

AD-R155 239

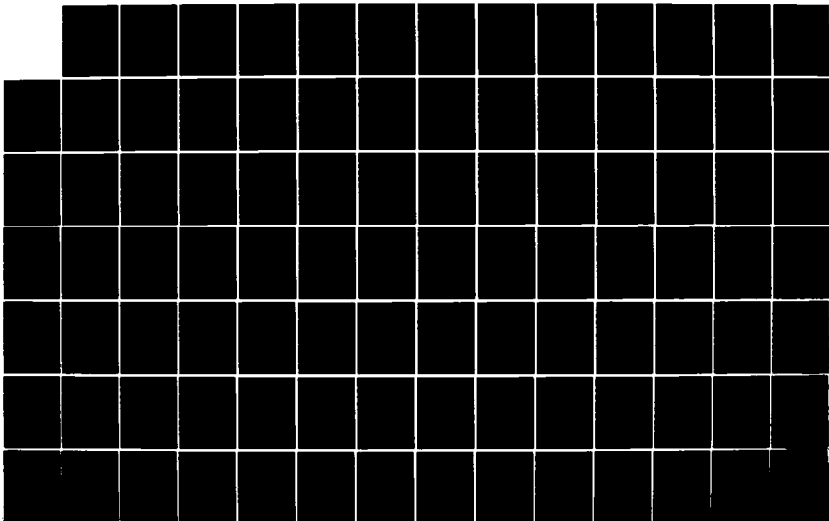
APPLICATIONS OF CORRELATION TECHNIQUES FOR BATTLEFIELD
IDENTIFICATION I(U) JET PROPULSION LAB PASADENA CA
D E HOCHMAN ET AL. 30 JUN 84 JPL-D-179 NAS7-918

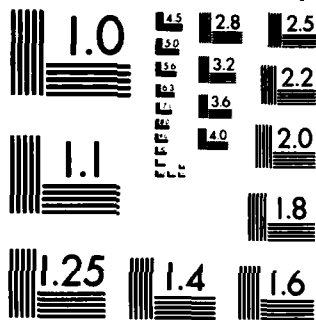
1/2

UNCLASSIFIED

F/G 12/1

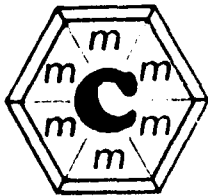
NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

D-179



THE MATHEMATICS CLINIC

AD-A155 239

APPLICATIONS OF CORRELATION TECHNIQUES FOR BATTLEFIELD IDENTIFICATION I

Conducted at

Institute of Decision Science
Claremont McKenna College

Clinic Participants:

Carolynn Black
Rick Clough
Jeanie Fiskin
Fif Ghobadian
Bob Hammett
Mary Hennried
Lee Leong
Joe Martinetto
Glenn Silberberg
Chandra Wahjudi

This document has been prepared for public release and sale. Distribution is unlimited.

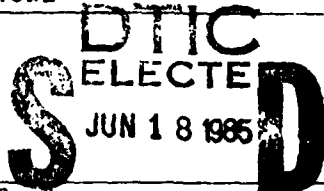
Daniel E. Hochman (Team Leader)
Professor Charles Bertness (Faculty Consultant)
Professor Janet Myhre (Faculty Supervisor)

FINAL REPORT

Jet Propulsion Laboratory
California Institute of Technology
June 1984

*Claremont Graduate School Claremont Men's College Kearney Mudd College
Pitzer College Pomona College Scripps College*

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER D-179	2. GOVT ACCESSION NO. D. 1155 737	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Applications of Correlation Techniques for Battlefield Identification I		5. TYPE OF REPORT & PERIOD COVERED FINAL
		6. PERFORMING ORG. REPORT NUMBER D-179
7. AUTHOR(s) D. Hochman, C. Black, R. Clough, J. Fiskin, F. Ghobadian, R. Hammett, M. Hernried, L. Leong, J. Martinetto, G. Silberberg, C. Wahjudi		8. CONTRACT OR GRANT NUMBER(s) NAS7-918
9. PERFORMING ORGANIZATION NAME AND ADDRESS Institute of Decision Science, Claremont McKenna College, Claremont, CA 91711		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS RE 182 AMEND # 187
11. CONTROLLING OFFICE NAME AND ADDRESS Commander, USAICS ATTN: ATSI-CD-SF Ft. Huachuca, AZ 85613-7000		12. REPORT DATE June 30, 1984
		13. NUMBER OF PAGES 100
14. MONITORING AGENCY NAME & ADDRESS (If different from Controlling Office) Jet Propulsion Laboratory ATTN: 171-209 California Institute of Technology 4800 Oak Grove, Pasadena, CA 91109		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE NONE
16. DISTRIBUTION STATEMENT (of this Report) Approved for Public Dissemination		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) A		
18. SUPPLEMENTARY NOTES Prepared by Jet Propulsion Laboratory for the US Army Intelligence Center and School's Combat Developer's Support Facility		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) SELF-CORRELATION ALGORITHM, MULTIVARIATE DISTRIBUTIONS, HOTELING'S T-SQUARED STATISTIC, ROBUSTNESS, MULTIVARIATE SKEWNESS		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) This report is the first in a series presenting a study of self-correlation algorithms in Intelligence and Electronic Warfare (IEW) systems. It was performed by the Mathematics Clinic of the Claremont Graduate School in support of the U. S. Army Intelligence Center and School (USAICS) Software Analysis and Management System (USAMS) task at Jet Propulsion Laboratory. The self-correlation algorithms use multivariate statistical tests to determine the equality of mean vectors from two different datasets.		

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

The statistical tests developed were variations of Hotelling's T-squared statistics. The main results deal with the analysis of the robustness of these statistics with respect normality and equal covariance matrices. To do this the Clinic developed mathematical techniques and simulation programs which generate multivariate normal, gamma, and lognormal distributions with a specified dependency between the vector components. Theoretical and sample multivariate skewness numbers of these distributions were also developed. The results indicate that the statistical tests are sensitive to skewness in the distributions but are not affected by slightly differing covariance matrices.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

U.S. ARMY INTELLIGENCE CENTER AND SCHOOL
Software Analysis and Management System

①

APPLICATIONS OF CORRELATION TECHNIQUES
FOR BATTLEFIELD IDENTIFICATION I

June 30, 1984

Concur:

Martha Ann Griesel

Martha Ann Griesel, Subgroup Leader
Algorithm Methodology Subgroup

F. Lesh

Fred Lesh, Manager
USAMS Task Team

J. P. McClure

J. P. McClure, Manager
Ground Data Systems Section

Julio Nakamura

Julio Nakamura, Manager
Defense Information Systems Program

JUN 13 1985

A

JET PROPULSION LABORATORY
California Institute of Technology
Pasadena, California

This document has been approved
for public release and sale; its
distribution is unlimited.

Acknowledgments

The Claremont Mathematics Clinic Team would like to express its appreciation to the Jet Propulsion Laboratory for the opportunity to be involved in this project. A very special thanks is expressed to our JPL liaisons Martha Ann Griesel and James W. Gillis. The achievement of the Clinic team is owing in large measure to their enthusiastic guidance and support.

The work described in this publication was performed at the Institute of Decision Science Claremont McKenna College, sponsored by the United States Army Intelligence Center and School. The writing and publication of this paper was supported by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Ac.	
NT	
DT	
Un	
Jun	
By	
Dis	
At	
Dist	
A-1	



Abstract

This report is the first in a series presenting a study of self-correlation algorithms in intelligence systems. It was performed by the Mathematics Clinic of the Claremont Graduate School in support of the Algorithm Analysis subtask of the U.S. Army Intelligence Center and School (USAICS) Software Analysis and Management System (USAMS) task at Jet Propulsion Laboratory. The self-correlation algorithms use multivariate statistical tests to determine the equality of mean vectors from two different datasets. The statistical tests developed were variations of Hotelling's T^2 -statistics. The main results deal with the analysis of the robustness of these statistics with respect to normality and equal covariance matrices. To do this the Clinic developed mathematical techniques and simulation programs which generate multivariate normal, gamma, and lognormal distributions with a specified dependency between the vector components. Theoretical and sample multivariate skewness numbers of these distributions were also developed. The results indicate that the statistical tests are sensitive to skewness in the distributions but are not affected by slightly differing covariance matrices.

self-correlation algorithms, multivariate distributions, Hotelling's T²-test, robustness, simulation

Key Words:

Self-correlation algorithm

Multivariate distributions

Hotelling's T^2 -statistic

Robustness

Multivariate skewness

Summary

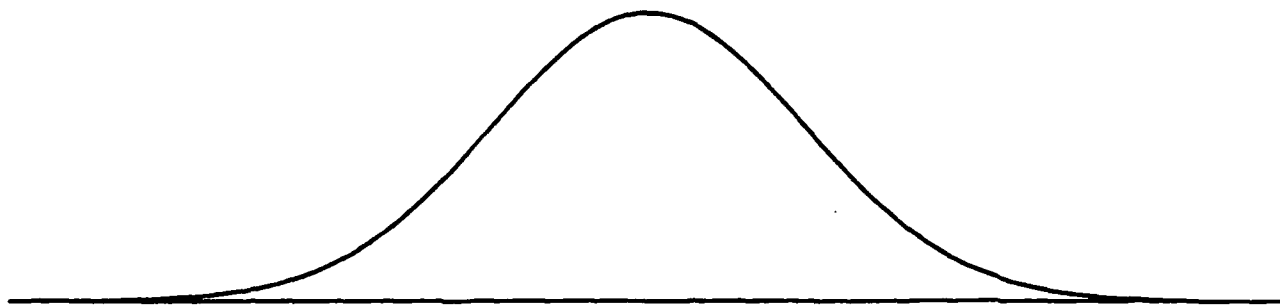
This study is the first in a series of reports involved in researching self-correlation and cross-correlation algorithms in intelligence systems. These algorithms are used to maintain a data base of current information about a battlefield. The initial view of the battlefield is stored in a central computer data base. As new data is received from the sensors on the battlefield, it is used to update the old data and formulate a new picture of the battlefield. The work on these algorithms reported here focuses on the sensitivity of the mathematical tests to changes and uncertainties in the data. This project was performed by the Mathematics Clinic of the Claremont Graduate School (CGS) in support of the Algorithm Analysis subtask of the U.S. Army Intelligence Center and School (USAICS) Software Analysis and Management System (USAMS) task at Jet Propulsion Laboratory (JPL). This is an ongoing task to study intelligence algorithms for the Combat Developers Support Facility at USAICS.

The analysis this year focused on statistical tests that are variations of Hotelling's T^2 -statistic. In many applications the Hotelling's T^2 -statistics revert to chi-square (χ^2) tests. (The χ^2 forms are used in many of the current systems.) These tests are used to check the equality of means in a multivariate setting. Such tests could be used, for example, to test if two location means or if two radar signal means are the same. The statistical tests have several variations depending on whether the mean of the old data and the dependency relationships of the old data and new data are assumed to be known or are estimated. Three main cases were studied:

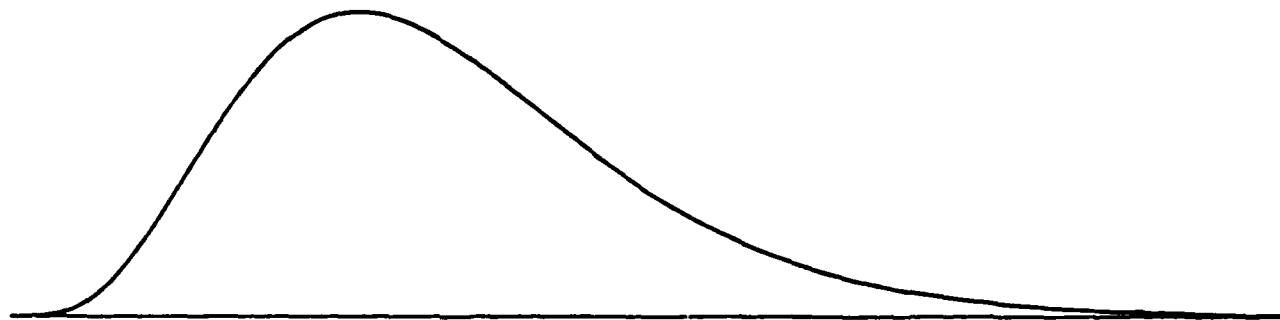
1. The mean of the old data is assumed to be known. The dependency relationships for both the old and new data are assumed to be known.
2. The mean of the old data is assumed to be known. The dependency relationship is estimated for the new data and is assumed to be known for the old data.
3. The mean of the old data is estimated. The dependency relationships for both the old and new data are estimated and are assumed to be equal.

Once the statistical tests were developed, the Clinic team began investigating the robustness of the tests, i.e. the sensitivity of the tests to the relaxation of the assumptions. All of the tests assume that the incoming data is normally distributed (Figure i). This assumption was deemed most likely to fail in two ways: the distributions are skewed or they have fat tails. To test the behavior of the tests when the distribution of the incoming data was skewed, the gamma (Figure ii) and lognormal (Figure iii) distributions were used. These distributions were chosen because they are representative of skewed distributions and have statistical properties which made the mathematical development possible. A major task of the Clinic team this year was to develop a software package to simulate representative multivariate data from these distributions. The results obtained indicate that if the actual distribution of the data is skewed, the statistical tests, developed under the normality assumptions, may be invalid. The simulation programs were also used to test the robustness of the assumption of equal covariance matrices while the investigation of fat-tailed distributions was left to future studies.

The development of simulation programs for determining the robustness of the intelligence algorithms and the analysis based on the statistical tests is being continued.



Univariate Normal
Figure i



Univariate Gamma
Figure ii



Univariate Lognormal
Figure iii

TABLE OF CONTENTS

			PAGE
Chapter	I	Introduction - - - - -	1
Chapter	II	Mathematical Models for the Self-correlation Stage - - - - -	9
Chapter	III	Multivariate Distributions with Component Dependency - - - - -	21
Chapter	IV	Skewness for Multivariate Distributions -	35
Chapter	V	Unequal Covariance Matrices - - - - -	51
Chapter	VI	Computer Simulation - - - - -	53
Chapter	VII	Robustness Results - - - - -	59
Chapter	VIII	Concluding Remarks - - - - -	65

PREVIOUS PAGE
IS BLANK

TABLE OF FIGURES

	PAGE
Figure 1 - - - - -	2
Figure 2 - - - - -	3
Figure 3 - - - - -	4
Figure 4 - - - - -	5
Figure 5 - - - - -	10
Figure 6 - - - - -	12
Figure 7 - - - - -	13
Figure 8 - - - - -	36
Figure 9 - - - - -	52
Figure 10 - - - - -	60
Figure 11 - - - - -	62
Figure 12 - - - - -	64

APPENDIXES

		PAGE
Appendix	I Finding the Maximum Likelihood Function for μ', Σ^{-1} - - - - -	67
Appendix	II The Distribution of T^2 for Case 2 - - -	71
Appendix	III Using Likelihood Ratio Criteria to Test the Equality of Variance-Covariance Matrices	77
Appendix	IV The Distribution of T^2 for Case 5 - - -	79
Appendix	V Elements of Covariance Matrix of a Lognormal	81
Appendix	VI Covariance Matrix Restrictions for Lognormal Case - - - - -	83
Appendix	VII Derivation of $\beta_{1,4}$ - - - - -	87
Appendix	VIII Generation of Univariate Random Samples -	93
Appendix	IX An Evaluation of Several Random Number Generators - - - - -	97
Bibliography	- - - - -	99



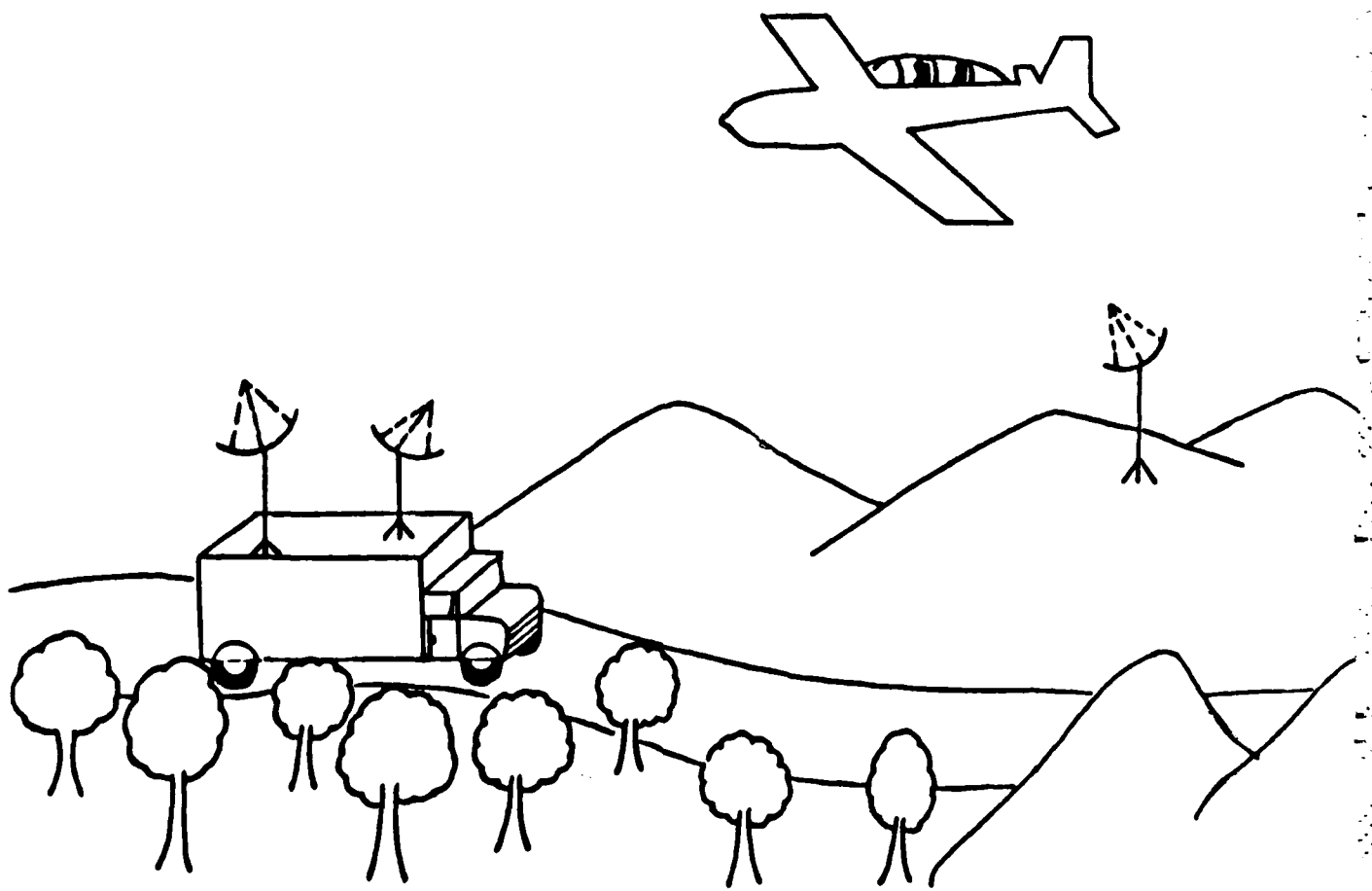
Chapter I - Introduction

The Mathematics Clinic project involved researching methods to maintain a data base of current information about a battlefield. The initial view of the battlefield is stored in a central computer data base. As new data is received from the sensors on the battlefield, it is used to update the data base and formulate a new picture of the battlefield.

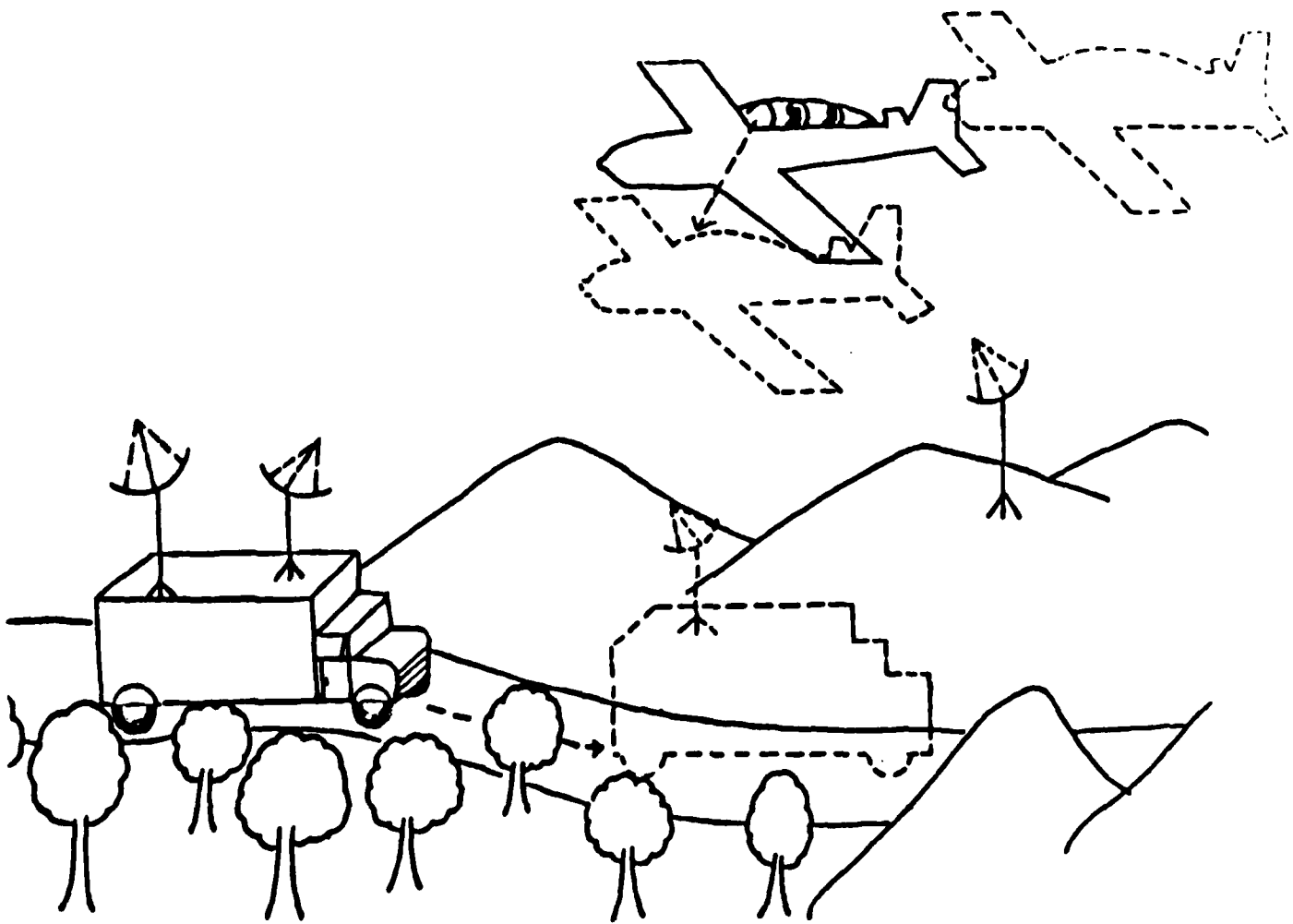
The process can be illustrated by Figures 1 and 2. The original view of the battlefield, illustrated by Figure 1, shows one plane, a truck with two radars, and a single radar on top of the hill. New data is collected from the battlefield, as represented by the dotted lines in Figure 2. It indicates that the previously identified truck and plane have moved to new locations, and that the truck now carries only one radar. It also shows a new incoming plane. The single radar on top of the hill, however, is still at the same location. The Clinic's goal was to develop statistical methods to analyze the incoming data in order to accurately update the data base.

The procedure followed to formulate the new estimate of the battlefield is illustrated in Figure 3. New data is first collected and filtered. This procedure prepares the data for the identification stage, where the source of the transmission can be determined. It is the next stage, self-correlation, on which the Clinic team concentrated its research efforts. Statistical tests are performed here to compare the new incoming data with the old estimate of the battlefield. The new estimate of the battlefield is then stored in the data base. Figure 4 shows the updated picture of the battlefield. The final stage is battlefield identification or cross-correlation which is the analysis of the configuration of the battlefield based on the enemies capabilities.

The statistical tests used in the self-correlation stage are variations of the Hotelling's T^2 -statistic, which is a multivariate extension of the student's t-statistic. The use of a multivariate test statistic is necessary

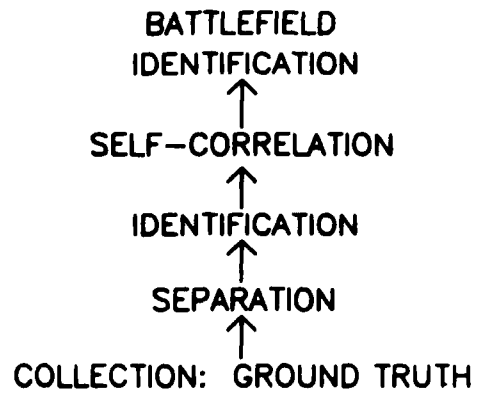


Original Battlefield Picture
Figure 1



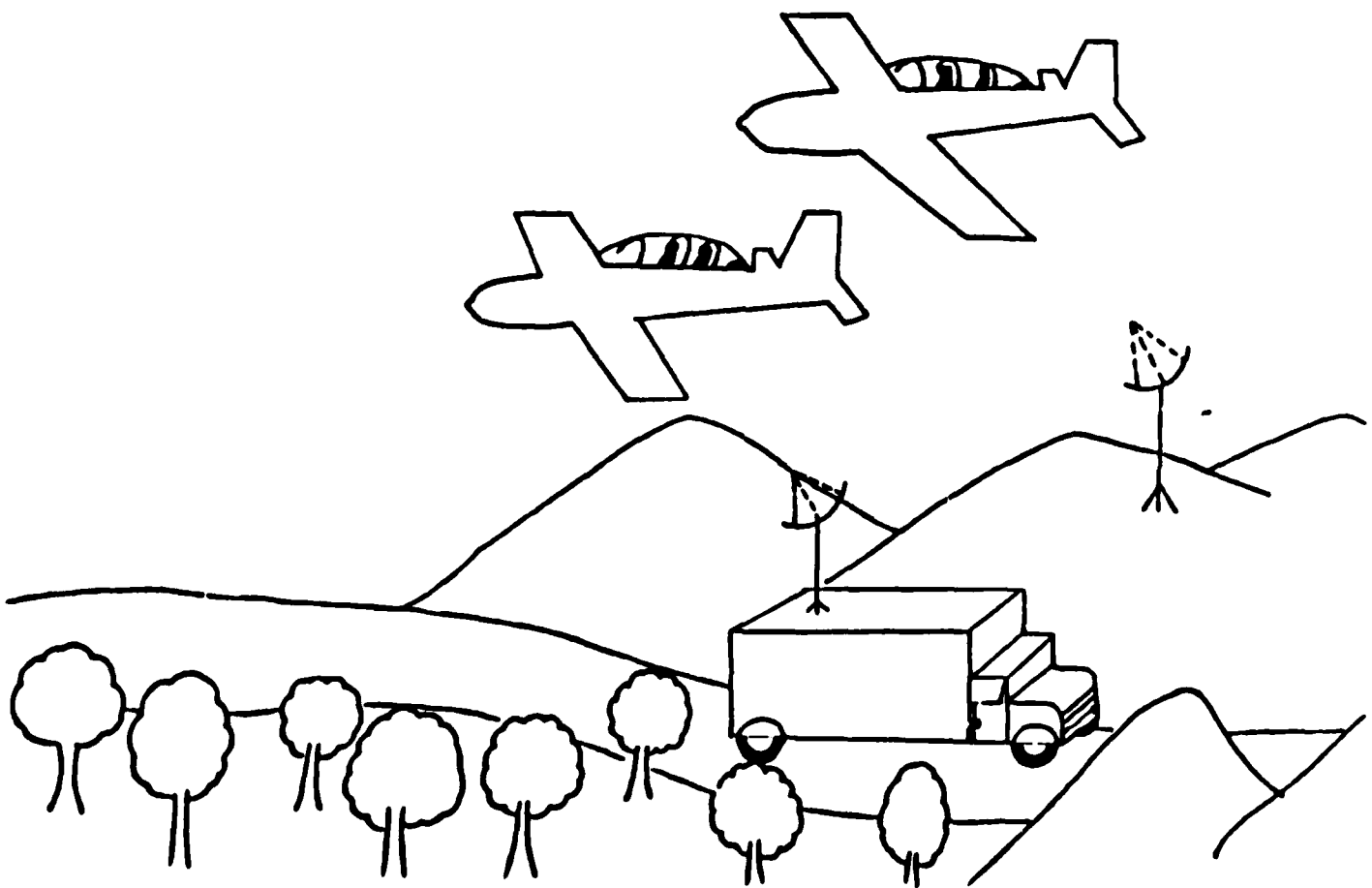
Original and New Battlefield Picture

Figure 2



Procedure to Update Data

Figure 3



Updated Battlefield Picture
Figure 4

because the data from the sensors comes in the form of vectors. (The location vector, for example, may have three components: longitude, latitude, and altitude. The signal parameter vector may have four components: pulse repetition interval, pulse width, scan rate, and frequency.) During the first semester, the Clinic team rederived the basic Hotelling's T^2 -statistic and its variations and then extended the results. Chapter II presents the mathematical models of the test statistics that the team developed.

The statistical tests used in this report, namely Hotelling's T^2 -statistic and its variations, were derived under certain assumptions. In some applications these assumptions may not be met. Thus it is desirable to investigate the robustness of the statistical tests used. (A statistical test is robust if the failure of assumptions does not invalidate its ability to yield correct results). One of the assumptions that was relaxed was that of normally distributed data. For example, the normality assumptions are violated when the data comes from skewed distributions such as gamma or lognormal. To measure the non-normality of multivariate distributions K. V. Mardia's generalizations of skewness and kurtosis were used. Chapter IV develops these concepts. A second assumption which was relaxed was that of equal covariance matrices. Chapter V contains a brief discussion of covariance matrices. To study the robustness of the above two assumptions, the Clinic team needed to generate multivariate distributions that had characteristics representative of the battlefield data. This involved generating random samples of vectors such that the vectors components have different statistical distributions which may not be independent of each other. That is although the vectors in the generated random sample are independent of each other, the statistical distributions of the vector components may be dependent. The dependence among the vector components is described by the covariance matrix of the multivariate

distribution. For example, if the covariance matrix has non-zero off-diagonal elements, then a dependency between the vector components exists. The method devised for generating these random samples consists of first specifying the desired component distributions and then specifying the component dependency (i.e. the covariance matrix). The components of the sample vectors are then assumed to be expressible as a combination of appropriately chosen univariate distributions. Using the assumed covariance matrix, it is sometime possible to solve for the coefficients of the combinations and hence obtain the desired random sample. Thus the problem of simulating data sets is reduced to finding the coefficient matrix of a system of equations. The Clinic team named the process of finding the coefficients matrix "backsolving" since the algorithm starts with the desired result (the covariance matrix) and ends with the means by which the problem is solved (the coefficient matrix). The details of the "backsolving" process are discussed in Chapter III. The associated computer simulation programs are described in Chapter VI.

The results of the Clinic team's investigations indicate that the statistical tests are sensitive to skewness in the distributions but are not affected by slightly differing covariance matrices. A detailed discussion of the robustness results are contained in Chapter VII.

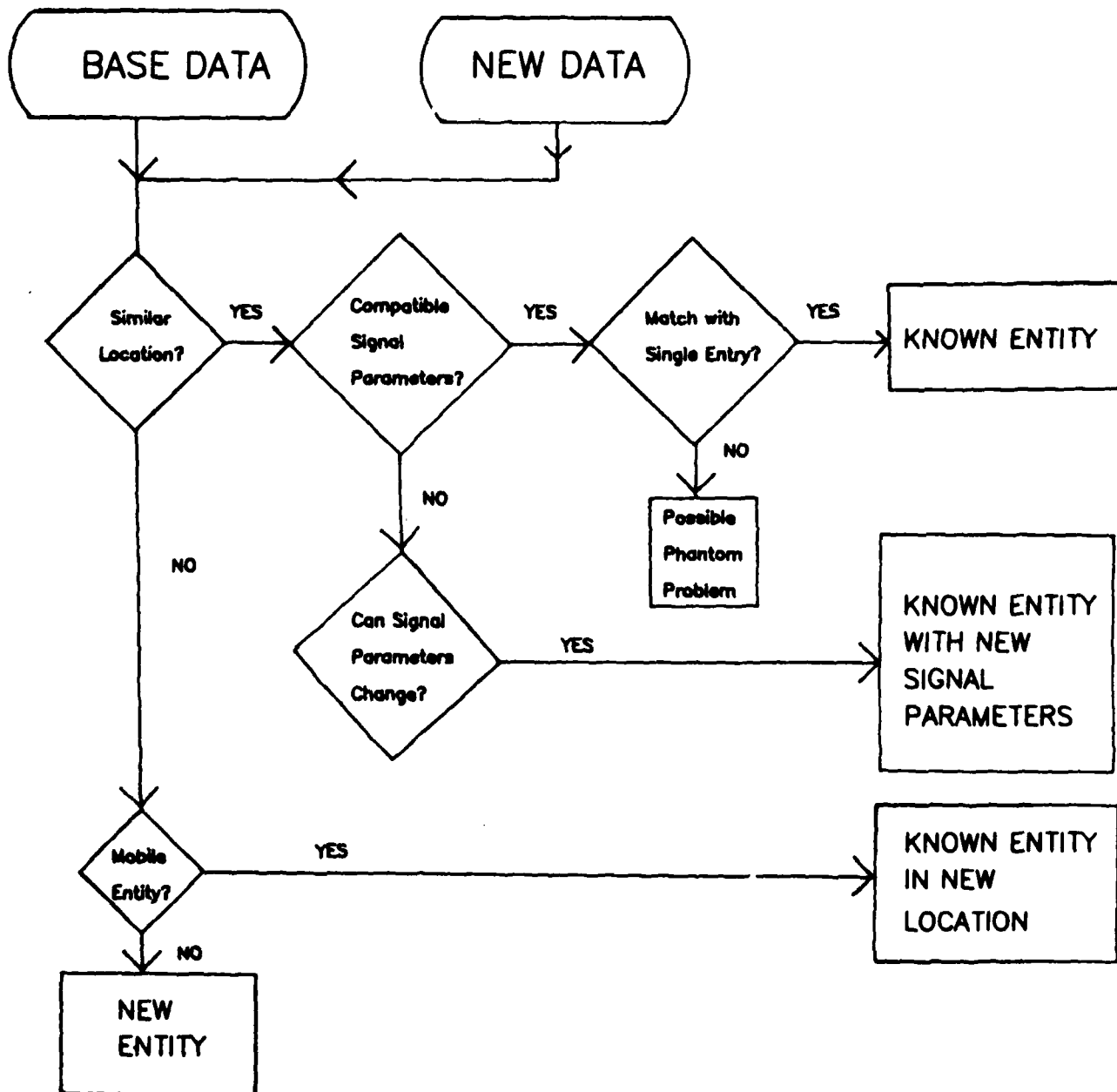
The report concludes with Chapter VIII which contains comments on possible future projects.

Chapter II - Mathematical Models for the Self-Correlation Stage

The Clinic Team analyzed various models for the self-correlation stage. In this chapter we briefly illustrate the self-correlation algorithm and then develop the appropriate mathematical models.

The self-correlation stage involves the testing for equality of mean vectors from two datasets. The first dataset represents the old estimate of the battlefield which has been stored in a central computer system and is referred to as the Base Data. The second dataset represents the new estimate of the battlefield and is referred to as the New Data. Information from both datasets is summarized and stored in the form of (mean) vectors and (covariance) matrices; denoted by $\underline{\mu}$ and Σ respectively. (Vectors will be denoted by an underscore.) The covariance matrix describes the relationships between the vector components. When referring to the Base Data we will use the subscript B and when referring to the New Data we will use the subscript N; for example, Σ_B denotes the covariance matrix from the Base Data.

Figure 5 shows how the self-correlation algorithm is implemented for the situations mostly likely to occur on a battlefield. For simplicity, not every possible case has been included; however, the logic used for most standard situations is illustrated. The first step in the algorithm is to test whether the candidate entity (New Data) is in the same location as a known entity (Base Data). If the locations of the two entities are determined to be the same, then their signal parameters are compared. If these two sets of parameters are compatible with one another, then we conclude that the observations refer to the same entity. However if the candidate entity is strongly associated with more than one known entity, then it is possible that either the candidate entity or the one of the known entities is a



Self-Correlation Algorithm
Figure 5

deceptive reading referred to as a phantom. Phantoms can be caused by failing to divide the observations of the two entities in the separation stage or by incorrectly matching a candidate entity with a known entity. When a candidate entity has the same location as a known entity but different and noncompatible signal parameters, the possibility that it is a known entity which has changed its signal parameters must be checked. The last possibility we consider is when the candidate entity's location is not the same as the location of any known entity. In this case it must be determined whether the candidate entity is a known entity with the same signal parameters that has moved (i.e. the known entity is mobile) or whether the candidate entity is a previously undetected entity.

The questions asked in the above decision process, as well as others in the self-correlation algorithm, are answered by using hypothesis tests of the equality of mean vectors. The desired statistical tests are based on Hotelling's T^2 -statistic or a variation. In the development of these tests there are two possibilities that must be considered; namely whether the mean vectors and covariance matrices are to be considered as known constants or as estimates. In all there are six variations of the model to be considered. These six cases are described verbally in Figure 6. In Figure 7 the six cases are enumerated and described using previously introduced notation.

As an example, under the assumption that μ_B is estimated, Σ_B is unknown but $\Sigma_B = \Sigma_N$ is assumed, case 5 is appropriate. Cases 1-5 lead to variations of Hotelling's T^2 -statistic to test

$$H_0: \mu_B = \mu_N$$

which is called the null hypothesis against

$$H_A: \mu_B \neq \mu_N$$

which is called the alternative hypothesis.

KNOWN BASE DATA	ESTIMATED BASE DATA	
Dependency relationship of New Data is known	Dependency relationships for Base and New Data are known and equal	Dependency relationships for Base and New Data are known and unequal
Dependency relationship of New Data is not known	Dependency relationships are not known, but are assumed to be equal	Dependency relationships are not known and are assumed to be unequal

Description of Mathematical Models

Figure 6

μ_B known	μ_B estimated	
<u>Case 1</u> Σ_N known	<u>Case 3</u> Σ_N known Σ_B known $\Sigma_B = \Sigma_N$	<u>Case 4</u> Σ_N known Σ_B known $\Sigma_B \neq \Sigma_N$
<u>Case 2</u> Σ_N estimated	<u>Case 5</u> Σ_N estimated Σ_B estimated $\Sigma_B = \Sigma_N$	<u>Case 6</u> Σ_N estimated Σ_B estimated $\Sigma_B \neq \Sigma_N$

Numeration of Mathematical Models
 Figure 7

Before we develop the mathematical models for these six cases, we give a more detailed description of the mean vector. A sensor observes p characteristics of a particular entity N times: therefore, there are N

vectors of the type $\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$. The incoming data is assumed to be from a

p -variate normally distributed population and the N observations are assumed independent of one another. The j th component of each vector satisfies the equation

$$x_j(t) = \mu_j + \epsilon_j(t)$$

where $x_j(t)$ is the value of the j th characteristic at time t , μ_j is the true value of the characteristic, which remains constant over time for a stationary entity and $\epsilon_j(t)$ is the error term at time t . The N vectors are averaged to obtain the sample mean vector

$$\bar{\underline{x}}(t) = \underline{\mu} + \underline{\epsilon}(t)$$

That is, $\bar{\underline{x}}(t)$ is the sample mean observed by the sensor at time t (average or final time of observations). We now proceed with development of the mathematical models for the self-correlation stage

Case 1 is the simplest and least likely case to occur, the reader is referred to Johnson and Leone (1964, pp. 294-295). The T^2 -statistic used in Case 2 is presented next.

Let $\underline{x}_1, \dots, \underline{x}_N$ be N observations, each with p characteristics.

$$\underline{x}_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1p} \end{pmatrix}$$

The distribution of \underline{x}_i is assumed to be normal with mean $\underline{\mu}$ and covariance matrix Σ (In the self-correlation algorithm $\underline{\mu} = \underline{\mu}_N$ and $\Sigma = \Sigma_N$.) Each observation vector has probability density function

$$f_{\underline{x}_i}(\underline{x}_i) = \frac{1}{(\sqrt{2\pi})^p |\Sigma|^{1/2}} \exp[-\frac{1}{2}(\underline{x}_i - \underline{\mu})' \Sigma^{-1}(\underline{x}_i - \underline{\mu})]$$

The joint density, or likelihood function, is

$$\begin{aligned} f_{\underline{x}_1, \dots, \underline{x}_N}(\underline{x}_1, \dots, \underline{x}_N) &= L(\underline{\mu}; \Sigma^{-1}) \\ &= \prod_{\alpha=1}^N [f_{\underline{x}_\alpha}(\underline{x}_\alpha)] \\ &= \frac{|\Sigma^{-1}|^{N/2}}{(\sqrt{2\pi})^{Np}} \exp[-\frac{1}{2} \sum_{\alpha=1}^N (\underline{x}_\alpha - \underline{\mu})' \Sigma^{-1}(\underline{x}_\alpha - \underline{\mu})] \end{aligned}$$

$L(\underline{\mu}', \Sigma^{-1})$ denotes the likelihood function of the parameters listed. For this case, the null hypothesis is $H_0: \underline{\mu} = \underline{\mu}_0$ with $\underline{\mu}_0$ known, and Σ is unknown and invertible. (In the self-correlation algorithm, $\underline{\mu}_0 = \underline{\mu}_B$, $\underline{\mu} = \underline{\mu}_N$, $\Sigma = \Sigma_N$.) In order to test H_0 we compute the likelihood ratio:

$$\lambda = \frac{\max_{\Sigma^{-1}} L(\underline{\mu}_0, \Sigma^{-1})}{\max_{\underline{\mu}, \Sigma^{-1}} L(\underline{\mu}', \Sigma^{-1})}$$

The numerator of this ratio is the maximum of the likelihood function with parameter space restricted by the null hypothesis, ω . The denominator is the corresponding function maximized over the entire parameter space, Ω . The test is to reject H_0 if the computed λ is small; that is, less than or equal to some λ_0 . The derivation of the maximum of the likelihood function over the entire parameter space, Ω , and over the restricted parameter space, ω , is shown in Appendix I. Using this derivation the likelihood ratio becomes

$$\lambda = \frac{|\hat{\Sigma}_{\Omega}|^{1/2N}}{|\hat{\Sigma}_{\omega}|^{1/2N}} = \frac{|\sum_{\alpha=1}^N (x_{\alpha} - \bar{x})(x_{\alpha} - \bar{x})'|^{N/2}}{|\sum_{\alpha=1}^N (x_{\alpha} - \mu_0)(x_{\alpha} - \mu_0)'|^{N/2}}$$

$$= \frac{|A|^{N/2}}{|A + N(\bar{x} - \mu_0)(\bar{x} - \mu_0)'|^{N/2}}$$

where $A = \sum_{\alpha=1}^N (x_{\alpha} - \bar{x})(x_{\alpha} - \bar{x})' = (N-1)S$, with $S = A/(N-1)$. Thus

$$\lambda^{2/N} = \frac{|A|}{|A + [\sqrt{N}(\bar{x} - \mu_0)][\sqrt{N}(\bar{x} - \mu_0)]'|}$$

or equivalently

$$\lambda^{2/N} = \frac{1}{1 + T^2/(N-1)} \quad (2.1)$$

where

$$\begin{aligned} T^2 &= (N - 1)N(\underline{x} - \underline{\mu}_0)'A^{-1}(\underline{x} - \underline{\mu}_0) \\ &= N(\bar{\underline{x}} - \underline{\mu}_0)'S^{-1}(\bar{\underline{x}} - \underline{\mu}_0) \end{aligned}$$

Equation 2.1 can also be expressed as

$$T^2 = (N-1)(\lambda^{-2/N} - 1).$$

The decision whether or not to reject the null hypothesis is determined by comparing λ against a critical λ_0 at a certain significance level. We reject H_0 when

$$\lambda^{2/N} \leq \lambda_0^{2/N} \quad (2.2)$$

Inverting Equation 2.2, subtracting 1, and multiplying by $N-1$, the critical region is redefined as

$$T^2 \geq T_0^2$$

In this case, T^2 can be shown to be distributed according to an F distribution with p and $N-p+1$ degrees of freedom. This result is proved in Appendix II.

For cases 3 through 6, the two sample tests, the T^2 -statistic does not follow one specific distribution; it varies depending on the assumptions on the covariance matrices. When the covariance matrices are unknown, and there is not enough data to assume that the estimates are equal to the true values, we use statistical methods to test whether or not two covariance matrices are equal. The derivation of this test is presented in Appendix III.

We will now examine the case of testing two observed sets of estimates, $\bar{x}^{(1)}$ and $\bar{x}^{(2)}$, for equality of mean vectors, i.e., $H_0: \underline{\mu}^{(1)} = \underline{\mu}^{(2)}$, or,

$$\begin{pmatrix} \mu_1^{(1)} \\ \mu_2^{(1)} \\ \vdots \\ \mu_p^{(1)} \end{pmatrix} = \begin{pmatrix} \mu_1^{(2)} \\ \mu_2^{(2)} \\ \vdots \\ \mu_p^{(2)} \end{pmatrix}$$

(In the self-correlation algorithm, superscript (1) quantities may be thought of as the New Data and superscript (2) as the Base Data.)

For case 3, to test H_0 , the statistic

$$T^2 = \frac{N_1 N_2}{N_1 + N_2} (\bar{x}^{(1)} - \bar{x}^{(2)})' \Sigma^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)})$$

is used, where $\Sigma = (\sigma_{ij})$; $\sigma_{ij}^{(1)} = \sigma_{ij}^{(2)} = \sigma_{ij}$ by assumption.

This statistic follows a noncentral χ^2 distribution with p degrees of freedom and non-centrality parameter

$$\delta = \frac{N_1 N_2}{N_1 + N_2} \sum_{i=1}^p \sum_{j=1}^p \sigma_{ij} (\mu_i^{(1)} - \mu_i^{(2)}) (\mu_j^{(1)} - \mu_j^{(2)}).$$

For case 4

$$T^2 = (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})' \bar{\Sigma}^{-1} (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)}) \text{ is used,}$$

where $\bar{\Sigma} = (\bar{\sigma}_{ij})$ and $\bar{\sigma}_{ij} = \frac{1}{N_1} \sigma_{ij}^{(1)} + \frac{1}{N_2} \sigma_{ij}^{(2)}$. This statistic follows a non-central χ^2 distribution with p degrees of freedom, and non-centrality parameter

$$\delta = \sum_{i=1}^p \sum_{j=1}^p \bar{\sigma}_{ij} (\mu_i^{(1)} - \mu_i^{(2)}) (\mu_j^{(1)} - \mu_j^{(2)})$$

The preceding results (cases 3 and 4) can be found in Johnson and Leone (1964, pp. 296-297).

Case 5 is analogous to case 3 except that Σ must be replaced with

$$S = \frac{1}{N_1 + N_2 - 2} \left[\sum_{\alpha=1}^{N_1} (\underline{x}_{\alpha}^{(1)} - \underline{\bar{x}}^{(1)}) (\underline{x}_{\alpha}^{(1)} - \underline{\bar{x}}^{(1)})' + \sum_{\alpha=1}^{N_2} (\underline{x}_{\alpha}^{(2)} - \underline{\bar{x}}^{(2)}) (\underline{x}_{\alpha}^{(2)} - \underline{\bar{x}}^{(2)})' \right]$$

Then T^2 becomes

$$T^2 = \frac{N_1 N_2}{N_1 + N_2} (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})' S^{-1} (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})$$

Appendix IV derives the distribution of this statistic, which leads to the critical region

$$T^2 > \frac{(N_1 + N_2 - 2)p}{N_1 + N_2 - p - 1} F_{p, N_1 + N_2 - p - 1}(\alpha)$$

where α is the level of significance. It is only in case 6, unequal and unknown covariance matrices, that there is not a precise statistical procedure yet developed to test the equality of mean vectors. In this case, other methods such as computer-intensive data handling techniques must be used.

The preceding tests will yield precise results only when all of the assumptions concerning the data are met. However, in some applications not all of the data will be normally distributed observations. The remainder of this report discusses techniques used to analyze the validity of the tests when assumptions are not valid.

Chapter III - Multivariate Distributions with Component Dependence

This chapter describes the mathematical methods used to generate data to test the robustness of the statistical tests of Chapter II. The robustness results are contained in Chapter VII while the computer techniques used are in Chapter VI.

To simulate relevant data, it is important to understand the characteristics of the observations made by a sensor. Examples of these observations are the signal parameter and location vectors. A signal parameter vector may have four characteristics or components: pulse repetition interval, pulse width, scan rate, and frequency. The location vector, on the other hand, may have longitude, latitude, and altitude as its components. Each data vector is independent of the others since the sensor takes readings from different locations at different times; however due to the nature of the type of data collected, the components of each data vector may not be independent of each other. A good example of such dependence in the signal parameter vector is the relationship between the pulse width and the pulse repetition interval. These two components are related through the peak-to-average power ratio.

When simulating data, we assume $p=4$. That is, we assume the mean vectors have 4 components and the covariance matrices have dimension 4×4 . (JPL has indicated this would be sufficient for the desired applications.) The Clinic team generated data from the multivariate normal, gamma, and lognormal distributions. The choice of these particular distributions was based on two criteria. First, these distributions have reasonable statistical properties and hence the mathematical development was possible. Secondly they have the desired properties for testing robustness. The multivariate gamma and lognormal provide examples of non-normal or skewed distributions while the multivariate normal is used as a "control" distribution.

The general approach in generating the desired multivariate random samples is as follows. We first specify the desired component distributions and component dependency (i.e., the covariance matrix, denoted by $\Sigma = (\sigma_{ij})$). To obtain the specified dependency the components of the vectors in the random sample are assumed to be expressible as a combination of appropriately chosen univariate distributions, we used the notation

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \quad \text{where } x_i = \sum_{j=1}^4 a_{ij} y_j, \quad i=1, \dots, 4$$

and the y_i are independent univariate random variables.

In matrix form,

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = A \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}$$

Thus the problem becomes one of finding a coefficient matrix A such that the multivariate distribution \underline{x} will have the specified distribution and covariance matrix. The Clinic team named the process of finding the coefficient matrix A "back-solving" since the algorithm starts with the desired result (the covariance matrix, Σ) and ends with the means (the coefficient matrix, A) by which the problem is solved. Once the coefficient matrix A has been found, the univariate random variables y_i are generated using standard

techniques which, together with A , produce the desired multivariate random samples.

Before looking at the specific distributions, several comments should be made. First since the covariance matrix is symmetric, we need only work with ten of its entries; therefore we assume that the coefficient matrix is lower triangular. Secondly when testing for robustness with respect to the normality assumption in case 5, we need to have $\Sigma_N = \Sigma_B$. Thus we need to generate multivariate random samples with equal covariances matrices. To do this we first show that we can backsolve for both the normal and gamma cases provided certain relationships among the entries of Σ are satisfied (see equations 3.9, 3.10 and 3.11). Since backsolving for the lognormal case appears to be very difficult mathematically, we avoid it by choosing our coefficient matrix A so that the resulting covariance matrix can be backsolved for the normal and gamma cases. Finally we note that although the Clinic team approached backsolving using basic algebraic manipulations, Cholesky decomposition techniques could have been used. We now develop the backsolving methods used for generating multivariate normal, gamma and lognormal data.

NORMAL

Constructing specified dependence between normal random variables was relatively simple due to the fact that any linear combination of normally distributed random variables is also normally distributed. By adding four univariate random variables, $y_i, i=1, \dots, 4$, which are $N(0,1)$, new random variables were created as follows:

$$x_i = a_{i1}y_1 + a_{i2}y_2 + a_{i3}y_3 + a_{i4}y_4 + \mu_i, \quad i = 1, \dots, 4$$

where $a_{ij} = 0$ for $i < j$

Each x_i is distributed normally with mean

$$E [x_i] = \mu_i,$$

$$\text{Var} [x_i] = \sum_{k=1}^i a_{ik}^2, \text{ and} \quad (3.1)$$

$$\text{Cov} [x_i, x_j] = \sum_{k=1}^4 a_{ik} a_{jk} \quad (3.2)$$

In matrix form, the new equations look like:

$$\underline{x} = A \underline{y} + \underline{u}$$

or equivalently,

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{bmatrix} a_{11} & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix}$$

Given a certain covariance matrix, $\Sigma = (\sigma_{ij})$, the objective is to compute the coefficient matrix, or the A matrix, so that the covariance matrix of \underline{x} is equal to Σ .

Using equations 3.1 and 3.2, the a_{ij} terms can be computed. For example, by equation 3.2,

$$\begin{aligned} \sigma_{41} &= a_{41} a_{11} + a_{42} a_{12} + a_{43} a_{13} + a_{44} a_{14} \\ &= a_{41} a_{11} \end{aligned}$$

since $a_{12} = a_{13} = a_{14} = 0$. And by equation 3.1 ,

$$\sigma_1^2 = a_{11}^2, \text{ or}$$

$$a_{11} = \sigma_1$$

Thus,

$$a_{41} = \frac{\sigma_{41}}{\sigma_1}$$

The rest of the a_{ij} terms are computed in a similar way involving numerous calculations, algebraic manipulations, and substitutions. They are

$$a_{11} = \sigma_1$$

$$a_{21} = \frac{\sigma_{12}}{\sigma_1}$$

$$a_{22} = \frac{\sqrt{\det B}}{\sigma_1}$$

$$a_{31} = \frac{\sigma_{13}}{\sigma_1}$$

$$a_{32} = \frac{\sigma_1^2 \sigma_{23} - \sigma_{12} \sigma_{13}}{\sigma_1 \sqrt{\det B}}$$

$$a_{33} = \sqrt{\frac{\det C}{\det B}}$$

$$a_{41} = \frac{\sigma_{14}}{\sigma_1}$$

$$a_{42} = \frac{\sigma_1^2 \sigma_{24} - \sigma_{12} \sigma_{14}}{\sigma_1 \sqrt{\det B}}$$

$$a_{43} = \frac{\sigma_1^2 \sigma_2^2 \sigma_{34} - \sigma_{12}^2 \sigma_{34} - \sigma_2^2 \sigma_{13} \sigma_{14} - \sigma_1^2 \sigma_{23} \sigma_{24} + \sigma_{12} \sigma_{13} \sigma_{24} + \sigma_{12} \sigma_{14} \sigma_{23}}{\sqrt{(\det B) (\det C)}}$$

$$a_{44} = \sqrt{\frac{\det D}{\det C}}$$

where

$$\det B = \det \begin{vmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{vmatrix} ,$$

$$\det C = \det \begin{vmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{vmatrix} \quad \text{and}$$

$$\det D = \det \begin{vmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{vmatrix}$$

GAMMA

The gamma distribution was chosen to represent a skewed distribution. Through its parameters, n and λ , the gamma has a great deal of versatility. The density function of a gamma distribution is

$$f_y(y) = \frac{\lambda^n y^{n-1}}{\Gamma(n)} e^{-\lambda y} \quad y, n, \lambda > 0,$$

Its mean and variance are

$$E[y] = \frac{n}{\lambda} \text{ and}$$

$$\text{Var}[y] = \frac{n}{\lambda^2} \text{ respectively.}$$

To simulate multivariate gamma random samples we assume, as in the normal case, that the coefficient matrix is lower triangular.

$$\begin{aligned} \text{Let } x_1 &= a_{11}y_1 \\ x_2 &= a_{21}y_1 + a_{22}y_2 \\ x_3 &= a_{31}y_1 + a_{32}y_2 + a_{33}y_3 \\ x_4 &= a_{41}y_1 + a_{42}y_2 + a_{43}y_3 + a_{44}y_4 \end{aligned}$$

where the y_i are independent gamma random variables with parameters (n_i, λ_i) , $i = 1, 2, 3, 4$. i.e., $y_i \sim G(n_i, \lambda_i)$.

In order for the x_i to be distributed as gamma, certain restrictions have to be imposed on the n_i and λ_i . These restrictions become clear when we look at the moment generating function. The moment generating function (MGF) of y_i is

$$M_{y_i}(t) = \left(1 - \frac{t}{\lambda_i}\right)^{-n_i}$$

Similarly,

$$M_{x_1}(t) = \left(1 - \frac{a_{11}t}{\lambda_1}\right)^{-n_1}$$

On the other hand, the MGF of x_2 is

$$\begin{aligned} M_{x_2}(t) &= E \left[e^{t(a_{21}y_1 + a_{22}y_2)} \right] \\ &= E \left[e^{a_{21}ty_1} \right] \cdot E \left[e^{a_{22}ty_2} \right] \quad (\text{by independence}) \\ &= \left(1 - \frac{a_{21}t}{\lambda_1}\right)^{-n_1} \cdot \left(1 - \frac{a_{22}t}{\lambda_2}\right)^{-n_2} \end{aligned}$$

The MGF of x_2 indicates that x_2 will be distributed as a $G(n_1+n_2, \frac{\lambda_1}{a_{21}} + \frac{\lambda_2}{a_{22}})$ only if

$$\frac{a_{21}}{\lambda_1} = \frac{a_{22}}{\lambda_2} \quad (3.3)$$

By a similar analysis, two more restrictions are imposed for x_3 and x_4 to be gamma. They are

$$\frac{a_{31}}{\lambda_1} = \frac{a_{32}}{\lambda_2} = \frac{a_{33}}{\lambda_3} \quad \text{and} \quad (3.4)$$

$$\frac{a_{41}}{\lambda_1} = \frac{a_{42}}{\lambda_2} = \frac{a_{43}}{\lambda_3} = \frac{a_{44}}{\lambda_4} \quad (3.5)$$

In summary, if restrictions 3.3, 3.4, and 3.5 hold, x_j will be distributed as

$$G \left(\sum_{k=1}^i n_k, \sum_{k=1}^i \frac{\lambda_k}{a_{ik}} \right)$$

The mean of x_i is

$$E [x_i] = \sum_{k=1}^4 \frac{n_k}{\lambda_k} a_{ik}$$

and the covariance matrix is

$$\sigma_{ij} = \text{cov}[x_i, x_j]$$

$$= E \left[\left(\sum_{i=1}^4 a_{ik} y_k - \sum_{k=1}^4 \frac{n_k}{\lambda_k} a_{ik} \right) \left(\sum_{k=1}^4 a_{jk} y_k - \sum_{k=1}^4 \frac{n_k}{\lambda_k} a_{jk} \right) \right]$$

$$= E \left[\left\{ \sum_{k=1}^4 a_{ik} \left(y_k - \frac{n_k}{\lambda_k} \right) \right\} \left\{ \sum_{k=1}^4 a_{jk} \left(y_k - \frac{n_k}{\lambda_k} \right) \right\} \right]$$

$$= \sum_{i=1}^4 a_{ik} a_{jk} \frac{n_k}{\lambda_k^2}$$

since $E \left[\left(y_k - \frac{n_k}{\lambda_k} \right) \left(y_j - \frac{n_j}{\lambda_j} \right) \right] = \begin{cases} 0 & k \neq j \\ \frac{n_k}{\lambda_k^2} & k = j \end{cases}$

Accordingly, the a_{ij} terms can be expressed in terms of σ_{ij} , λ_i , and n_i . After tedious algebraic manipulations

$$a_{ij} = 0, \text{ for } i < j$$

$$a_{11} = \frac{\sigma_{11}}{\sqrt{\sigma_{11} n_1}} \lambda_1$$

$$a_{21} = \frac{\sigma_{21} \lambda_1}{\sqrt{\sigma_{11} n_1}}$$

$$a_{22} = \frac{\sigma_{22}^{\lambda_2}}{\sqrt{\sigma_{22}(n_1+n_2)}}$$

$$a_{31} = \frac{\sigma_{31}^{\lambda_1}}{\sqrt{\sigma_{11}n_1}}$$

$$a_{32} = \frac{\sigma_{32}^{\lambda_2}}{\sqrt{\sigma_{22}(n_1+n_2)}}$$

$$a_{33} = \frac{\sigma_{33}^{\lambda_3}}{\sqrt{\sigma_{33}(n_1+n_2+n_3)}}$$

$$a_{41} = \frac{\sigma_{41}^{\lambda_1}}{\sqrt{\sigma_{11}n_1}}$$

$$a_{42} = \frac{\sigma_{42}^{\lambda_2}}{\sqrt{\sigma_{22}(n_1+n_2)}}$$

$$a_{43} = \frac{\sigma_{43}^{\lambda_3}}{\sqrt{\sigma_{33}(n_1+n_2+n_3)}}$$

$$a_{44} = \frac{\sigma_{44}^{\lambda_4}}{\sqrt{\sigma_{44}(n_1+n_2+n_3+n_4)}}$$

The covariance matrix, Σ is constrained due to the restrictions imposed on the relationships between the a_{ij} terms by equations 3.3 , 3.4 , and 3.5 .

From equation 3.3 ,

$$\frac{\frac{\sigma_{21}^{\lambda_1}}{\sqrt{\sigma_{11}n_1}}}{\lambda_1} = \frac{\frac{\sigma_{22}^{\lambda_2}}{\sqrt{\sigma_{22}(n_1+n_2)}}}{\lambda_2}$$

$$\frac{\sigma_{21}}{\sqrt{\sigma_{11}n_1}} = \frac{\sigma_{22}}{\sqrt{\sigma_{22}(n_1+n_2)}} \quad (3.6)$$

Similarly, from equation 3.4 ,

$$\frac{\sigma_{31}}{\sqrt{\sigma_{11}n_1}} = \frac{\sigma_{32}}{\sqrt{\sigma_{22}(n_1+n_2)}} = \frac{\sigma_{33}}{\sqrt{\sigma_{33}(n_1+n_2+n_3)}} \quad (3.7)$$

and from equation 3.5 ,

$$\frac{\sigma_{41}}{\sqrt{\sigma_{11}n_1}} = \frac{\sigma_{42}}{\sqrt{\sigma_{22}(n_1+n_2)}} = \frac{\sigma_{43}}{\sqrt{\sigma_{33}(n_1+n_2+n_3)}} = \frac{\sigma_{44}}{\sqrt{\sigma_{44}(n_1+n_2+n_3+n_4)}} \quad (3.8)$$

Combining 3.6 , 3.7 , and 3.8 , the restrictions on σ_{ij} become

$$\frac{\sigma_{12}}{\sigma_{22}} = \frac{\sigma_{13}}{\sigma_{23}} = \frac{\sigma_{14}}{\sigma_{24}} \quad (3.9)$$

$$\frac{\sigma_{23}}{\sigma_{33}} = \frac{\sigma_{24}}{\sigma_{34}} \quad (3.10)$$

$$\frac{\sigma_{33}\sigma_{44}}{\sigma_{43}^2} > 1 \quad (3.11)$$

LOGNORMAL

The lognormal distribution is a skewed distribution which is related to the normal distribution. Given that y is normally distributed with mean μ and variance σ^2 then $x = \exp[y]$ is defined as a lognormal distribution. The expected value of x is

$$E[x] = \exp[\mu] \cdot \exp[\sigma^2/2]$$

while the variance is

$$\text{Var}[x] = \exp[2\mu] \cdot \exp[\sigma^2] \cdot (\exp[\sigma^2] - 1).$$

Since $x = \exp[y]$, the condition $0 < x < \infty$ must hold.

Let

$$x_i = \exp \left[\sum_{j=1}^i a_{jj} y_j + b_i \right] \quad i = 1, \dots, 4$$

Thus

$$x_1 = \exp [a_{11} y_1 + b_1]$$

$$x_2 = \exp [a_{11} y_1 + a_{22} y_2 + b_2]$$

$$x_3 = \exp [a_{11} y_1 + a_{22} y_2 + a_{33} y_3 + b_3]$$

$$x_4 = \exp [a_{11} y_1 + a_{22} y_2 + a_{33} y_3 + a_{44} y_4 + b_4]$$

where the a_{jj} and b_i are scalar constants, and the y_i are independent standard normal distributions. i.e. $y_i \sim N(0,1)$. Each x_i is lognormally distributed

since $\left[\sum_{j=1}^i a_{jj} y_j + b_i \right]$ is distributed normally with mean b_i and variance $\sum_{j=1}^i a_{jj}^2$.

Using the model shown above, the expected value of x_i is

$$E[x_i] = \exp[b_i] \cdot \exp \left[\frac{\sum_{k=1}^i a_{kk}^2}{2} \right]$$

The variances of the x_i , which are the main diagonal elements of the covariance matrix, are

$$\text{Var}[x_i] = \exp[2b_i + \left(\sum_{k=1}^i a_{kk}^2\right)] \left[\exp\left(\sum_{k=1}^i a_{kk}^2\right) - 1\right]$$

Thus

$$\sigma_{ij} = \exp \left[b_i + b_j + 1/2 \left(\sum_{k=1}^i a_{kk}^2 + \sum_{k=1}^j a_{kk}^2 \right) \right] \left[\exp \left(\sum_{k=1}^i a_{kk}^2 \right) - 1 \right]$$

Please see appendix V for the actual equations for each σ_{ij} .

The relationships needed between the entries of the covariance matrix $\Sigma = (-$
in order to backsolve for the normal and gamma cases are

$$\frac{\sigma_{12}}{\sigma_{22}} = \frac{\sigma_{13}}{\sigma_{23}} = \frac{\sigma_{13}}{\sigma_{24}}$$

$$\frac{\sigma_{23}}{\sigma_{33}} = \frac{\sigma_{24}}{\sigma_{34}}$$

$$\frac{\sigma_{33}\sigma_{44}}{\sigma_{43}^2} > 1$$

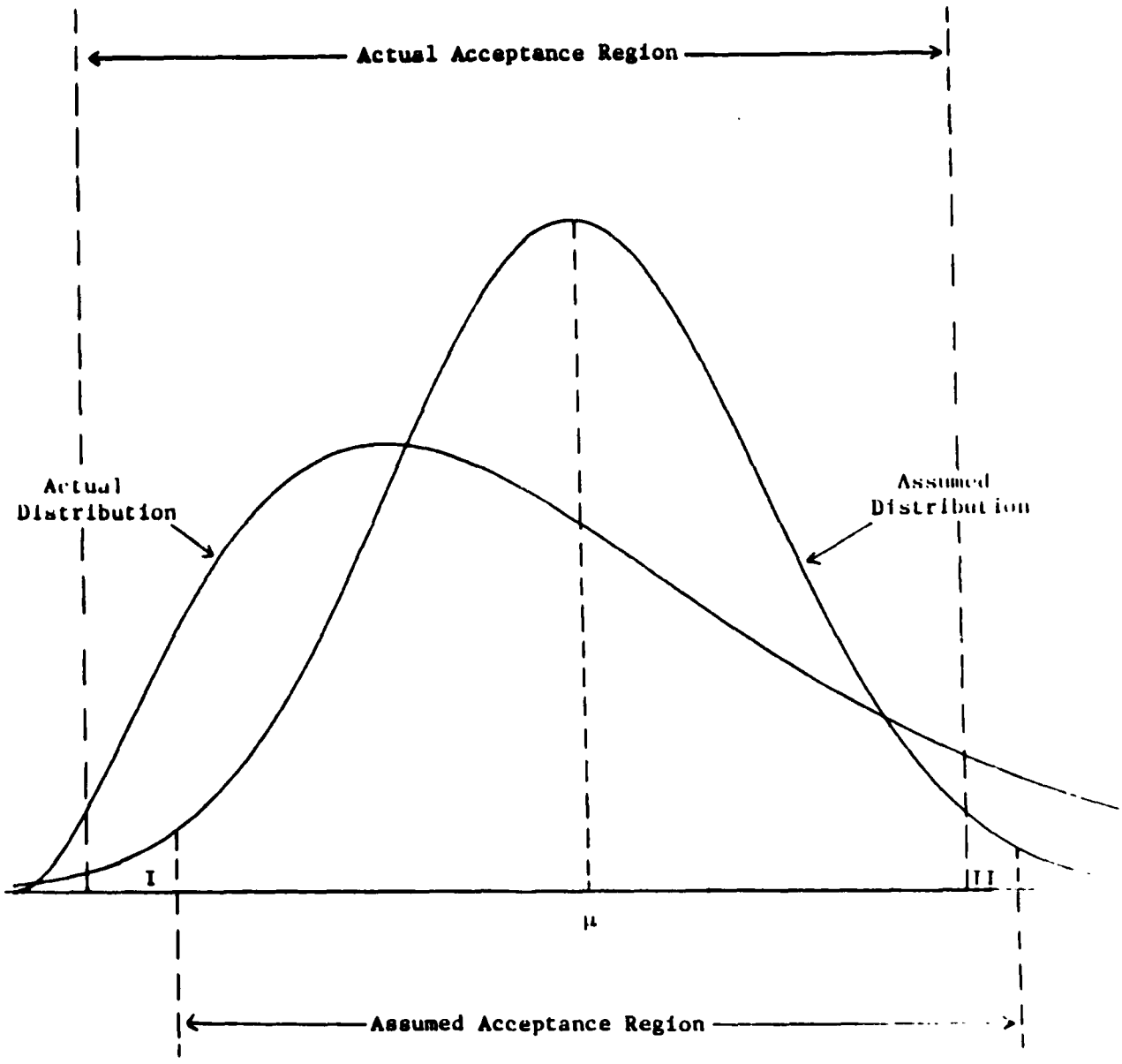
Appendix VI shows these relationships will always hold for a multivariate lognormal random variable defined as above.



Chapter IV - Skewness for Multivariate Distributions

The previously discussed distributions (Chapter III) were chosen in order to study the robustness of the Hotelling's T^2 -statistics. The Clinic team decided to analyze the effect of relaxing the assumptions of normality and equal covariance matrices. The relaxing of equal covariance matrices is discussed in Chapter V. To analyze robustness with respect to normality, the normal distribution is used as a control since it is the distribution on which the T^2 -statistics are based. Two descriptive statistics commonly used in the literature to measure the non-normality of multivariate distributions are skewness and kurtosis. Skewness is a measure of how symmetric a distribution is. A symmetric distribution, such as normal, will have a skewness of zero. Kurtosis, or excess, is a measure of the probability density in the tails of a distribution. The normal distribution is said to have a standard level of kurtosis. By using these two measures in statistical tests, one can determine whether a distribution is normal or not [Kres (1983, Table 26)] These tests for multivariate normality, developed by K. V. Mardia, were not studied by the Clinic team; however, Mardia's theories of multivariate skewness and kurtosis were used extensively. In a Monte-Carlo study of the T^2 -statistic in a univariate case, Mardia found that the level of skewness had a much more significant effect on the test statistic than the level of kurtosis [Mardia (1975,p.167)]. Similar results are indicated throughout the literature.

The following example will illustrate the effects of skewness. Figure 8 shows a non-normal skewed distribution centered at the correct mean and the assumed normal distribution. This case can lead to two types of errors. If a candidate entity falls outside the assumed acceptance region, it is considered a new entity and is added to the Base Data. Thus when the candidate



Skewed Distributions
Figure 8

entity falls in Region I, an error will be made since the assumed acceptance region is too small. This error will create a phantom entity in the Data Base. The opposite problem occurs in Region II. Since the assumed acceptance region is too large compared to the actual acceptance region, the tests will incorrectly associate a candidate entity with a known entity in the Data Base. Hence a new entity will not be detected when it enters the battlefield.

To analyze the effect of skewness, the gamma and lognormal distributions were examined along with the normal. The gamma and lognormal are both skewed distributions common to statistical analysis. This chapter will develop the skewness model in general terms and then present the skewness formulas for the gamma and lognormal distributions used in the Clinic teams's analysis.

Most distributions can be characterized by their moments about the mean. Specifically, the third moment is the measure of skewness:

$$\beta_{1,p} = E \left\{ [(\underline{x}-\underline{\mu})^{-1}(\underline{y}-\underline{\mu})] \right\} \quad (4.1)$$

Here, \underline{x} and \underline{y} are independently and identically distributed random variables, and p is the number of characteristics in each vector. For $p=2$, the following theorem gives an alternate expression for $\beta_{1,p}$.

Theorem Let $\underline{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $\underline{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$, $\Sigma = (\sigma_{ij})$, $\Sigma^{-1} = (\sigma^{ij})$, and $\rho = \text{corr}(x_1, x_2) = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$.

Also for $1 \leq r, s, t \leq 2$ let

$$\mu_{rst} = E[(x_r - \mu_r)(x_s - \mu_s)(x_t - \mu_t)] \quad (4.2)$$

and

$$\gamma_{rst} = \frac{\mu_{rst}}{\sigma_r \sigma_s \sigma_t} \quad (4.3)$$

then

$$\begin{aligned} \beta_{1,2} = & \frac{1}{(1-\rho^2)^3} \left[\gamma_{111}^2 + \gamma_{222}^2 + 3(1+2\rho^2)(\gamma_{122}^2 + \gamma_{112}^2) \right. \\ & - 2\rho^3 \gamma_{111} \gamma_{222} + 6\rho \left\{ \gamma_{111}(\rho \gamma_{122} - \gamma_{112}) \right. \\ & \left. \left. + \gamma_{222}(\gamma_{112}\rho - \gamma_{122}) - (2+\rho^2)\gamma_{122}\gamma_{112} \right\} \right] \end{aligned}$$

Proof:

$$\beta_{1,2} = E \left[\left[\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{pmatrix} \right]^3 \right]$$

Let

$$a = ((x_1 - \mu_1)\sigma^{11} + (x_2 - \mu_2)\sigma^{21})(y_1 - \mu_1) \quad \text{and}$$

$$b = ((x_1 - \mu_1)\sigma^{12} + (x_2 - \mu_2)\sigma^{22})(y_2 - \mu_2)$$

then

$$\begin{aligned} \beta_{1,2} &= E[(a + b)^3] \\ &= E[(a^3 + 3a^2b + 3ab^2 + b^3)] \\ &= E[a^3] + E[b^3] + E[3a^2b] + E[3ab^2] \end{aligned} \quad (4.4)$$

The elements of Σ^{-1} can be expressed in terms of ρ . They are

$$\sigma^{11} = \frac{1}{\sigma_1^2 \sigma_2^2 - \sigma_{12}^2} \quad \cdot \quad \sigma_2^2 = \frac{1}{\sigma_1^2 \left(1 - \frac{\sigma_{12}^2}{\sigma_1^2 \sigma_2^2} \right)} = \frac{1}{\sigma_1^2 (1 - \rho^2)} \quad (4.5)$$

$$\sigma^{22} = \frac{1}{\sigma_1^2 \sigma_2^2 - \sigma_{12}^2} \cdot \sigma_1^2 = \frac{1}{\sigma_2^2 \left(1 - \frac{\sigma_{12}^2}{\sigma_1^2 \sigma_2^2}\right)} = \frac{1}{\sigma_2^2 (1 - \rho^2)} \quad (4.6)$$

$$\sigma^{12} = \sigma^{21} = \frac{1}{\sigma_1^2 \sigma_2^2 - \sigma_{12}^2} \cdot (-\sigma_{12}) = \frac{-\sigma_{12}}{\sigma_1^2 \sigma_2^2 \left(1 - \frac{\sigma_{12}^2}{\sigma_1^2 \sigma_2^2}\right)} = \frac{-\sigma_{12}}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)} \quad (4.7)$$

It can easily be shown that

$$\begin{aligned} a^3 = & (y_1 - \mu_1)^3 \left[(x_1 - \mu_1)^3 (\sigma^{11})^3 + 3(x_1 - \mu_1)^2 (\sigma^{11})^2 (x_2 - \mu_2) \sigma^{21} \right. \\ & + 3(x_1 - \mu_1) \sigma^{11} (x_2 - \mu_2)^2 (\sigma^{21})^2 \\ & \left. + (x_2 - \mu_2)^3 (\sigma^{21})^3 \right] \end{aligned}$$

By using equation 4.2, the expected value of a^3 can be written as

$$\begin{aligned} E[a^3] = & \mu_{111} [\mu_{111} (\sigma^{11})^3 + 3\mu_{112} (\sigma^{11})^2 \sigma^{21} + 3\mu_{122} \sigma^{11} (\sigma^{21})^2 \\ & + \mu_{222} (\sigma^{21})^3] \end{aligned}$$

Substitution of equations 4.5, 4.6, and 4.7 into σ^{11} , σ^{22} , and σ^{12} , of the above equation yields

$$\begin{aligned} E[a^3] = & \mu_{111} \left[\mu_{111} \left(\frac{1}{\sigma_1^2 (1 - \rho^2)} \right)^3 + 3\mu_{112} \left(\frac{1}{\sigma_1^4 (1 - \rho^2)^2} \right) \left(\frac{-\sigma_{12}}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)} \right) \right. \\ & \left. + 3\mu_{122} \left(\frac{1}{\sigma_1^2 (1 - \rho^2)} \cdot \frac{\sigma_{12}^2}{\sigma_1^4 \sigma_2^4 (1 - \rho^2)^2} \right) + \mu_{222} \frac{-\sigma_{12}^3}{\sigma_1^6 \sigma_2^6 (1 - \rho^2)^3} \right] \end{aligned}$$

$$= \frac{1}{(1-\rho^2)^3} \left[\mu_{111}^2 \frac{1}{\sigma_1^6} - \frac{3\mu_{111}\mu_{112}\sigma_{12}}{\sigma_1^6\sigma_2^2} + \frac{3\mu_{111}\mu_{122}\sigma_{12}^2}{\sigma_1^6\sigma_2^4} - \frac{\mu_{111}\mu_{222}\sigma_{12}^3}{\sigma_1^6\sigma_2^6} \right]$$

And by equation 4.3,

$$E[a^3] = \frac{1}{(1-\rho^2)^3} \left[\gamma_{111}^2 - 3\gamma_{111}\gamma_{112}\rho + 3\gamma_{111}\gamma_{122}\rho^2 - \gamma_{111}\gamma_{222}\rho^3 \right]$$

Similarly,

$$E[b^3] = \mu_{222} \left[\mu_{111}(\sigma^{12})^3 + 3\mu_{112}(\sigma^{12})^2\sigma^{22} + 3\mu_{122}\sigma^{12}(\sigma^{22})^2 + \mu_{111}(\sigma^{22})^2 \right]$$

By equations 4.3, 4.5 - 4.7,

$$E[b^3] = \frac{1}{(1-\rho^2)^3} \left[\gamma_{222}\gamma_{111}\rho^3 + 3\gamma_{222}\gamma_{112}\rho^2 - 3\gamma_{222}\gamma_{122}\rho + \gamma_{222}\gamma_{222} \right]$$

The same procedures also apply in calculating $E[a^2b]$. First

$$a^2b = (y_1 - \mu_1)^2 (y_2 - \mu_2) \left[(x_1 - \mu_1)^3 (\sigma^{11})^2 \sigma^{12} + (x_2 - \mu_2)^3 (\sigma^{21})^2 \sigma^{22} + (x_1 - \mu_1)^2 (x_2 - \mu_2) 2\sigma^{11}(\sigma^{12})_2 + (\sigma^{11})^2 \sigma^{22} + (x_1 - \mu_1)(x_2 - \mu_2)^2 \left\{ 2\sigma^{11}\sigma^{21}\sigma^{22} + (\sigma^{12})^3 \right\} \right]$$

Its expected value is

$$\begin{aligned}
 E[a^2b] &= \mu_{112} \left[\mu_{111} (\sigma^{11})^2 \sigma^{12} + \mu_{222} (\sigma^{21})^2 \sigma^{22} \right. \\
 &\quad + \mu_{112} \left\{ 2\sigma^{11} (\sigma^{12})^2 + (\sigma^{11})^2 \sigma^{22} \right\} \\
 &\quad \left. + \mu_{112} \left\{ 2\sigma^{11} \sigma^{21} \sigma^{22} + (\sigma^{12})^3 \right\} \right]
 \end{aligned}$$

It can be expressed in terms of ρ such that

$$\begin{aligned}
 E[a^2b] &= \mu_{112} \left[\mu_{111} \cdot \left(\frac{-\sigma_{12}}{\sigma_1^6 \sigma_2^2 (1-\rho^2)^3} \right) + \mu_{222} \left(\frac{\sigma_{12}^2}{\sigma_1^4 \sigma_2^6 (1-\rho^2)^3} \right) \right. \\
 &\quad + \mu_{112} \left(\frac{2\sigma_{12}^2}{(1-\rho^2)^3 \sigma_1^6 \sigma_2^4} + \frac{1}{\sigma_1^4 \sigma_2^2 (1-\rho^2)^3} \right) \\
 &\quad \left. + \mu_{112} \left(\frac{-2\sigma_{12}}{(1-\rho^2)^3 \sigma_1^4 \sigma_2^4} - \frac{\sigma_{12}^3}{\sigma_1^6 \sigma_2^6 (1-\rho^2)^3} \right) \right]
 \end{aligned}$$

The final substitution yields

$$\begin{aligned}
 E[a^2b] &= \frac{1}{(1-\rho^2)^3} \left[-\gamma_{112} \gamma_{111} \rho + \gamma_{112} \gamma_{222} \rho^2 + \gamma_{112}^2 (2\rho^2 + 1) \right. \\
 &\quad \left. + \gamma_{112} \gamma_{122} (-2\rho - \rho^3) \right] \quad (4.10)
 \end{aligned}$$

Similarly,

$$\begin{aligned}
 ab^2 &= (y_1 - \mu_1)(y_2 - \mu_2)^2 \left[(x_1 - \mu_1)^3 (\sigma^{12})^2 \sigma^{11} + (x_2 - \mu_2)^3 \sigma^{21} (\sigma^{22})^2 \right. \\
 &\quad + (x_1 - \mu_1)^2 (x_2 - \mu_2) \left\{ 2\sigma^{12} \sigma^{22} \sigma^{11} + (\sigma^{12})^3 \right\} \\
 &\quad \left. + (x_1 - \mu_1)(x_2 - \mu_2)^2 \left\{ 2(\sigma^{12})^2 \sigma^{22} + \sigma^{11} (\sigma^{22})^2 \right\} \right]
 \end{aligned}$$

and therefore,

$$\begin{aligned}
 E[ab^2] &= \left[\mu_{122} \mu_{111} (\sigma^{12})^2 \sigma^{11} + \mu_{222} \sigma^{21} (\sigma^{22})^2 \right. \\
 &\quad \left. + \left\{ \mu_{112} 2\sigma^{12} \sigma^{22} \sigma^{11} + (\sigma^{12})^3 \right\} \right. \\
 &\quad \left. + \left\{ \mu_{112} 2(\sigma^{12})^2 \sigma^{22} + \sigma^{11} (\sigma^{22})^2 \right\} \right] \\
 &= \frac{\mu_{122}}{(1-\rho^2)^3} \left[\mu_{111} \frac{\sigma_{12}^2}{\sigma_1^6 \sigma_2^4} - \frac{\mu_{222} \sigma_{12}}{\sigma_1^2 \sigma_2^6} + \mu_{112} \left(\frac{-2\sigma_{12}}{\sigma_1^4 \sigma_2^4} - \frac{\sigma_{12}^3}{\sigma_1^6 \sigma_2^6} \right) \right. \\
 &\quad \left. + \mu_{122} \left(\frac{2\sigma_{12}^2}{\sigma_1^4 \sigma_2^6} + \frac{1}{\sigma_1^2 \sigma_2^4} \right) \right] \\
 &= \frac{1}{(1-\rho^2)^3} \left[\gamma_{122} \gamma_{111} \rho^2 - \gamma_{122} \gamma_{222} \rho + \gamma_{122} \gamma_{112} (-2\rho - \rho^3) \right. \\
 &\quad \left. + \gamma_{122}^2 (2\rho^2 + 1) \right] \quad (4.11)
 \end{aligned}$$

Finally, results from 4.8 , 4.9 , 4.10 , and 4.11 are substituted back into 4.4 to give

$$\begin{aligned}
 \beta_{1,2} &= \frac{1}{(1-\rho^2)^3} \left[\gamma_{111}^2 - 6\gamma_{111} \gamma_{112} \rho + 6\gamma_{111} \gamma_{122} \rho^2 \right. \\
 &\quad - 2\gamma_{111} \gamma_{222} \rho^3 + 6\gamma_{222} \gamma_{112} \rho^2 - 6\gamma_{222} \gamma_{122} \rho \\
 &\quad + \gamma_{222}^2 + 3\gamma_{112}^2 (2\rho^2 + 1) + 6\gamma_{122} \gamma_{112} (-2\rho - \rho^3) \\
 &\quad \left. + 3\gamma_{122}^2 (2\rho^2 + 1) \right]
 \end{aligned}$$

Regrouping, we obtain the desired result:

$$\begin{aligned}
 \beta_{1,2} &= \frac{1}{(1-\rho^2)^3} \left[\gamma_{111}^2 + \gamma_{222}^2 + 3(1+2\rho^2)(\gamma_{122}^2 + \gamma_{112}^2) \right. \\
 &\quad - 2\rho^3 \gamma_{111} \gamma_{222} + \left\{ 6\rho \gamma_{111} (\rho \gamma_{122} - \gamma_{112}) \right. \\
 &\quad \left. \left. + \gamma_{222} (\gamma_{112} \rho - \gamma_{122}) - (2+\rho^2) \gamma_{122} \gamma_{112} \right\} \right]
 \end{aligned}$$

This result corresponds to the work of Mardia in the two dimensional case. [Mardia (1970, p. 523.)] Mardia gives a general formula for the computing of theoretical skewness of a multivariate random variable. To check Mardia's results, the Clinic team developed a formula for the case $p = 4$. This derivation is presented in Appendix VII. The team then implemented both Mardia's formula and the team's in a computer program to compare results for $p = 4$. It was found that Mardia's general skewness formula is correct. The formula is:

$$\sum_{\substack{\sigma \\ r,s,t \\ 1 \leq r,s,t,r',s',t' \leq p}} \sum_{\substack{\sigma \\ r',s',t'}} \sigma_{rr'} \sigma_{ss'} \sigma_{tt'} \mu_{rst} \mu_{r's't'}$$

Thus, for a random sample from a multivariate population, S or Σ is computed, and the theoretical moments inherent in each specific distribution are used to set a measure of theoretical skewness. It is clear that if the distribution type changes, so will the moments, and hence the value of skewness. The theoretical measures of skewness, including the derivation of moments, for the multivariate normal, gamma, and lognormal distributions constructed in Chapter III are presented next.

NORMAL

The normal distribution is symmetric, and has a third moment equal to zero; thus, its measure of skewness should also equal zero.

Given:

$$x_1 = a_{11}y_1 + \mu_1$$

$$x_2 = a_{21}y_1 + a_{22}y_2 + \mu_2$$

$$x_3 = a_{31}y_1 + a_{32}y_2 + a_{33}y_3 + \mu_3$$

$$x_4 = a_{41}y_1 + a_{42}y_2 + a_{43}y_3 + a_{44}y_4 + \mu_4$$

It can easily be shown that the following elements of Mardia's skewness equation which apply to the multivariate normal distribution are all equal to zero.

$$E [(x_i - \mu_i)^3] = 0 \quad (4.12)$$

$$E [(x_i - \mu_i)^2(x_j - \mu_j)] = 0 \quad (4.13)$$

$$E [(x_i - \mu_i)(x_j - \mu_j)(x_k - \mu_k)] = 0 \quad (4.14)$$

In equation, 4.12

$$\begin{aligned} E [(x_i - \mu_i)^3] &= E [(a_{i1}y_1 + a_{i2}y_2 + a_{i3}y_3 + a_{i4}y_4)^3] \\ &= E \left[\sum_{j=1}^4 (a_{ij})^3 y_j + 3 \sum_{j=1}^4 \sum_{k \neq j} (a_{ij})^2 a_{ik} y_j^2 y_k \right. \\ &\quad \left. + 6 \sum_{j=1}^4 \sum_{k \neq j} a_{ik} y_k \right] \end{aligned}$$

The first and third terms of the equation on the right are equal to zero since the mean and third moment of y_i are equal to zero. Due to independence among the y_i , and that $E[y_k] = 0$, the second term will also be equal to zero.

In equation 4.13,

$$\begin{aligned}
 E [(x_i - \mu_i)^2 (x_j - \mu_j)] &= E [(a_{i1}y_1 + a_{i2}y_2 + a_{i3}y_3 + a_{i4}y_4)^2 \cdot \sum_{k=1}^4 a_{jk}y_k] \\
 &= E \left[\sum_{k=1}^4 (a_{ik})^2 (y_k)^2 \left(\sum_{k=1}^4 a_{jk}y_k \right) + \right. \\
 &\quad \left. \left(\sum_{k=1}^4 \sum_{\substack{\ell=1 \\ \ell \neq k}}^4 a_{ik}y_k a_{i\ell}y_\ell \right) \left(\sum_{k=1}^4 a_{jk}y_k \right) \right] \\
 &= E \left[\sum_{k=1}^4 (a_{ik})^2 a_{jk} y_k^3 + \sum_{k=1}^4 \sum_{\substack{\ell=1 \\ \ell \neq k}}^4 (a_{ik})^2 y_k^2 a_{j\ell} y_\ell \right. \\
 &\quad \left. + 2 \sum_{k=1}^4 \sum_{\substack{\ell=1 \\ \ell \neq k}}^4 a_{ik} a_{jk} y_k^2 y_\ell + \sum_{k=1}^4 \sum_{\substack{\ell=1 \\ \ell \neq k}}^4 \sum_{m=1}^4 a_{ik} a_{i\ell} a_{jm} y_k y_\ell y_m \right]
 \end{aligned}$$

Once again, every term on the right side of the equality sign is equal to zero since the mean and the third moment of each y_i are equal to zero, and all the y 's are independent. Similarly,

$$E [(x_i - \mu_i)(x_j - \mu_j)(x_k - \mu_k)] = 0 .$$

Substituting equations 4.13, 4.14, and 4.15 into the formula derived for $\beta_{1,4}$ in Appendix VII, it can easily be seen that $\beta_{1,p} = 0$ for $p = 4$.

GAMMA

Given

$$x_1 = a_{11}y_1$$

$$x_2 = a_{21}y_1 + a_{22}y_2$$

$$x_3 = a_{31}y_1 + a_{32}y_2 + a_{33}y_3$$

$$x_4 = a_{41}y_1 + a_{42}y_2 + a_{43}y_3 + a_{44}y_4$$

where each y_i is distributed as gamma (n_i, λ_i) $i=1, \dots, 4$. The mean variance, and third moment of the y_i are respectively

$$E[y_i] = \frac{n_i}{\lambda_i},$$

$$\text{var}[y_i] = \frac{n_i}{\lambda_i^2} \quad \text{and}$$

$$E[(y_i - E[y_i])^3] = \frac{2n_i}{\lambda_i^3}$$

Define $w_i = (y_i - E[y_i])$

Since the y_i are independent, the w_i are also independent. Some properties of w_i are

$$E[w_i] = 0$$

$$E[w_i^2] = \frac{n_i}{\lambda_i^2}$$

$$E[w_i^3] = \frac{2n_i}{\lambda_i^3}$$

The third moment of x_i is

$$\begin{aligned}
 E [(x_i - E[x_i])^3] &= E \left[\left(\sum_{k=1}^i a_{ik} y_k - \sum_{k=1}^i a_{ik} E[y_k] \right)^3 \right] \\
 &= E \left[\left(\sum_{k=1}^i a_{ik} (y_k - E[y_k]) \right)^3 \right] \\
 &= E \left[\left(\sum_{k=1}^i a_{ik} w_k \right)^3 \right] \\
 &= \left[E \left(\sum_{k=1}^i a_{ik}^3 w_k^3 \right) + \text{interaction terms involving one } w_i \text{ to the first power} \right]
 \end{aligned}$$

Since the w_i are independent, the interaction terms are equal to zero ($E[w_i] = 0$). Thus,

$$\begin{aligned}
 E [(x_i - E[x_i])^3] &= E \left[\sum_{k=1}^i a_{ik}^3 w_k^3 \right] \\
 &= \sum_{k=1}^i E [a_{ik}^3 w_k^3] \\
 &= \sum_{k=1}^i a_{ik}^3 \cdot \frac{2n_k}{\lambda_k^3} \\
 &= 2 \sum_{k=1}^i \left(\frac{a_{ik}}{\lambda_k} \right)^3 n_k
 \end{aligned}$$

Similarly,

$$E [(x_i - E[x_i])^2 (x_j - E[x_j])] = \sum_{k=1}^j a_{ik}^2 a_{jk} E[w_k^3] \text{ for } i \neq j$$

and for $i < j < k$,

$$E [(x_i - E[x_i])(x_j - E[x_j])(x_k - E[x_k])] = \sum_{\ell=1}^{\min(i,j,k)} a_{i\ell} a_{j\ell} a_{k\ell} E[w_\ell^3]$$

LOGNORMAL

Given that y is $N(0, \sigma^2)$, then $x = \exp[y]$ is a lognormal distribution and the coefficient of skewness is

$$\frac{E[(x - E(x))^3]}{(\text{var}[x]^{3/2})}$$

For computational purposes, the important elements of Mardia's skewness equation, $\beta_{1,4}$, are based on the following dependencies among the components of the \underline{x} vector:

$$x_1 = \exp[a_{11}y_1 + b_1]$$

$$x_2 = \exp[a_{11}y_1 + a_{22}y_2 + b_2]$$

$$x_3 = \exp[a_{11}y_1 + a_{22}y_2 + a_{33}y_3 + b_3]$$

$$x_4 = \exp[a_{11}y_1 + a_{22}y_2 + a_{33}y_3 + a_{44}y_4 + b_4]$$

where the y_i are independently identically distributed $N(0,1)$.

The elements of Mardia's skewness equation

$$E[x_i] = \exp[b_i] \cdot \exp\left[\frac{1}{2}(a_{11}^2 + \dots + a_{ii}^2)\right]$$

$$E[x_i^2] = \exp[2b_i] \cdot \exp\left[\frac{4}{2}(a_{11}^2 + \dots + a_{ii}^2)\right]$$

$$E[x_i^3] = \exp[3b_i] \cdot \exp\left[\frac{9}{2}(a_{11}^2 + \dots + a_{ii}^2)\right]$$

$$E[(x_i - \mu_i)(x_j - \mu_j)(x_k - \mu_k)]$$

$$= E[x_i x_j x_k] - \mu_i E[x_j x_k] - \mu_j E[x_i x_k] - \mu_k E[x_i x_j] + 2\mu_i \mu_j \mu_k$$

$$E[(x_i - \mu_i)^2(x_j - \mu_j)] = E[x_i^2 x_j] - 2\mu_i E[x_i x_j] - \mu_j E[x_i^2] + 2\mu_i^2 \mu_j.$$

$$E[(x_i - \mu_i)^3] = E[x_i^3] - 3\mu_i E[x_i^2] + 2\mu_i^3.$$

For $i < j$,

$$E[x_i^2 x_j] = \exp[2b_i + b_j] \cdot \exp\left[\frac{1}{2}\left(\sum_{k=1}^i 9a_{kk}^2 + \sum_{k=i+1}^j a_{kk}^2\right)\right]$$

$$E[x_i x_j] = \exp[b_i + b_j] \cdot \exp\left[\frac{1}{2}\left(\sum_{k=1}^i 4a_{kk}^2 + \sum_{k=i+1}^j a_{kk}^2\right)\right]$$

For $i > j$,

$$E[x_i^2 x_j] = \exp[2b_i + b_j] \cdot \exp\left[\frac{1}{2}\left(\sum_{k=1}^j 9a_{kk}^2 + \sum_{k=j+1}^i 4a_{kk}^2\right)\right]$$

$$E[x_i x_j] = \exp[b_i + b_j] \cdot \exp\left[\frac{1}{2}\left(\sum_{k=1}^j 4a_{kk}^2 + \sum_{k=j+1}^i a_{kk}^2\right)\right]$$

For $i < j < l$,

$$E[x_i x_j x_l] = \exp[b_i + b_j + b_l] \cdot \exp\left[\frac{1}{2}\left(\sum_{k=1}^i 9a_{kk}^2 + \sum_{k=i+1}^j 4a_{kk}^2 + \sum_{k=j+1}^l a_{kk}^2\right)\right]$$

These measures of skewness are specific to each distribution type since each distribution has different values for the moments needed in the skewness equation. These results, then, are theoretical. It is also possible to obtain an indication of sample skewness, $b_{1,p}$, which is not dependent on distribution type. This measure is:

$$b_{1,p} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N [(x_i - \bar{x})^{-1} (x_j - \bar{x})]^3$$

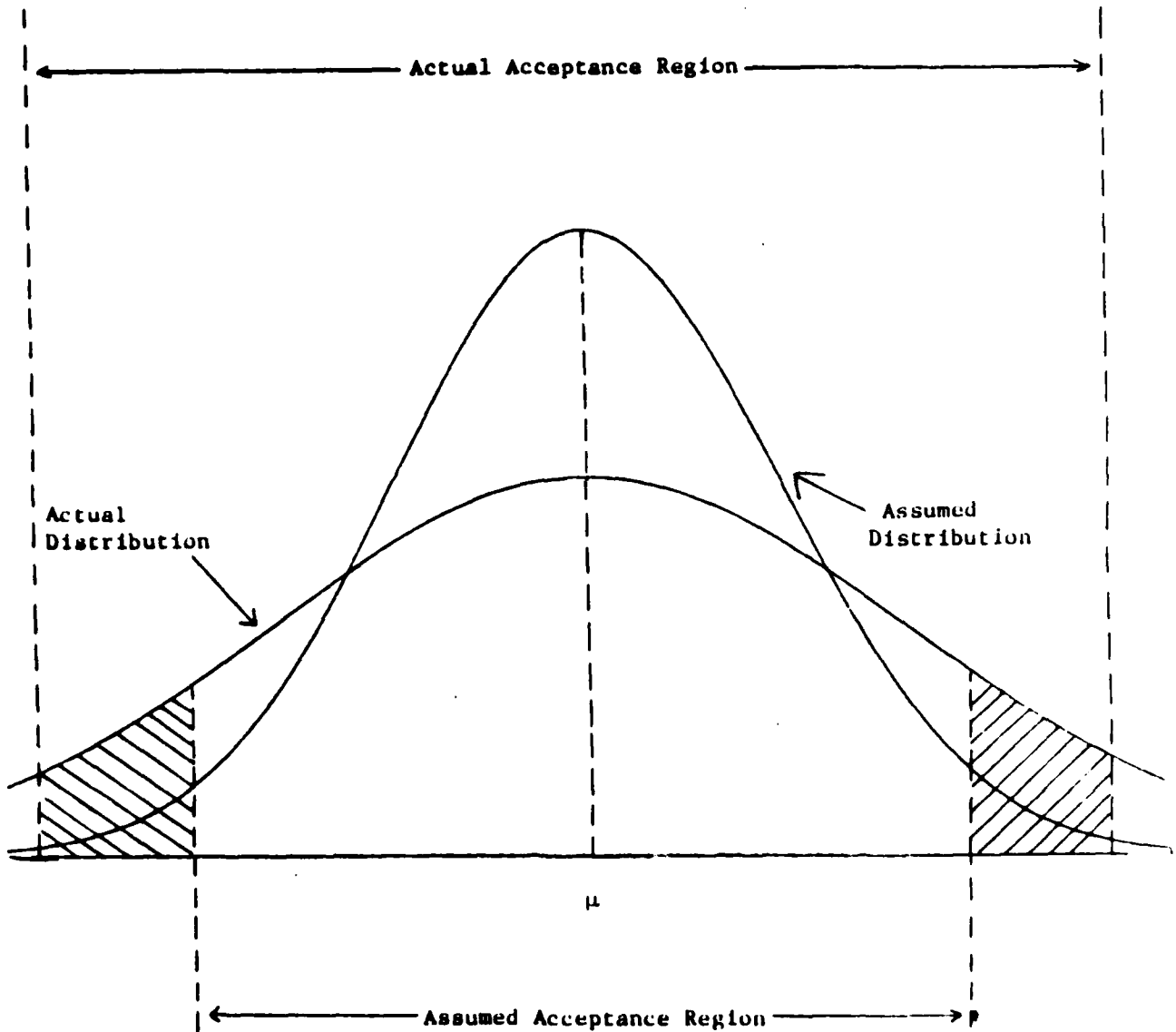
$b_{1,p}$ can be tested against zero to see if the sample is from a symmetric distribution. This test is presented in [Kres, (1983, Table 26)]. By using both the sample and theoretical measures of skewness, the Clinic team examined the robustness of Hotelling's T^2 statistic with respect to distribution type.

Chapter V - Unequal Covariance Matrices

Another assumption that the team relaxed was that of the equality of the covariance matrices. It is important to analyze the robustness of the T^2 -statistic when this assumption fails because there is no statistic derived for the case of unknown and unequal covariance matrices (Case 6). If the test statistic for case 5 is robust, then it would be acceptable for use in case 6; if the statistic is not robust, however, using the wrong statistic would lead to phantoms or other types of misinformation in the Base Data. To illustrate this situation in the univariate case, Figure 9 shows two symmetric distributions with the same means, but the actual distribution of the data has a much greater variance than is assumed, i.e. the actual distribution is "fat-tailed." The effect of this incorrect assumption is an acceptance region which is too small compared to that of the actual distribution. As in the case of Region I, in Figure 8, phantoms may occur because an estimate that falls in the shaded region is from a known entity, but the test will indicate that any estimate that falls outside the assumed acceptance region is a new entity.

The Clinic team altered the equality of covariance matrices by allowing the input of an additional positive definite matrix to be added to the covariance matrix of the test distribution. In this way it is possible to make either a larger or a smaller covariance matrix for the test.

Chapter VII contains the results obtained by relaxing the assumption of equal covariance matrices.



Fat-Tailed Distributions
Figure 9

Chapter VI - Computer Simulation

The programs written for the Claremont Graduate School Mathematics Clinic under the auspices of the Jet Propulsion Laboratory are designed to simulate data from various probability distributions and perform the Hotelling's T^2 -tests to determine whether or not two sets of data are from the same distribution. As described in the report, the T^2 -test assumes that the data from the two populations are normally distributed. This program is used to study the robustness of the test when the data is not properly distributed; in particular, when the data follows a skewed distribution, i.e. a gamma or lognormal. The mathematics of these tests have already been presented. This chapter constitutes a "user's manual" for the programs written by the Clinic team.

Since simulation requires a great deal of computer time, the program INPUT was written to create input data for the simulation program, SIMULATE. With data generated from INPUT, SIMULATE is not interactive and requires no user attention.

Specifically, INPUT sets all of the necessary parameters for SIMULATE to run. These parameters are as follows:

1. The output file for the information from the simulations.
2. The distribution types for the test distribution. Here we have 1=Uniform; *2=Normal; 3=Exponential; *4=Gamma; *5=Lognormal; 6=Weibull; *7=Cauchy; *8=Gamma and Lognormal with the same covariance matrix. The starred numbers are implemented completely in SIMULATE. The other distributions are not complete and are left for further development. The default is a Normal distribution, i.e. if no distribution is specified a normal distribution will be used.

3. The test number depending on the assumptions made about the covariance matrices. The test number is 1 if the covariance matrices are known and equal, 2 if the covariance matrix is estimated from the test data, and 5 if the covariance matrices are unknown and equal. Tests 3, 4, and 6 require further work and are not implemented.
4. The parameters necessary to generate random observations from the specific distribution. The specific method for obtaining observations will be discussed later.
5. The amount by which the theoretical means of the Normal base distribution should be varied from those of the test distribution.
6. The positive definite matrix by which the covariance matrix of the base distribution should be varied from the test distribution.
7. The theoretical mean of the base distribution.
8. The theoretical covariance matrix of the base distribution.
9. The confidence level (1-significance level) for the test where 1=99% and 2=95%.
10. The sample size of the generated distributions. A large sample of 100 or a small sample of 20 are the sizes used. If other sample sizes are desired, then it is necessary to alter the critical test values in the code to reflect the sample size.
11. The number of simulations using the parameters as described in 2 through 10 above.

INPUT allows the user to input data for as many runs as desired so SIMULATE may generate many sets of results. Given the above parameters, SIMULATE may generate the prescribed data and perform the specified tests.

The flow of control of SIMULATE is straightforward. The parameters are

read in, the observations from the specified distributions are generated, and the particular test is performed. Clearly, there are many details to be filled in, but rather than describing every line of code, a description of each procedure is given here.

PROCEDURE RANDOMIZE

This procedure generates a seed integer to be used subsequently in the function, RND, described below. This seed is generated by calling the external system Math Library function FOR\$SECONDS. The output of FOR\$SECONDS is a real value corresponding to the current time given by the computer's internal clock. Since this value is given in milliseconds, it is improbable that the same seed will be returned twice. This greatly improves the "randomness" of this random number generator.

FUNCTION RND

This function is called throughout the program to return a random number in the range (0,1). After every 500 calls, RND will generate a new seed to be used in subsequent calls. RND calls the external system Math Library function MTH\$RANDOM to return the random number. See Appendix IX.

PROCEDURE INPUT_PARAMS

INPUT_PARAMS is the procedure which gets all of the necessary information from the file created by INPUT. The actual parameters read in differ according to the distributions to be generated.

PROCEDURE MAT_OUT_ONE

This procedure outputs an NxN matrix to the output file.

PROCEDURE MAT_OUT_TWO

This procedure outputs two NxN matrices to the output file, side by side.

PROCEDURE INVERT

This is an external procedure written in BASIC which computes the inverse of the matrix passed to it. The BASIC matrix handling capabilities are much

easier to use than trying to invert a matrix using brute force. At this point, no investigation has been made concerning the accuracy of the BASIC inverse routine. If a better routine is found (such as an IMSL routine) it could replace the current inversion procedure.

FUNCTION MAT_MULT_ROW_COLUMN

A row vector and a column vector are sent to this function, and the scalar result is returned.

FUNCTION MAT_MULT_ROW_MATRIX

A row vector and a matrix are sent to this function. The result, a row vector, is returned.

FUNCTION TRANSPOSE

This function returns the row vector which is the transpose of the column vector passed.

FUNCTION DETERM

This function computes the determinant of a 4x4 matrix along with its sub-determinants. The particular determinant calculated depends on whether a 2, 3, or 4 is passed into the parameter SIZE. Each calculation is done by following the general formula for the determinant of a matrix. No recursion is used.

PROCEDURE UNISETS

This procedure generates observations from a bivariate Uniform distribution with parameters (A1, B1) and (A2, B2). This procedure was inserted at the beginning of the program development and was recently updated to build 4-characteristic samples.

PROCEDURE NORMSETS

This procedure generates observations from a Normal distribution with four characteristics. This procedure requires the mean about which the observations

are centered and the matrix which describes the dependencies between the characteristics.

PROCEDURE EXPOSETS

Like UNISETS, this procedure is incomplete. It generates independent observations from an exponential distribution with parameter λ .

PROCEDURE GAMMASETS

This procedure generates 4-variate observations from a Gamma distribution using a sum of exponential random variables.

PROCEDURE LOGNORMAL

This procedure generates 4-variate observations from a lognormal distribution. The methodology used to build dependence between the parameters described in Chapter III is used here. The local procedure AMAKER is used to generate an A matrix and a B vector. These entities are generated randomly between specified upper and lower bounds. These are used to generate the covariance matrix which must be positive definite in order to solve for the dependence matrix required to generate the data from the normal distribution used as the base. A's and B's are generated until the covariance is positive definite.

PROCEDURE CAUCHYSETS

This procedure generates 4-variate observations from a Cauchy distribution. This distribution has "fat-tails" and may be used to test the robustness of the T^2 -squared test on a fat-tailed distribution.

PROCEDURE WEIBULL

This procedure may be used to generate data from a Weibull distribution, but it is not completely coded.

PROCEDURE STAT_TESTS

This is the main procedure of the program. It performs the statistical tests described earlier. The local procedures GAMSKEW and LOGSKEW compute

the theoretical skewness using the moments of the distributions according to Mardia. Within STAT_TESTS, a count of the number of acceptances and rejections is kept to summarize the results of the simulations. This procedure also computes the estimated skewness and kurtosis of the test distribution.

PROCEDURE INFO_OUT

This is the output procedure of the program. It outputs all of the vital information needed to analyze the test statistics. This information includes the test type, the test distribution, the mean vector, and the covariance matrices of the test and base distribution, the parameters used to generate the data, and the dependence matrix. After the summary information is printed once, the user can suppress repeating the summary; in this case, only information that changes is reported after the summary.

As mentioned above, the flow of SIMULATE is rather straightforward. The only section of code that may seem unclear is case 8 in the FOR Z:=1 TO RUNS loop. When the program was originally developed, it was designed to generate data for only one test distribution. When the decision was made to run LOGNORMAL and GAMMA data with the same covariance matrix, it was easier to simply generate lognormal data, perform the tests, generate the gamma data, and then let the program continue normally. Thus, the necessary code was exactly duplicated for running the lognormal tests. This is not necessarily efficient, but it was the best solution given the time constraints.

As a final note, the reader may refer to Appendix VIII for the simulation techniques for generating the univariate random samples.

Chapter VII - Robustness Results

The self-correlation stage uses statistical tests to check if one set of data (New Data) is from the same population as another set of data (Base Data). The tests are based on assumptions about the data, such as distribution type and dependency relationships. The study of robustness concentrates on a statistical test's ability to perform as expected even though the data does not conform to the assumptions upon which the test is based; specifically, if the data from the battlefield are not normal, or the assumptions on the covariance matrices are incorrect, will the tests still perform accurately? These assumptions were addressed by the implementation of the mathematical techniques and simulation program previously described (Chapters III-VI).

The first way in which the clinic team analyzed the robustness of Hotelling's T^2 -statistics was through the use of non-normal distributions. As mentioned in Chapter IV, skewness has a detrimental effect on the performance of these tests in the univariate case. The team used back-solving techniques (Chapter III) to generate multivariate skewed distributions to be used to test the T^2 -statistics for cases 1, 2, and 5. (These were the only cases simulated since case 3 and case 4 do not appear to have relevant applications and no test statistic was derived for case 6.) A description of the 6 cases is contained in Figure 7, Chapter II. There are two ways that the tests could fail to perform properly: the test could reject the null hypothesis when the null hypothesis is really true (Type I error), or it could accept the null hypothesis when it is really false (Type II error).

The Clinic team tested for Type I error by simulating Base Data and New Data with equal mean vector and covariance matrices. At a 5% significance level the null hypothesis should be accepted approximately 95% of the time when it is true. Figure 10 shows the results of the simulation. Since the statistical test

$$\mu_B = \mu_N, \sigma_B \neq \sigma_N$$

EXPECT \geq 95%

DIST. TYPE	CASE		
	1	2	5
NORMAL	94	94	99
GAMMA	0	16	26
LOGNORMAL	73	9	95

Simulation Results 1

Figure 10

are based on the assumption that the data follow normal distributions, they should be accurate when normally distributed data is used. That is, the null hypothesis should be accepted approximately 95% of the time when it is true. The first row shows that in each of the three cases the percentages are as expected. Since the gamma and lognormal distributions are skewed distributions (i.e., not normal) we would expect the tests to fail. The second and third rows indicate that the tests do fail for skewed distributions. Except for the lognormal distribution in case 5, neither the gamma nor the lognormal distributions lead to the acceptance of the null hypothesis more than 63% of the time when it is true; in fact the gamma distribution never did better than 28%. The reasons why the lognormal distribution performed well for case 5 are unclear at this point in time. It is suspected that the use of pooled estimates for covariance matrices from both the New Data and the Base Data could have compensated for the effect of a slightly skewed distribution.

The Type II error, accepting the null hypothesis when it is not true, was tested by generating data with equal covariance matrices but unequal mean vectors. Hence we would expect the null hypothesis of equality of mean vectors to be rejected a large percentage of the time. (That is, the null hypothesis would be accepted a small percentage of the time.) As shown in Figure 11, the null hypothesis was accepted 0% of the time for all three distribution types in cases 1, 2, and 5. Thus, for Type II error simulation, the statistical tests performed well for both the normal and skewed distributions. In other words, the statistical tests appear to be robust with respect to Type II error for the distributions considered.

The final simulation performed generated data from the three distributions with equal mean vectors but with different dependency relationships; thus $\mu_N = \mu_B$ but $\Sigma_N \neq \Sigma_B$. This simulation was used to check the robustness of

$$\mu_B \neq \mu_N, \sigma_B = \sigma_N$$

EXPECT = 0%

DIST. TYPE	CASE		
	1	2	5
NORMAL	0	0	0
GAMMA	0	0	0
LOGNORMAL	0	0	0

Simulation Results 2

Figure 11

equal covariance matrices in case 5. Once again, with equal mean vectors we would expect the null hypothesis to be accepted approximately 95% of the time. Figure 12 shows the results of this simulation. While the normally distributed data performed well, the two skewed distributions did not (98% vs. 23% and 71%). We note that these results are similar to those shown in Figure 10 for case 5. Thus, these preliminary results seem to indicate that small to moderate differences in the two covariance matrices have little effect on the performance of the tests. These results imply that it is the skewness of the distributions that causes the non-robustness of the tests. We note that it is important to analyse robustness of equal covariance matrices since there is no test statistic derived for the case of unknown and unequal covariance matrices (case 6). If the test statistic for case 5 is robust, then it would be acceptable for use in case 6.

In summary, the Clinic team used a simulation approach to test the robustness of Hotelling's T^2 -statistics with respect to normality and equal covariance matrices. Three separate simulations were performed involving normal, gamma, and lognormal distributions. The results of the simulations show that the T^2 -statistical tests are not robust with respect to the normality (or symmetry) assumptions. In other words, if the data being analyzed comes from a skewed distribution and a variation of Hotellings T^2 -statistic is used, the decision made by the test will probably be incorrect. More studies need to be made on the relationship between skewness and the T^2 -statistics. Using T^2 -statistics that assume equal covariance matrices on datasets that actually have slightly different covariance matrices, however, seems to have an insignificant effect on the performance of the tests.

$$\mu_B = \mu_N, \quad \Sigma_B = \Sigma_N$$

EXPECT \geq 95%

DIST. TYPE	CASE
	5
NORMAL	98
GAMMA	23
LOGNORMAL	71

Simulation Results 3

Figure 12

Chapter VIII - Concluding Remarks

In this report the Clinic team first developed the mathematical techniques and simulation program required to generate multivariate distributions with component dependency. These results were then used to investigate the effects of skewed distributions and unequal covariance matrices on the statistical tests in the self-correlation algorithm. (The results of these investigations are contained in Chapter VII.) The Clinic team plans to continue working in this area. In particular, since preliminary results indicate that skewed distributions cause inaccurate conclusions in the statistical tests, the relationship between the robustness of these tests with respect to multivariate skewness and kurtosis will be studied further. This will be the main objective of the next report, Applications of Correlation Techniques in Battlefield Identification II.

There are at least two other areas that clearly should be studied in future reports. One topic would be the investigation of the robustness of the statistical tests with respect to other assumptions. These assumptions include

1. The sensor data is unbiased.
2. The error term of the mean vector is independent of time.
3. The New Data describes only one entity.
4. The entities are stationary.

The Clinic also plans to develop a final computer package to implement results. This package would be a user-friendly process which, among other things, performs goodness-of-fit tests on the data to determine whether or not the proposed statistical tests are appropriate.

APPENDIX I

Finding the Maximum Likelihood Function for $\underline{\mu}, \Sigma^{-1}$

PREVIOUS PAGE IS BLANK 

The likelihood function $L(\underline{\mu}; \Sigma^{-1}; x_1, \dots, x_N)$ is given by

$$L = \frac{1}{(2\pi)^{\frac{1}{2}pN} |\Sigma|^{-\frac{1}{2}N}} \exp \left[-\frac{1}{2} \sum_{\alpha=1}^N (x_{\alpha} - \underline{\mu})' \Sigma^{-1} (x_{\alpha} - \underline{\mu}) \right] \quad (A1.1)$$

To find the maximum likelihood estimates for $\underline{\mu}$ and Σ^{-1} in Equation A1.1 we begin by taking the natural logarithm of Equation A1.1. Denoting $\underline{\mu}^*$ and ψ^* to be the maximum likelihood estimates for $\underline{\mu}$ and Σ^{-1} respectively, we get

$$\ln L = -\frac{1}{2} p N \ln(2\pi) + \frac{1}{2} N \ln |\psi^*| - \frac{1}{2} \sum_{\alpha=1}^N (x_{\alpha} - \underline{\mu}^*)' \psi^* (x_{\alpha} - \underline{\mu}^*) \quad (A1.2)$$

where $\underline{\mu}^*$ and ψ^* maximize $\ln L$.

The following lemma will be useful in solving the above equation for $\underline{\mu}^*$ and ψ^* .

Lemma 1: Let x_1, \dots, x_N be N (p -characteristic) vectors. Then for any vector \underline{b} ,

$$\sum_{\alpha=1}^N (x_{\alpha} - \underline{b})(x_{\alpha} - \underline{b})' = \sum_{\alpha=1}^N (x_{\alpha} - \bar{x})(x_{\alpha} - \bar{x})' + N(\bar{x} - \underline{b})(\bar{x} - \underline{b})'$$

$$\begin{aligned} \text{Proof: } \sum_{\alpha=1}^N (x_{\alpha} - \underline{b})(x_{\alpha} - \underline{b})' &= \sum_{\alpha=1}^N [(x_{\alpha} - \bar{x}) + (\bar{x} - \underline{b})][(x_{\alpha} - \bar{x}) + (\bar{x} - \underline{b})]' \\ &= \sum_{\alpha=1}^N [(x_{\alpha} - \bar{x})(x_{\alpha} - \bar{x})' + (x_{\alpha} - \bar{x})(\bar{x} - \underline{b})' + (\bar{x} - \underline{b})(x_{\alpha} - \bar{x})' + (\bar{x} - \underline{b})(\bar{x} - \underline{b})'] \\ &= \sum_{\alpha=1}^N (x_{\alpha} - \bar{x})(x_{\alpha} - \bar{x})' + \left[\sum_{\alpha=1}^N (x_{\alpha} - \bar{x}) \right] (\bar{x} - \underline{b})' + (\bar{x} - \underline{b}) \sum_{\alpha=1}^N (x_{\alpha} - \bar{x})' + N(\bar{x} - \underline{b})(\bar{x} - \underline{b})' \end{aligned}$$

The second and third terms are equal to zero since $\sum_{\alpha=1}^N (x_{\alpha} - \bar{x}) = \sum_{\alpha=1}^N x_{\alpha} - N(\frac{\sum x_{\alpha}}{N}) = 0$.

Thus, let $\underline{b} = \underline{\mu}^*$ and apply Lemma 1, to obtain that

$$\sum_{\alpha=1}^N (x_{\alpha} - \underline{\mu}^*)(x_{\alpha} - \underline{\mu}^*)' = \sum_{\alpha=1}^N (x_{\alpha} - \bar{x})(x_{\alpha} - \bar{x})' + N(\bar{x} - \underline{\mu}^*)(\bar{x} - \underline{\mu}^*)' = A + N(\bar{x} - \underline{\mu}^*)(\bar{x} - \underline{\mu}^*)' \quad (A1.3)$$

$$\text{where } A = \sum_{\alpha=1}^N (\underline{x}_{\alpha} - \bar{\underline{x}})(\underline{x}_{\alpha} - \bar{\underline{x}})'$$

Recall the properties of the trace of a matrix:

If C is $m \times n$ and D is $n \times m$ then

$$\text{tr}(CD) = \sum_{i=1}^m \sum_{j=1}^n c_{ij} d_{ji} = \text{tr}(DC)$$

$$\text{tr}(C+D) = \text{tr}(C) + \text{tr}(D)$$

Applying Equation AI.3 and the properties of the trace of a matrix to the third term of the right side of Equation AI.2, we have

$$\begin{aligned} \sum_{\alpha=1}^N (\underline{x}_{\alpha} - \underline{\mu}^*)' \psi^* (\underline{x}_{\alpha} - \underline{\mu}^*) &= \text{tr} \sum_{\alpha=1}^N (\underline{x}_{\alpha} - \underline{\mu}^*)' \psi^* (\underline{x}_{\alpha} - \underline{\mu}^*) \\ &= \text{tr} \sum_{\alpha=1}^N \psi^* (\underline{x}_{\alpha} - \underline{\mu}^*) (\underline{x}_{\alpha} - \underline{\mu}^*)' \\ &= \text{tr} \psi^* \sum_{\alpha=1}^N (\underline{x}_{\alpha} - \underline{\mu}^*) (\underline{x}_{\alpha} - \underline{\mu}^*)' \\ &= \text{tr} \{ \psi^* [A + N(\bar{\underline{x}} - \underline{\mu}^*) (\bar{\underline{x}} - \underline{\mu}^*)'] \} \\ &= \text{tr} \{ \psi^* A \} + \text{tr} \{ \psi^* N(\bar{\underline{x}} - \underline{\mu}^*) (\bar{\underline{x}} - \underline{\mu}^*)' \} \\ &= \text{tr} \{ \psi^* A \} + N(\bar{\underline{x}} - \underline{\mu}^*)' \psi^* (\bar{\underline{x}} - \underline{\mu}^*) \end{aligned} \quad (\text{AI.4})$$

With the result of Equation AI.4, Equation AI.2 can be rewritten as

$$\ln L = -\frac{1}{2} p N \ln(2\pi) + \frac{1}{2} N \ln |\psi^*| - \frac{1}{2} \text{tr}(\psi^* A) - \frac{1}{2} N(\bar{\underline{x}} - \underline{\mu}^*)' \psi^* (\bar{\underline{x}} - \underline{\mu}^*) \quad (\text{AI.5})$$

We only need to maximize the second and third terms of the right side of Equation (AI.5) because the first term is a constant, and the fourth term is equal to zero then $\underline{\mu}^* = \bar{\underline{x}}$.

To maximize the second and third terms, which are $\frac{1}{2} N \ln |\psi^*|$ and $-\frac{1}{2} \text{tr}(\psi^* A)$ respectively, we must use the following lemma.

Lemma 2: Let $f(c) = \frac{1}{2}N \ln |c| - \frac{1}{2} \sum_{i,j=1}^p c_{ij} d_{ij}$ where

$c = (c_{ij})$ is positive semidefinite and $D = (d_{ij})$ is positive definite. Then the maximum of $f(c)$ is taken at $C=ND^{-1}$ and the maximum is

$$f(ND^{-1}) = \frac{1}{2}pN(\ln N) - \frac{1}{2}N \ln |D| - \frac{1}{2}pN.$$

Applying Lemma 2 to Equation AI.5 by letting $C=\psi^*$ and $D=A$, we obtain

$$\ln L = -\frac{1}{2}pN \ln (2\pi) + \frac{1}{2}pN \ln N - \frac{1}{2}N \ln |A| - \frac{1}{2}pN. \quad (AI.6)$$

or, when the second and third term of the right side are combined, we obtain

$$\ln L = -\frac{1}{2}pN \ln (2\pi) + \frac{1}{2}N \left[\ln \frac{Np}{|A|} \right] - \frac{1}{2}pN \quad (AI.7)$$

After taking the exponential of Equation AI.7, the result is

$$\max_{\mu, \Sigma^{-1}} L(\underline{\mu}, \Sigma^{-1}) = \frac{1}{(2\pi)^{\frac{1}{2}pN} |\hat{\Sigma}|^{\frac{1}{2}N}} \exp [-\frac{1}{2}pN]$$

$$\text{where } |\hat{\Sigma}|^{\frac{1}{2}N} = \left| \frac{A}{N} \right|^{-\frac{1}{2}N} = N^{\frac{1}{2}pN} |A|^{-\frac{1}{2}N}.$$

APPENDIX II

The Distribution of T^2 for Case 2

Unknown Variance-Covariance Matrix, $H_0: \underline{\mu} = \underline{\mu}_0$.

We wish to find the distribution for $T^2 = N(N-1)(\bar{\underline{x}} - \underline{\mu}_0)' A^{-1} (\bar{\underline{x}} - \underline{\mu}_0)$, where $\bar{\underline{x}}$ is distributed as $N(\underline{\mu}_0, \frac{\Sigma}{N})$. This will be done by several transformations which associate T^2 with a simple and known statistical distribution.

Remember that the matrix A is defined as

$$\begin{aligned} A &= \sum_{\alpha=1}^N (\underline{x}_{\alpha} - \bar{\underline{x}})(\underline{x}_{\alpha} - \bar{\underline{x}})' \\ &= \sum_{\alpha=1}^N \begin{pmatrix} x_{\alpha 1} - \bar{x}_1 \\ \vdots \\ x_{\alpha p} - \bar{x}_p \end{pmatrix} (x_{\alpha 1} - \bar{x}_1, \dots, x_{\alpha p} - \bar{x}_p) \quad j=1, \dots, p \end{aligned}$$

Note that $\begin{pmatrix} x_{\alpha 1} - \bar{x}_1 \\ \vdots \\ x_{\alpha p} - \bar{x}_p \end{pmatrix}$ is not independent because each \bar{x}_j is dependent on the

corresponding $x_{\alpha j}$. This dependency can be eliminated by summing up to $N-1$, f.e., A is distributed independently as $\sum_{\alpha=1}^{N-1} z_{\alpha} z_{\alpha}'$. z_{α} is independent and is distributed as $N(\underline{0}, \Sigma)$. We denote the sample covariance matrix by S.

Our first transformation is to let $T^2 = \underline{y}' S^{-1} \underline{y}$, where $\underline{y} = \sqrt{N}(\bar{\underline{x}} - \underline{\mu}_0)$ and \underline{y} is distributed as $N(\underline{y}, \Sigma)$. The objective of this transformation is to simplify the mean so that it will become zero under the null hypothesis. In notation,

$$\begin{aligned} \underline{y} &= E[\underline{y}] = E[\sqrt{N}(\bar{\underline{x}} - \underline{\mu}_0)] = \sqrt{N} E[(\bar{\underline{x}} - \underline{\mu}_0)] \\ &= \sqrt{N}(E[\bar{\underline{x}}] - \underline{\mu}_0) \\ &= \sqrt{N}(\underline{\mu} - \underline{\mu}_0) \end{aligned}$$

PREVIOUS PAGE
IS BLANK

$$\begin{aligned}
\text{Var}[\underline{y}] &= \text{Var} [\sqrt{N}(\bar{\underline{x}} - \underline{\mu}_0)] = N \text{Var} [(\bar{\underline{x}} - \underline{\mu}_0)] \\
&= N[\text{Var}(\bar{\underline{x}}) + \text{Var}(\underline{\mu}_0) - \text{Cov}(\bar{\underline{x}}, \underline{\mu}_0)] \\
&= N \cdot \text{Var}(\bar{\underline{x}}) = N \cdot \frac{\Sigma}{N} = \Sigma
\end{aligned}$$

Our second transformation is to let D be a nonsingular matrix such that $D\Sigma D' = I$, and define

$$\underline{y}^* = D\underline{y}$$

$$S^* = DSD'$$

$$\underline{\gamma}^* = D\underline{\gamma}$$

T^2 is now equal to $\underline{y}^{*'} S^{*-1} \underline{y}^*$,

$$\begin{aligned}
\text{because } T^2 &= \underline{y}' S^{-1} \underline{y} \\
&= (\underline{y}' I) S^{-1} (I \underline{y}) \\
&= \underline{y}' [D' (D^*)^{-1}] S^{-1} (D^{-1} D) \underline{y} \\
&= (D\underline{y})' (DSD')^{-1} (D\underline{y}) \\
&= \underline{y}^{*'} S^{*-1} \underline{y}^*
\end{aligned}$$

\underline{y}^* is distributed as $N(\underline{\gamma}^*, I)$.

The third transformation is to let the first row of a $p \times p$ orthogonal matrix, Q , be defined by

$$q_{1i} = \frac{y_i^*}{\sqrt{\underline{y}^{*'} \underline{y}^*}} \quad i=1, \dots, p$$

In other words,

$$Q = \left[\begin{array}{cccc} \frac{y_1^*}{\sqrt{\underline{y}^{*'} \underline{y}^*}} & \frac{y_2^*}{\sqrt{\underline{y}^{*'} \underline{y}^*}} & \dots & \frac{y_p^*}{\sqrt{\underline{y}^{*'} \underline{y}^*}} \\ \text{anything so that } Q \text{ is orthogonal} & & & \end{array} \right]$$

The rest of the matrix Q , other than row 1, can be anything as long as orthogonality is maintained.

$$\sum_{i=1}^p q_{1i}^2 = 1 \text{ because } \frac{\sum_{i=1}^p y_i^{*2}}{y^{*'}y^*} = \frac{\sum_{i=1}^p y_i^{*2}}{\sum_{i=1}^p y_i^{*2}} = 1$$

Q , as defined, is a random matrix. What we want to do is to express T^2 in a scalar form rather than in a matrix form. This can be done by letting

$$\begin{aligned} \underline{U} &= Qy^* \\ \underline{B} &= Q(N-1)S^* Q' \end{aligned}$$

With the above definitions, the first component of \underline{U} now becomes

$$\begin{aligned} U_1 &= \sum_{i=1}^p q_{1i} y_i^* \\ &= \sum_{i=1}^p q_{1i} (q_{1i} \sqrt{y^{*'}y^*}) \\ &= \sqrt{y^{*'}y^*} \sum_{i=1}^p (q_{1i})^2 \\ &= \sqrt{y^{*'}y^*} \cdot 1 = \sqrt{y^{*'}y^*} \end{aligned}$$

The remaining components of \underline{U} become

$$\begin{aligned} U_j &= \sum_{i=1}^p q_{ji} y_i^* \quad j \neq 1 \\ &= \sum_{i=1}^p q_{ji} (q_{1i} \sqrt{y^{*'}y^*}) \\ &= \sqrt{y^{*'}y^*} \left(\sum_{i=1}^p q_{ji} q_{1i} \right) \\ &= \sqrt{y^{*'}y^*} \cdot 0 \\ &= 0 \end{aligned}$$

Therefore, $\underline{U} = \begin{pmatrix} U_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$.

Then, $\frac{T^2}{N-1} = \underline{U}' B^{-1} \underline{U}$ (AII.1)

because $\underline{U}' B^{-1} \underline{U} = (QY^*)' (Q(N-1)S^*Q')^{-1} (QY^*)$

$$= Y^{*'} [Q'(Q')^{-1}] \frac{(S^*)^{-1}}{N-1} [Q^{-1}Q] Y^*$$

$$= Y^{*'} \frac{(S^*)^{-1}}{N-1} Y^* \quad (\text{recall } y^* = Dy \text{ and } S^* = DSD)$$

$$= (Dy)' \frac{(DSD')^{-1}}{N-1} (Dy)$$

$$= \frac{y'D'(D')^{-1}S^{-1}D^{-1}Dy}{N-1}$$

$$= \frac{y'S^{-1}y}{N-1} = \frac{T^2}{N-1}$$

Re-expressing Equation AII.1

$$\frac{T^2}{N-1} = (U_1 \ 0 \dots 0) \begin{pmatrix} b^{11} & b^{12} & \dots & b^{1p} \\ b^{21} & b^{22} & \dots & b^{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b^{p1} & b^{p2} & \dots & b^{pp} \end{pmatrix} \begin{pmatrix} U_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$= (U_1 b^{11} \ U_1 b^{12} \ \dots \ U_1 b^{1p}) \begin{pmatrix} U_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$= U_1^2 b^{11} \quad \text{where } (b^{ij}) = B^{-1}$$

In Anderson's notation,

$$\frac{1}{b^{11}} = b_{11} - b_{(1)} B_{22}^{-1} b'_{(1)}$$

$$= b_{11 \cdot 2, \dots, p} \quad \text{where } B = \begin{pmatrix} b_{11} & b_{(1)} \\ b'_{(1)} & b_{22} \end{pmatrix}$$

is a partitioned matrix.

Hence,

$$\frac{T^2}{N-1} = \frac{U_1^2}{b_{11 \cdot 2, \dots, p}} = \frac{y^{*'} y^*}{b_{11 \cdot 2, \dots, p}}$$

The denominator is conditionally distributed as χ^2 (chi-square) with $N-p$ degrees of freedom; it is the sum from 1 to $N-p$ of the square of w where w is distributed as $N(0,1)$. The numerator, on the other hand, is distributed as a noncentral χ^2 with p degrees of freedom and noncentrality parameter $\gamma^* \gamma^* = \gamma^* \Sigma^{-1} \gamma^*$. Thus, $\frac{T^2}{N} \cdot \frac{(N-p+1)}{p}$ is distributed as a noncentral F with p and $N-p+1$ degrees of freedom. The noncentrality parameter is $\delta = \gamma^* \Sigma^{-1} \gamma^*$.

APPENDIX III

Using Likelihood Ratio Criteria to Test the Equality of Variance-Covariance Matrices

Let $\underline{x}_\alpha^{(g)}$ be an observation from the g^{th} population where $\alpha=1, \dots, N_g$ and $g=1, \dots, q$. (q is the total number of populations.)

$\underline{x}^{(g)}$ is $N(\underline{\mu}^{(g)}, \Sigma_g)$ and is a column vector of size p . Let

p = number of characteristics,

N_g = number of observations in the g^{th} population,

N = total number of observations,

$$A_g = \prod_{\alpha=1}^{N_g} (\underline{x}_\alpha^{(g)} - \underline{\bar{x}}^{(g)})' (\underline{x}_\alpha^{(g)} - \underline{\bar{x}}^{(g)}) \quad g=1, \dots, q, \text{ and}$$

A = Sum of A_g 's.

We want to test $H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_q$

The likelihood function is

$$L = \prod_{g=1}^q \frac{1}{(2\pi)^{\frac{1}{2}pN_g} |\Sigma_g|^{\frac{1}{2}N_g}} \exp \left[-\frac{1}{2} \sum_{\alpha=1}^{N_g} (\underline{x}_\alpha^{(g)} - \underline{\mu}^{(g)})' \Sigma_g^{-1} (\underline{x}_\alpha^{(g)} - \underline{\mu}^{(g)}) \right].$$

Define Ω as the parameter space where each Σ_g is positive definite and $\underline{\mu}^{(g)}$ is any vector. Define ω as the parameter space where $\Sigma_1 = \Sigma_2 = \dots = \Sigma_q$ and $\underline{\mu}^{(g)}$ is any vector. Then the maximizing values are

$$\hat{\underline{\mu}}^{(g)} = \underline{\bar{x}}^{(g)}, \hat{\Sigma}_g = \frac{A_g}{N_g} \text{ over } \Omega$$

and

$$\hat{\underline{\mu}}^{(g)} = \underline{\bar{x}}^{(g)}, \hat{\Sigma}_g = \frac{A}{N} \text{ over } \omega$$

Therefore, the likelihood ratio criterion to test H_0 is

$$\lambda = \frac{\prod_{g=1}^q |\hat{\Sigma}_g|^{N_g/2}}{|\hat{\Sigma}_\omega|^{N/2}} = \frac{\prod_{g=1}^q \left| \frac{A_g}{N_g} \right|^{N_g/2}}{\left| \frac{A}{N} \right|^{N/2}}$$

$$= \frac{\prod_{g=1}^q |A_g|^{1/2 N_g}}{|A|^{1/2 N}} \cdot \frac{N^{1/2 p N}}{\prod_{g=1}^q N_g^{1/2 p N_g}} \quad (\text{AIII.1})$$

$\lambda \leq \lambda(\alpha)$ where $\lambda(\alpha)$ is defined such that Equation AIII.1 holds with probability α when H_0 is true.

APPENDIX IV

The Distribution of T^2 for Case 5
Unknown but Equal Variance-Covariance Matrices

$$H_0: \underline{\mu}^{(1)} = \underline{\mu}^{(2)}$$

This derivation uses the results of Appendix II to show that

$$T^2 = \frac{N_1 N_2}{N_1 + N_2} (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})' S^{-1} (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})$$

follows an F distribution with p and $N_1 + N_2 - p - 1$ degrees of freedom. Recall that

Set

$$S = \frac{1}{N_1 + N_2 - 2} \left[\begin{array}{c} 2 \\ \sum_{k=1}^{N_k} \end{array} \sum_{\alpha=1}^p (\underline{x}_{\alpha}^{(k)} - \underline{\bar{x}}^{(k)}) (\underline{x}_{\alpha}^{(k)} - \underline{\bar{x}}^{(k)})' \right]$$

$$\underline{z} = \sqrt{\frac{N_1 N_2}{N_1 + N_2}} (\underline{\bar{x}}^{(1)} - \underline{\bar{x}}^{(2)})$$

Under the null hypothesis \underline{z} is distributed as $N(0, \Sigma)$. Defining \underline{z} this way accomplishes the same goal as the first transformation in Appendix II. Therefore, the remaining transformations of Appendix II, when applied to \underline{z} will yield the desired result. Specifically:

$$T^2 \geq \frac{(N_1 + N_2 - 2)p}{N_1 + N_2 - p - 1} F_{p, N_1 + N_2 - p - 1}(\alpha)$$

where α is the level of significance.

APPENDIX V

Elements of Covariance Matrix of a Lognormal

PREVIOUS PAGE
IS BLANK

Given:

$$x_1 = \exp[a_{11}y_1 + b_1]$$

$$x_2 = \exp[a_{11}y_1 + a_{22}y_2 + b_2]$$

$$x_3 = \exp[a_{11}y_1 + a_{22}y_2 + a_{33}y_3 + b_3]$$

$$x_4 = \exp[a_{11}y_1 + a_{22}y_2 + a_{33}y_3 + a_{44}y_4 + b_4]$$

and the y_i are i.i.d. $N(0,1)$,

and $(\sum_{j=1}^i a_{jj}y_j) + b_i \sim N(b_i, \sum_{j=1}^i a_{jj}^2)$ for $i=1, \dots, 4$.

The covariance matrix elements are:

$$\sigma_{11} = [\exp(2b_1 + a_{11}^2)] \cdot [\exp(a_{11}^2) - 1]$$

$$\sigma_{22} = [\exp(2b_2 + a_{11}^2 + a_{22}^2)] \cdot [\exp(a_{11}^2 + a_{22}^2) - 1]$$

$$\sigma_{33} = [\exp(2b_3 + a_{11}^2 + a_{22}^2 + a_{33}^2)] \cdot [\exp(a_{11}^2 + a_{22}^2 + a_{33}^2) - 1]$$

$$\sigma_{44} = [\exp(2b_4 + a_{11}^2 + a_{22}^2 + a_{33}^2 + a_{44}^2)] \cdot [\exp(a_{11}^2 + a_{22}^2 + a_{33}^2 + a_{44}^2) - 1]$$

$$\sigma_{12} = \sigma_{21} = [\exp(b_1 + b_2 + \frac{2a_{11}^2 + a_{22}^2}{2})] \cdot [\exp(a_{11}^2) - 1]$$

$$\sigma_{23} = \sigma_{32} = [\exp(b_2 + b_3 + \frac{2a_{11}^2 + 2a_{22}^2 + a_{33}^2}{2})] \cdot [\exp(a_{11}^2 + a_{22}^2) - 1]$$

$$\sigma_{24} = \sigma_{42} = [\exp(b_2 + b_4 + \frac{2a_{11}^2 + 2a_{22}^2 + a_{33}^2 + a_{44}^2}{2})] \cdot [\exp(a_{11}^2 + a_{22}^2) - 1]$$

$$\sigma_{34} = \sigma_{43} = \left[\exp(b_3 + b_4 + \frac{2a_{11}^2 + 2a_{22}^2 + 2a_{33}^2 + a_{44}^2}{2}) \right] \cdot$$

$$[\exp(a_{11}^2 + a_{22}^2 + a_{33}^2) - 1]$$

$$\sigma_{13} = \sigma_{31} = \left[\exp(b_1 + b_3 + \frac{2a_{11}^2 + a_{22}^2 + a_{33}^2}{2}) \right] \cdot [\exp(a_{11}^2) - 1]$$

$$\sigma_{14} = \sigma_{41} = \left[\exp(b_1 + b_4 + \frac{2a_{11}^2 + a_{22}^2 + a_{33}^2 + a_{44}^2}{2}) \right] \cdot [\exp(a_{11}^2) - 1]$$

APPENDIX VI

Covariance Matrix Restrictions for Lognormal Case

Proof that the restrictions on the covariance matrix also hold true in the lognormal case. Let

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}$$

The following relationships must be satisfied in order to backsolve for the Gamma or Normal distributions.

$$\frac{\sigma_{12}}{\sigma_{22}} = \frac{\sigma_{13}}{\sigma_{23}} = \frac{\sigma_{14}}{\sigma_{24}} \quad (\text{AVI.1})$$

$$\frac{\sigma_{23}}{\sigma_{33}} = \frac{\sigma_{24}}{\sigma_{34}} \quad (\text{AVI.2})$$

$$\frac{\sigma_{33}\sigma_{44}}{\sigma_{43}^2} > 1 \quad (\text{AVI.3})$$

Using the results of Appendix V, AVI.1 becomes

$$\begin{aligned} & \frac{[\exp(b_1 + b_2 + a_{11}^2 + \frac{a_{22}^2}{2})] \cdot [\exp(a_{11}^2) - 1]}{[\exp(2b_2 + a_{11}^2 + a_{22}^2)] \cdot [\exp(a_{11}^2 + a_{22}^2) - 1]} \\ = & \frac{[\exp(b_1 + b_3 + \frac{2a_{11}^2 + a_{22}^2 + a_{33}^2}{2})] \cdot [\exp(a_{11}^2) - 1]}{[\exp(b_2 + b_3 + \frac{2a_{11}^2 + 2a_{22}^2 + a_{33}^2}{2})] \cdot [\exp(a_{11}^2 + a_{22}^2) - 1]} \end{aligned}$$

$$= \frac{[\exp(b_1 + b_4 + \frac{2a_{11}^2 + a_{22}^2 + a_{33}^2 + a_{44}^2}{2})] \cdot [\exp(a_{11}^2) - 1]}{[\exp(b_2 + b_4 + \frac{2a_{11}^2 + 2a_{22}^2 + a_{33}^2 + a_{44}^2}{2})] \cdot [\exp(a_{11}^2 + a_{22}^2) - 1]}$$

After simplifying each ratio,

$$\exp[b_1 - b_2 - \frac{a_{22}^2}{2}] = \exp[b_1 - b_2 - \frac{a_{22}^2}{2}] = \exp[b_1 - b_2 - \frac{a_{22}^2}{2}]$$

Clearly, relationship AVI.1 holds in lognormal case. For the relationship

$$\frac{\sigma_{23}}{\sigma_{33}} = \frac{\sigma_{24}}{\sigma_{34}}, \quad (\text{AVI.2})$$

$$\begin{aligned} & \frac{[\exp(b_2 + b_3 + \frac{2a_{11}^2 + 2a_{22}^2 + a_{33}^2}{2})] \cdot [\exp(a_{11}^2 + a_{22}^2) - 1]}{[\exp(2b_3 + a_{11}^2 + a_{22}^2 + a_{33}^2)] \cdot [\exp(a_{11}^2 + a_{22}^2 + a_{33}^2) - 1]} \\ &= \frac{[\exp(b_2 + b_4 + \frac{2a_{11}^2 + 2a_{22}^2 + a_{33}^2 + a_{44}^2}{2})] \cdot [\exp(a_{11}^2 + a_{22}^2) - 1]}{[\exp(b_3 + b_4 + \frac{2a_{11}^2 + 2a_{22}^2 + a_{33}^2 + a_{44}^2}{2})] \cdot [\exp(a_{11}^2 + a_{22}^2 + a_{33}^2) - 1]} \end{aligned}$$

or

$$\exp[b_2 - b_3 - \frac{a_{33}^2}{2}] = \exp[b_2 - b_3 - \frac{a_{33}^2}{2}]$$

^ }

For the relationship

$$\frac{\sigma_{33} \sigma_{44}}{\sigma_{43}^2} > 1, \quad (\text{AVI.3})$$

$$\begin{aligned} & \left([\exp(2b_3 + a_{11}^2 + a_{22}^2 + a_{33}^2)] \cdot [\exp(a_{11}^2 + a_{22}^2 + a_{33}^2) - 1] \right) \cdot \\ & \left[\exp(2b_4 + a_{11}^2 + a_{22}^2 + a_{33}^2 + a_{44}^2) \right] \cdot [\exp(a_{11}^2 + a_{22}^2 + a_{33}^2 + a_{44}^2) - 1] \\ & \quad \div \\ & \left([\exp(b_3 + b_4 + a_{11}^2 + a_{22}^2 + a_{33}^2 + \frac{a_{44}^2}{2})] \cdot [\exp(a_{11}^2 + a_{22}^2 + a_{33}^2) - 1] \right) \\ & = \frac{\exp(a_{11}^2 + a_{22}^2 + a_{33}^2 + a_{44}^2) - 1}{\exp(a_{11}^2 + a_{22}^2 + a_{33}^2) - 1} > 1 \end{aligned}$$

Since all three relationships are valid, backsolving from a lognormally generated covariance matrix to a normal or gamma distribution, is possible.

APPENDIX VII

Derivation of $\beta_{1,4}$

Let \underline{x} and \underline{z} be independently and identically distributed random variables Define:

$$\beta_{1,4} = E [(\underline{x} - \underline{\mu})' \Sigma^{-1} (\underline{z} - \underline{\mu})^3]$$

$$= E \left[\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ x_3 - \mu_3 \\ x_4 - \mu_4 \end{pmatrix}' \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix} \begin{pmatrix} z_1 - \mu_1 \\ z_2 - \mu_2 \\ z_3 - \mu_3 \\ z_4 - \mu_4 \end{pmatrix} \right]^3$$

Let $y_i = x_i - \mu_i$ and $w_i = z_i - \mu_i$ for $i = 1, \dots, 4$, then

$$\beta_{1,4} = E \left[\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}' \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} \right]^3$$

$$= E \left[\left(w_1 \sum_{i=1}^4 y_i \sigma^{i1} + w_2 \sum_{i=1}^4 y_i \sigma^{i2} + w_3 \sum_{i=1}^4 y_i \sigma^{i3} + w_4 \sum_{i=1}^4 y_i \sigma^{i4} \right)^3 \right]$$

by matrix multiplication.

Let

$$a = w_1 \sum_{i=1}^4 y_i^{\sigma i1} \quad (\text{AVII.1})$$

$$b = w_2 \sum_{i=1}^4 y_i^{\sigma i2} \quad (\text{AVII.2})$$

$$c = w_3 \sum_{i=1}^4 y_i^{\sigma i3} \quad (\text{AVII.3})$$

$$d = w_4 \sum_{i=1}^4 y_i^{\sigma i4} \quad (\text{AVII.4})$$

The problem restated is

$$\beta_{1,4} = E[(a + b + c + d)^3], \quad (\text{AVII.5})$$

It can be shown that

$$\begin{aligned} (a + b + c + d)^3 &= (a^3 + b^3 + c^3 + d^3) + (3a^2b + 3ab^2 + 3c^2d + 3cd^2 \\ &\quad + 3a^2c + 3ad^2 + 3bc^2 + 3bd^2 + 3ac^2 + 3a^2d \\ &\quad + 3b^2c + 3b^2d) + (6abc + 6abd + 6acd + 6bcd) \end{aligned} \quad (\text{AVII.6})$$

Note that equation AVII.6 consists of three groups of like terms.

Substituting equations AVII.1 through AVII.4 into AVII.6 and expanding:

$$a^3 = w_1^3 \left[\sum_{i=1}^4 (y_i^{\sigma i1})^3 + 3 \sum_{i=1}^4 \sum_{\substack{j=i \\ j \neq i}}^4 (y_i^{\sigma i1})^2 y_j^{\sigma ji} + 6 \sum_{j=1}^4 \sum_{\substack{i=1 \\ i \neq j}}^4 y_i^{\sigma i1} y_j^{\sigma ji} \right]$$

The first group becomes

$$\begin{aligned}
 (a^3 + b^3 + c^3 + d^3) &= \sum_{k=1}^4 w_k^3 \left[\sum_{i=1}^4 (y_{i\sigma}^{ik})^3 + 3 \sum_{i=1}^4 \sum_{\substack{j=1 \\ j \neq i}}^4 (y_{i\sigma}^{ik})^2 y_{j\sigma}^{jk} \right. \\
 &\quad \left. + 6 \sum_{j=1}^4 \sum_{\substack{i=1 \\ i \neq j}}^4 y_{i\sigma}^{ik} \right] \quad (\text{AVII.7})
 \end{aligned}$$

In the second group,

$$\begin{aligned}
 a^2 b &= w_1^2 \left(\sum_{i=1}^4 y_{i\sigma}^{i1} \right)^2 w_2^4 \sum_{i=1}^4 y_{i\sigma}^{i2} \\
 &= w_1^2 w_2^4 \left[\sum_{i=1}^4 (y_{i\sigma}^{i1})^2 + 2 \sum_{j=2}^4 y_{i\sigma}^{i1} y_{j\sigma}^{j1} + 2 \sum_{j=3}^4 y_{2\sigma}^{21} y_{j\sigma}^{j1} \right. \\
 &\quad \left. + 2 y_{3\sigma}^{31} y_{4\sigma}^{41} \right] \sum_{i=1}^4 y_{i\sigma}^{i2}
 \end{aligned}$$

Multiplication and collection of terms yields

$$a^2 b = w_1^2 w_2^4 \left[\sum_{i=1}^4 \sum_{j=1}^4 y_i^{2(\sigma i1)} y_{j\sigma}^{j2} + \sum_{k=1}^4 \sum_{i=1}^4 \sum_{\substack{j=1 \\ j \neq i}}^4 y_{i\sigma}^{i1} y_{j\sigma}^{j1} y_{k\sigma}^{k2} \right]$$

The remaining terms in the second group of equation AVII.7 are calculated similarly, and

$$3[a^2b + a^2c + a^2d + b^2a + b^2c + b^2d + c^2a + c^2b + c^2d + d^2a + d^2b + d^2c]$$

$$= 3 \left[\sum_{i=1}^4 \sum_{\substack{j \neq i \\ j=1}}^4 w_i^2 w_j \left(\sum_{\ell=1}^4 \sum_{k=1}^4 y_k^2 y_{\ell} (\sigma^{ki})^2 \sigma^{\ell j} + \sum_{\ell=1}^4 \sum_{\substack{m=1 \\ k \neq m}}^4 \sum_{k=1}^4 y_m y_k y_{\ell} \sigma^{mi} \sigma^{ki} \sigma^{\ell j} \right) \right] \quad (\text{AVII.8})$$

Finally,

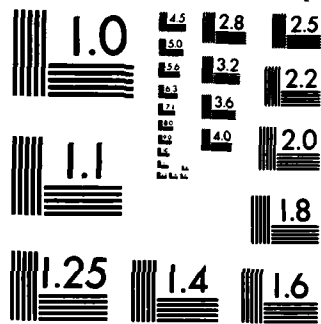
$$\begin{aligned} abc &= w_1 w_2 w_3 \left(\sum_{i=1}^4 y_i \sigma^{i1} \right) \left(\sum_{j=1}^4 y_j \sigma^{j2} \right) \left(\sum_{k=1}^4 y_k \sigma^{k3} \right) \\ &= w_1 w_2 w_3 \sum_{i=1}^4 \sum_{j=1}^4 \sum_{k=1}^4 y_i y_j y_k \sigma^{i1} \sigma^{j2} \sigma^{k3} \end{aligned}$$

Therefore, the last group is

$$\begin{aligned} 6[abc + abd + acd + bcd] &= 6 \left[w_1 w_2 w_3 \sum_{i=1}^4 \sum_{j=1}^4 \sum_{k=1}^4 y_i y_j y_k \sigma^{i1} \sigma^{j2} \sigma^{k3} \right. \\ &\quad + w_1 w_2 w_4 \sum_{i=1}^4 \sum_{j=1}^4 \sum_{k=1}^4 y_i y_j y_k \sigma^{i1} \sigma^{j2} \sigma^{k4} \\ &\quad + w_1 w_3 w_4 \sum_{i=1}^4 \sum_{j=1}^4 \sum_{k=1}^4 y_i y_j y_k \sigma^{i1} \sigma^{j3} \sigma^{k4} \\ &\quad \left. + w_2 w_3 w_4 \sum_{i=1}^4 \sum_{j=1}^4 \sum_{k=1}^4 y_i y_j y_k \sigma^{i2} \sigma^{j3} \sigma^{k4} \right] \quad (\text{AVII.9}) \end{aligned}$$

Substituting equations AVII.7, AVII.8, and AVII.9 into AVII.5, and using the properties of the expected value operator, then

$$\begin{aligned}
 \beta_{1,4} = & \sum_{k=1}^4 E[w_k^3] \left\{ \sum_{i=1}^4 E[y_i^3] (\sigma^{ik})^3 + 3 \sum_{i=1}^4 \sum_{\substack{j=1 \\ j \neq i}}^4 E[y_i^2 y_j] (\sigma^{ik})^2 \sigma^{jk} \right. \\
 & + 6 (E[y_1 y_2 y_3] \sigma^{1k} \sigma^{2k} \sigma^{3k} + E[y_1 y_2 y_4] \sigma^{1k} \sigma^{2k} \sigma^{4k} + E[y_1 y_3 y_4] \sigma^{1k} \sigma^{3k} \sigma^{4k} + \\
 & \left. + E[y_2 y_3 y_4] \sigma^{2k} \sigma^{3k} \sigma^{4k}) \right\} \\
 & + 3 \left\{ \sum_{i=1}^4 \sum_{\substack{j=1 \\ j \neq i}}^4 E[w_i^2 w_j] \left(\sum_{\ell=1}^4 \sum_{k=1}^4 E[y_k^2 y_\ell] (\sigma^{ki})^2 \sigma^{\ell j} \right. \right. \\
 & \left. \left. + \sum_{\ell=1}^4 \sum_{m=1}^4 \sum_{\substack{k=1 \\ k \neq m}}^4 E[y_m y_k y_\ell] \sigma^{mi} \sigma^{ki} \sigma^{\ell j} \right) \right\} \\
 & + 6 \left\{ E[w_1 w_2 w_3] \sum_{i=1}^4 \sum_{j=1}^4 \sum_{k=1}^4 E[y_i y_j y_k] \sigma^{i1} \sigma^{j2} \sigma^{k3} \right. \\
 & + E[w_1 w_2 w_4] \sum_{i=1}^4 \sum_{j=1}^4 \sum_{k=1}^4 E[y_i y_j y_k] \sigma^{i1} \sigma^{j2} \sigma^{k4} \\
 & + E[w_1 w_3 w_4] \sum_{i=1}^4 \sum_{j=1}^4 \sum_{k=1}^4 E[y_i y_j y_k] \sigma^{i1} \sigma^{j3} \sigma^{k4} \\
 & \left. + E[w_2 w_3 w_4] \sum_{i=1}^4 \sum_{j=1}^4 \sum_{k=1}^4 E[y_i y_j y_k] \sigma^{i2} \sigma^{j3} \sigma^{k4} \right\}
 \end{aligned}$$



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

APPENDIX VIII

Generation of Univariate Random Samples

Most computer systems only supply random numbers from the Uniform distributions: $U[0,1)$. The best of these random number generators is discussed in Appendix IX. To generate data from multivariate distributions, it is necessary to use observations from other univariate distributions. This appendix presents methods to generate data from the following distributions: $U(a,b)$, $E(\lambda)$, $W(b,c)$, $N(\mu,\sigma)$, $LN(\gamma,\beta)$, and $C(\alpha,\beta)$.

UNIFORM

The uniform distribution has a distribution function

$$U(a,b) = \frac{x-a}{b-a}$$

A standard technique used for many distributions is Inversion. Recall that any distribution, $r = F(x)$, is distributed $U[0,1)$. Therefore if F is invertible, $x = F^{-1}(r)$ will be from the desired distribution. Specifically

$$x = a + (b-a)r$$

EXPONENTIAL

The exponential distribution also uses the Inversion technique to generate a random observation given an observation from a $U(0,1)$. The exponential is

$$E(\lambda) = 1 - \exp(-\lambda x)$$

Set $r = F(x)$. Since r is random between 0 and 1, so is $1-r$. Thus

$$r^* = 1-r = \exp(-\lambda x)$$

$$\ln(r^*) = -\lambda x$$

$$x = \frac{-\ln(r^*)}{\lambda}$$

WEIBULL

The Weibull distribution is given by

$$W(b,c) = 1 - \exp\left[-\left(\frac{x}{b}\right)^c\right]$$

By the same analysis as before

$$x = b(-\ln r)^{1/c}$$

NORMAL

Some distributions cannot be inverted easily. These must be handled by other methods. The Normal distribution is the most important in our study. Let r_1 , and r_2 be observations from a $U[0,1)$.

$$x_1 = \sigma\sqrt{-2\ln r_1} \cos(2\pi r_2) + \mu$$

$$x_2 = \sigma\sqrt{-2\ln r_1} \sin(2\pi r_2) + \mu$$

are two observations from a $N(\mu, \sigma)$. (Note that the operators sin and cos act on radian arguments.)

LOGNORMAL

The Lognormal distribution can be generated easily once we have two normal observations. Thus, to generate from a $\text{Ln}(\gamma, \beta)$, set

$$\mu = \ln \gamma - \frac{1}{2} \ln\left(\frac{\beta}{\gamma^2} + 1\right)$$

$$\sigma^2 = \ln\left(\frac{\beta}{\gamma^2} + 1\right)$$

Then, if $n_1, n_2 \sim N(\mu, \sigma)$

$$y_1 = e^{n_1}$$

and

$$y_2 = e^{n_2}$$

are from the desired distribution.

GAMMA

The gamma distribution is generated by summing observations from the exponential distribution. Specifically

$$\sum_{i=1}^r E(\lambda) \sim G(\lambda, r)$$

CAUCHY

The Cauchy is a symmetric, fat tailed distribution with parameters median α and scale β . α gives a measure of location while β provides an indication of dispersion. The probability density function for the Cauchy is

$$C(\alpha, \beta) = \frac{1}{\pi\beta} \left[1 + \left(\frac{x - \alpha}{\beta} \right)^2 \right]^{-1}$$

The following generation techniques will lead to the generation of $C(r, q)$, where r is real and q is rational.

Let N_1 and N_2 be independently identically distributed $N(0, 1)$. Then

$$\frac{N_1}{N_2} \sim C(0, 1)$$

Next, use the fact that

$$\sum_{i=1}^n C(a_i, b_i) \sim C\left(\sum_{i=1}^n a_i, \sum_{i=1}^n b_i\right)$$

to see that

$$\sum_{i=1}^k C(0, 1) \sim C(0, k)$$

Another useful relationship is that

$$[C(0, b)]^{-1} \sim C\left(0, \frac{1}{b}\right)$$

By combining the preceding techniques, a $C(0, q)$ where $q = \frac{\ell}{k}$ is rational can be generated as follows:

$$\sum_{i=1}^{\ell} \left[\sum_{i=1}^k C(0, 1) \right]^{-1} \sim C\left(0, \frac{\ell}{k}\right)$$

The final transformation is that for any real number, r ,

$$C(0, q) + r \sim C(r, q)$$

Thus, in order to generate a random sample from a $C(r, q)$, use the following transformation

$$\left(\sum_{i=1}^{\ell} \left[\sum_{i=1}^k \frac{N_1}{N_2} \right]^{-1} \right) + r \sim C(r, q)$$

APPENDIX IX

An Evaluation of Several Random Number Generators

When we began performing the simulations of normally distributed data, some anomalies were present in our results. Consequently, it was decided to test the quality of the random number generator used in the simulation program, and determine if there were any alternative generators which performed better.

The following random number generators were studied:

- 1) The standard VAX BASIC random number generator.
- 2) MTH-RANDOM a VAX Run-Time Library procedure.
- 3) GGUW, the IMSL routine for generating random numbers with shuffling.
- 4) The algorithm $RND=(25173*RND+13849) \text{ MOD } 65536$, which is included in Peter Grogono's book PROGRAMMING IN PASCAL.
- 5) The algorithm $RND=(1061*RND+9533) \text{ MOD } 65536$, which was developed by G. Silberberg for use as part of a previous project.

A Kolmogorov-Smirnov test was used to measure the randomness of the generators. The Kolmogorov-Smirnov test is a standard statistical procedure to determine whether a set of data can be generated from a specified distribution. The K-S statistic D is defined as

$$D = \max \{F(i/n) - F_0(i), F_0(i) - F((i-1)/n)\}$$

where $F_0(i/n)$ is the i th ordered observation in the data set and F is the distribution function which the data is to be tested against.

If D is greater than $1.22/\sqrt{n}$ (where n is the number of observations in the data set), then it can be stated with 90% confidence that the data

does not follow the given distribution.

The following results were obtained using a sample of 199 numbers. 50 trials were performed for each algorithm.

Generator	Average D	# Trials Randomness Rejected
BASIC	.06143	4
VAX RTL	.06308	8
IMSL	.06058	5
Grogono's	.25837	50
Silberberg's	.07324	10

Another set of trials was performed, this time with a sample size of 2000.

Generator	Average D	# Trials Rejected out of 20
BASIC	.01269	0
VAX RTL	.01419	0
IMSL	.02314	4
Grogono's	.31935	20
Silberberg's	.07074	20

Conclusions: The BASIC random number generator showed the best performance, and was placed in our simulation program. Unfortunately, we found problems in sending the random numbers from BASIC to Pascal. The system generator used by BASIC is MTH\$RANDOM; therefore, we used this system function to generate random numbers with the Pascal program. The RTL generator performed nearly as well, and may be appropriate for use in certain circumstances. The others, especially Grogono's algorithm, should be avoided.

Bibliography

T^2 -Statistics --

Anderson, T. W., An Introduction to Multivariate Statistical Analysis.
New York: John Wiley & Sons, Inc., 1958.

This is perhaps the best source for deriving the T^2 statistics. Cases 2, 4, and 5 are discussed in depth. Additionally, the text presents the methods for testing the equality of variance-covariance matrices.

Johnson, Norman and Leone, Fred, Statistics and Experimental Design,
Vol. II. New York: John Wiley & Sons, Inc., 1964.

First text studied by the Clinic Team. Cases 1 and 3 are developed, although the proofs are somewhat cursory.

Simulation --

Naylor, Thomas H.; Balintfy, Joseph L.; Burdick, Donald S. and Chu, Kong,
Computer Simulation Techniques.

New York: John Wiley & Sons, Inc., 1958.

This book presents the univariate simulation techniques to generate the multivariate distributions in Chapter III.

Skewness --

Kres, Heinz, Statistical Tables for Multivariate Analysis.

New York: Springer - Verlag, 1983 ed.

This book contains a summary of results on sample skewness and kurtosis. It also has useful references.

Mardia, K. V., "Measures of Multivariate Skewness and Kurtosis with Applications." Biometrika, Vol. 57, pp. 519-530, 1970.

A standard reference on theoretical measures of multivariate skewness and kurtosis.

Mardia, K. V., "Assessment of Multinormality and Robustness of Hotelling's T^2 Test." Applied Statistics, Vol. 24, pp. 163-171, 1975.

This article indicates that skewness has more detrimental effects on the T^2 tests than kurtosis.

Other Useful Sources --

Chatfield, C., and Collins, A. J., Introduction to Multivariate Analysis. London: Chapman and Hall, 1980

Chapter two gives a conceptual understanding of multivariate use and application. Chapter six presents linear algebra proofs and applications of multivariate techniques.

Gillis, J. W., Griesel, M. A., and Radbill, J. R., Correlation Algorithm Report. Pasadena, Ca.: Jet Propulsion Laboratory, 1982.

Report generated at JPL to catalogue the progress made analysing intelligence gathering systems.

Grogono, Peter, Programming in PASCAL. Reading, Mass.: Addison-Wesley Publishing Company, Inc., 1980

This is the Clinic Team's choice for a Pascal language reference source.

Hastings, N. A. J. and Peacock, J. B., Statistical Distributions. Halsted Press, New York, 1975 ed.

This handbook contains definitions, properties and other useful information concerning the statistical distributions used in this report.

Scheffe, Henry, The Analysis of Variance. New York: John Wiley and Sons, Inc., 1959.

This text will aid us in analyzing what happens when assumptions on the model fail. Specifically, Chapter 10 handles some effects of nonnormality, and statistical dependence.

END

FILMED

7-85

DTIC