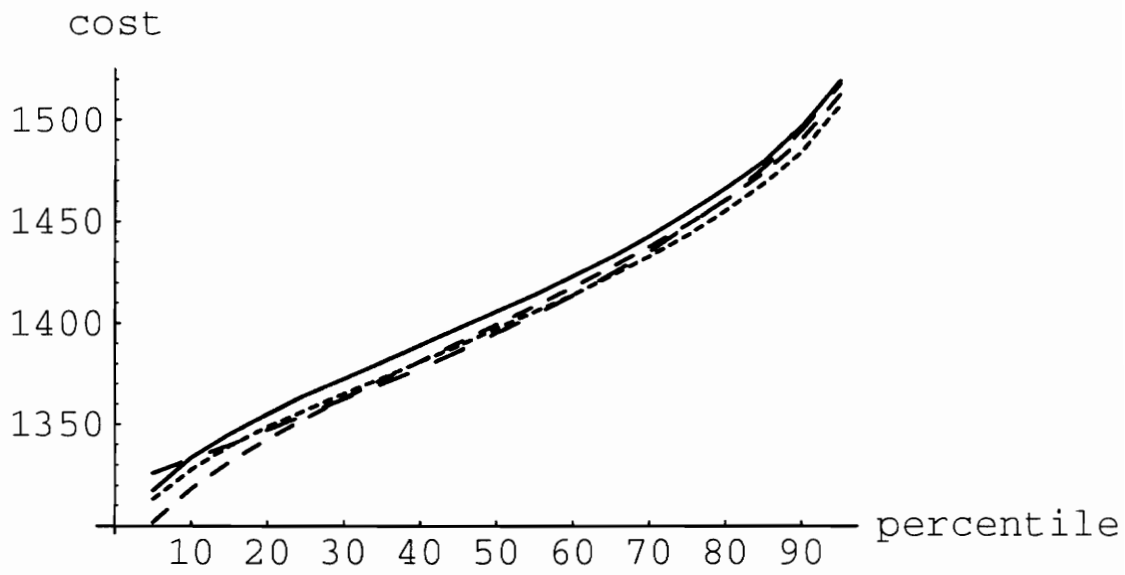
- - - - - - Rank Correlation

- - - - - - - - - - - - Beta

— — — — — Eigenvalue

————————Choleski

Figure 3.1: Direct Order



- - - - - - Rank Correlation

- - - - - - - - - - - - Beta

— — — — — Eigenvalue

————————Choleski

Figure 3.2: Reverse Order

Because of its accuracy in the first example, the output for the rank correlation method is used as the control method. In both graphs, the Choleski method is farthest from the rank correlation method. Comparing the beta and eigenvalue methods at the 50th and 70th percentiles of each graph, the beta method is closer to the control method at the 50th percentile but the eigenvalue method is closer at the 70th percentile. Overall, all four methods again appeared to be reasonably close to each other. The last example involves three components with different orders of magnitude.

3.3 Three Component Case

The following marginal distributions and correlations used are given along with the results in table form:

item	1	2	3
min	0	0	0
mode	1	10	100
max	1	10	100

and the correlations are

$$corr(x_i, x_j) = 1 \ \ if \ \ i = j$$

and

$$corr(x_i, x_j) = .9 \ \ if \ \ i \neq j.$$

Table 3.5: Table 5 Risk Methods and their Accuracy

	Test	Rank	Beta	Choleski	Choleski	Eigenvalue
Prob	Case	Correlation		Direct	Reverse	
0.05	25.33	26.46	30.00	55.19	6.43	25.88
0.10	35.48	36.37	39.59	59.17	21.56	36.11
0.15	43.32	44.18	46.65	62.30	32.01	43.03
0.20	49.86	50.46	52.49	64.88	40.60	49.63
0.25	55.64	56.24	57.57	67.10	47.74	55.25
0.30	60.82	61.28	62.13	68.96	54.26	60.52
0.35	65.64	66.04	66.33	70.68	60.06	65.46
0.40	70.18	70.18	70.23	72.42	65.24	70.07
0.45	74.38	74.13	73.92	74.02	71.06	74.04
0.50	78.38	78.07	77.42	75.59	76.27	77.89
0.55	82.16	81.52	80.79	77.00	81.30	81.64
0.60	85.78	85.37	84.05	78.29	86.31	85.09
0.65	89.22	88.58	87.23	79.62	91.25	88.44
0.70	92.57	91.94	90.35	80.93	96.47	91.72
0.75	95.79	95.72	93.44	82.20	102.23	94.68
0.80	98.92	98.85	96.52	83.38	108.34	97.97
0.85	101.98	101.95	99.65	84.53	115.29	101.31
0.90	104.95	104.97	102.88	85.59	123.54	104.83
0.95	107.96	107.96	106.35	86.96	134.48	109.02
5/95 Norm		0.0445	0.1841	1.1785	0.7462	0.0216
20/80 Norm		0.0120	0.0526	0.3011	0.1858	0.0116
5/95 % Norm		4.45%	18.41%	117.85%	74.62%	2.16%
20/80		1.20%	5.25%	30.11%	18.58%	1.16%

Unlike the previous two examples, the different methods in this example gave some very different and inaccurate results. The rank correlation and eigenvalue methods did not appear to be affected by the different orders of magnitude. The beta method, on the other hand, did not give better results than the rank correlation

and eigenvalue methods while the Choleski method gave the worst results. As can be seen by the overall norms, the order made a big difference in the results of the Choleski method. All methods, however, gave reasonably good estimates at the 50th percentile level.

3.4 Limitations on Correlation Coefficients

As mentioned earlier, given a set of marginal distributions, limitations exist on their correlation matrix. Examples of marginal distributions and corresponding maximum and minimum correlations are given below:

item	marginal distribution
x_1	$f_1(x_1) = 2x_1$
x_2	$f_2(x_2) = 3x_2^2$
x_3	$f_3(x_3) = 5x_3^4$
x_4	$f_4(x_4) = 10x_4^9$
x_5	$f_5(x_5) = 20x_5^{19}$

and

$Corr(x_i, x_j)$	min	max
(x_1, x_2)	-.901	.996
(x_1, x_3)	-.867	.984
(x_1, x_4)	-.834	.968
(x_1, x_5)	-.830	.957

Each example has demonstrated varying degrees of accuracy of the four correlation methods. Each method has advantges and disadvantages. Overall, the rank correlation method produced the most accurate results. The beta method produced fairly good results. Both of these methods are also commercially available. Unfortunately, the beta method lacks the theory to back up its good results and was impacted somewhat by different orders of magnitudes of components. The eigenvalue method results are good and this method preserves the order. Based on the three component case, the eigenvalue and rank correlation methods were the only methods not affected by the orders of magnitude of the components. Both these two methods are expensive, especially the eigenvalue method. Its factorization is based on a numerical method called a QR factorization and requires $\frac{4}{3}n^3$ multipli-

cations and the like number of additions. The Choleski method is also expensive, but its factorization has $\frac{n^3}{6}$ multiplications and the like number of additions. These operation counts are summarized in table 6. Further, these advantages and disadvantages, including a few more, are in table 7.

Table 3.6: Table 6: Operation Counts for Different Correlation Methods

Method	Order of number of additions/multiplications
Beta	$\frac{n^2}{2}$
Rank Correlation	$\frac{1}{6}n^3$
Choleski	$\frac{1}{6}n^3$
Eigenvalue	$\frac{4}{3}n^3$

Table 3.7: Table 7 Advantages and Disadvantages

Method	Advantages	Disadvantages
Rank Correlation	1. Gives best overall results 2. Do not need to know the distribution types 3. Commercially available in risk software (Crystal Ball)	1. Expensive 2. Order dependent
Beta	1. Cheap 2. Gives closed form solution 3. Commercialy available (spreadsheet)	1. Correlation matrix (if not positive definite) is not a true correlation matrix 2. No theory to back up the "good" results 3. Dependent somewhat on different orders of magnitude of components
Choleski	1. Available in standard software packages such as Matlab	1. Order dependent 2. Expensive 3. Dependent on orders of magnitude of components
Eigenvalue	1. Order independent 2. Not dependent on order of magnitude of components	1. Powerful P.C. needed 2. Very expensive

Chapter 4

Conclusion

The preceding discussion has highlighted four different approximating risk correlation methods that quantify uncertainty inherent in cost analysis. Next, each was applied and compared to a developed test case, and applied to a cost project decomposed into 30 elements and one with three elements. These were followed by a theoretical discussion with a few examples of restrictions to correlations. Finally, the advantages and disadvantages of each method were discussed.

In general, someone wishing to choose a correlation method should consider various issues. Since the rank correlation is available in a commercial software package known as Crystal Ball, a user who does not wish to do any additional programming may use this method. For projects with many cost components, however, it is very tedious to input the correlations. A person who does not wish to purchase this additional software and has a standard spreadsheet, may prefer to fit a beta distribution. The user, however, must be careful to insure that their correlation matrix is positive definite since unlike other methods, the beta method will not flag this. A person with some programming capabilities and with some access to software packages such as Matlab may choose the Choleski method. This should only be done, however, if the components have the same magnitude. Further, if the person has the capabil-

ity to use rank correlation, this research has shown that this method has definite advantages and is less restrictive than the Choleski method. Finally, if a user must choose between the Choleski and eigenvalue methods, and no time constraints exist, then the eigenvalue method is preferred even though the Choleski method is eight times quicker. The Choleski method would be preferred if quick turnarounds are crucial, although this method has the potential to give less than accurate solutions while the eigenvalue method is more consistent in its capability to compute accurate solutions. Further, the eigenvalue method is order-independent while the Choleski method is not.

Further research areas exist in risk correlation. First, additional research on the development of correlation coefficients is required. Although this study has shown that accurate risk correltion methods exist, in practice it is extremely difficult to estimate good correlation coefficients. Second, cases where one item is dependent on another rather than simply being correlated is needed. With the analysis of the results of these areas, risk methodology can be more complete in quantifying uncertainty inherent in a system.

Bibliography

[1] W. Mendenhall, D. Wackerly, R. Scheaffer, *Mathematical Statistics with Applications*, PWS-KENT Publishing Company, Boston, MA, 1990.

[2] M. Degroot, *Probability and Statistics*, Addison-Wesley Publishing Company, Menlo Park, CA, 1975.

[3] S. Gupta, D. Olsen, D. Hudak, J. Keenan, *Cost Risk Analysis of the Strategic Defense System*, The Analytic Science Corporation, Arlington, VA, 1992.

[4] M. Gunzburger, *Topics in Applied Math*, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1992.

[5] S. Book, P. Young, *Monte-Carlo Generation of Total Cost Distributions when WBS-Element Costs are Correlated*, The Aerospace Corporation, Los Angeles, CA, 1990.

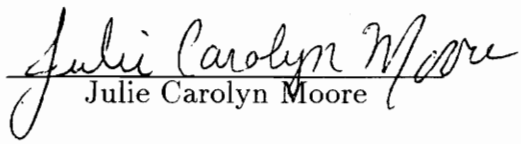
[6] G. Golub, C. Van Loan, *Matrix Computations*, John Hopkins University Press, Baltimore, MD, 1989.

[7] R. L. Iman and W. J. Conover, "A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables.",*Communications in Statistics*,Vol. B11, No.3, 1982.

[8] "Bivariate Distributions with Given Marginals", *The Annals of Statistics*, Ward Whitt, Vol. 4, 1976.

Vita

Julie Carolyn Moore was born in Alexandria, VA and attended Thomas Edison High School. This is where she developed her love for mathematics through the guidance of Andy Blount, her math teacher. She received her Bachelor's and Master's Degree at Virginia Polytechnic and State University where she not only met John but found math to be fun and frustrating. So now she needs to move on.


Julie Carolyn Moore