Estimation of the covariance matrix in presence of non-detects

Shanmugapriya Saravanan
*University of Nevada, Las Vegas*

Follow this and additional works at: [https://digitalscholarship.unlv.edu/rtds](https://digitalscholarship.unlv.edu/rtds)

Repository Citation

[https://digitalscholarship.unlv.edu/rtds/1906](https://digitalscholarship.unlv.edu/rtds/1906)
ESTIMATION OF THE COVARIANCE MATRIX IN PRESENCE OF NON-DETECTS

by

Shanmugapriya Saravanan

Master of Science
Bharathiar University, India
2003

A thesis submitted in partial fulfillment
of the requirements for the

Master of Science Degree in Mathematical Sciences
Department of Mathematical Sciences
College of Sciences

Graduate College
University of Nevada, Las Vegas
December 2005
The Thesis prepared by

Shanmugapriya Saravanan

Entitled

Estimation of the Covariance Matrix in Presence of Non-detects

is approved in partial fulfillment of the requirements for the degree of

Master of Science

Examination Committee Chair

Dean of the Graduate College
ABSTRACT

Estimation of the Covariance Matrix in Presence of Non-detects

by

Shanmugapriya Saravanan

Dr. Ashok K. Singh, Examination Committee Chair
Department of Mathematical Sciences
University of Nevada, Las Vegas

Environmental engineers and scientists often encounter trace level concentrations of contaminants that fall below the detection limit (DL) are reported as ‘less than DL’. The common method of dealing with non-detects in a multivariate sample is the substitution method in which each observation ‘less than DL’ is replaced by a suitable value (0, DL/2 or DL) with DL/2 being the recommended method.

In this thesis, the applicability of Kaplan-Meier method, which was developed for estimating the survival function in the presence of right censored data, is investigated. Some simulated examples are presented.
# TABLE OF CONTENTS

ABSTRACT ................................................................. iii

LIST OF TABLES .......................................................... v

LIST OF FIGURES ......................................................... vi

ACKNOWLEDGMENTS ................................................... vii

CHAPTER 1 INTRODUCTION ........................................ 1
  The Problem Considered in This Thesis ....................... 3
  Thesis Outline ....................................................... 4

CHAPTER 2 BACKGROUND CONCEPTS ....................... 5
  Survival Analysis .................................................... 5
  Kaplan Meier Method .............................................. 7
    Notations and Definitions ................................... 8
    Example ........................................................... 8
    Computer Implementation .................................... 11

CHAPTER 3 METHODOLOGY .......................................... 13
  Motivation ............................................................ 13
  Concepts ............................................................. 14
  Approach ............................................................. 15
    Explanation of the Program ................................ 15
    Estimation of $\hat{\mu}_i$ ....................................... 16
    Estimation of $\hat{\mu}'_{ij}$ ...................................... 16
    Pseudo Code for the Program .............................. 18

CHAPTER 4 RESULTS ................................................... 21
  Data Description .................................................. 21
  Example ............................................................. 22
  Comparison of the Results .................................... 24

CHAPTER 5 CONCLUSIONS AND FUTURE WORK ............. 46

BIBLIOGRAPHY .......................................................... 47

VITA ................................................................. 48
LIST OF TABLES

2.1 Example of a Recorded Life testing Data Set .............................................. 6
2.2 Left-Censored Environmental Data Set ..................................................... 8
2.3 Right-Censored Environmental Data Set ................................................... 9
2.4 K-M Implementation ................................................................. 10

4.1 Complete Sample ........................................................................... 22
4.2 Sample With Non-detects ..................................................................... 23
4.3 Estimates of \( \Sigma_1 \) and \( \mu_1 \) for 50 observations ......................... 27
4.4 Estimates of \( \Sigma_1 \) and \( \mu_1 \) for 100 observations ......................... 29
4.5 Estimates of \( \Sigma_2 \) and \( \mu_2 \) for 25 observations ......................... 31
4.6 Estimates of \( \Sigma_2 \) and \( \mu_2 \) for 50 observations ......................... 34
4.7 Estimates of \( \Sigma_2 \) and \( \mu_2 \) for 100 observations ......................... 37
4.8 Estimates of \( \Sigma_3 \) and \( \mu_3 \) for 25 observations ......................... 40
4.9 Estimates of \( \Sigma_3 \) and \( \mu_3 \) for 50 observations ......................... 42
4.10 Estimates of \( \Sigma_3 \) and \( \mu_3 \) for 100 observations ......................... 44

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
LIST OF FIGURES

2.1 K-M Estimates using Minitab14 .............................................. 12
2.2 Survival Plot using Minitab .................................................... 12

4.1 Comparison Graph for Estimates of $\mu_1$ for 25 Observations .......... 26
4.2 Comparison Graph for Estimates of $\sigma_1$ for 25 Observations .......... 26
4.3 Estimates of $\mu_1$ for 50 Observations ....................................... 28
4.4 Estimates of $\Sigma_1$ for 50 Observations ....................................... 28
4.5 Estimates of $\mu_1$ for 100 Observations ....................................... 30
4.6 Estimates of $\Sigma_1$ for 100 Observations ....................................... 30
4.7 Estimates of $\mu_2$ for 25 Observations ....................................... 32
4.8 Estimates $\sigma_{11}$ of $\Sigma_2$ for 25 Observations .............................. 32
4.9 Estimates of $\sigma_{12}$ and $\sigma_{13}$ of $\Sigma_2$ for 25 Observations ............. 33
4.10 Remaining estimates of $\Sigma_2$ for 25 Observations ....................... 33
4.11 Estimates of $\mu_2$ for 50 Observations ....................................... 35
4.12 Estimates $\sigma_{11}$ of $\Sigma_2$ for 50 Observations .............................. 35
4.13 Estimates of $\sigma_{12}$ and $\sigma_{13}$ of $\Sigma_2$ for 50 Observations ............. 36
4.14 Remaining estimates of $\sigma_2$ for 50 Observations ....................... 36
4.15 Estimates of $\mu_2$ for 100 Observations ....................................... 38
4.16 Estimates $\sigma_{11}$ of $\Sigma_2$ for 100 Observations .............................. 38
4.17 Estimates of $\sigma_{12}$ and $\sigma_{13}$ of $\Sigma_2$ for 100 Observations ............. 39
4.18 Remaining estimates of $\Sigma_2$ for 100 Observations ....................... 39
4.19 Estimates of $\mu_3$ for 25 Observations ....................................... 41
4.20 Estimates of $\Sigma_3$ for 25 Observations ....................................... 41
4.21 Estimates of $\mu_3$ for 50 Observations ....................................... 43
4.22 Estimates of $\Sigma_3$ for 50 Observations ....................................... 43
4.23 Estimates of $\mu_3$ for 100 Observations ....................................... 45
4.24 Estimates of $\Sigma_3$ for 100 Observations ....................................... 45
ACKNOWLEDGMENTS

It is a great pleasure to thank the many people whose help and suggestions were so valuable in completing this thesis. I am grateful to my advisor, Dr. Ashok K. Singh for his support, encouragement and guidance throughout my work, without his help being provided, the completion of this thesis would not have been possible.

I would like to give my special thanks to Dr. Rohan Dalpatadu, Dr. Dieudonne D. Phanord and Dr. Laxmi P. Gewali for their participation in my thesis committee. Their advice and patience is appreciated. I would also like to thank the College of Sciences at the University of Nevada Las Vegas for offering me the financial support throughout my graduate studies.

This thesis is dedicated to my parents, my brother and my uncle for their constant support, encouragement and prayers. I would also like to thank my friends Ellen, Shankar and Atit for their encouragement and help throughout my studies here. I would like to thank all my friends for their support and friendship.
CHAPTER 1

INTRODUCTION

Statistics is the methodology, which scientists and mathematicians have developed for interpreting and drawing conclusions from data. In nature many of the data are multivariate in the sense that any one particular phenomenon we would like to study most often depends on a number of other factors. For example the environmental impact of exhaust gases from vehicles and recognition of a chemical structure certainly depend on several factors. In almost all disciplines of science and engineering the data is often multivariate. Hence there arises a need to understand the relationship between many variables, which can be done by multivariate data analysis.

In many cases of data analysis in various fields of applications the data subject to analysis are not acquired as initially planned. Data that have not been obtained in complete form as intended are called incomplete data or censored data. Censored data are those whose measured properties are not known precisely, but are known to lie above or below some limit of sensitivity called detection limit, which is the lowest value that a measurement can be detected with a reasonable degree of accuracy[5].

A dataset that contains a percentage of observations whose measurements are less than DL is called left censored data and a data set that contains a percentage of
observations whose measurements are greater than DL are called right-censored data. In environmental studies censored data are commonly encountered as values below a detection limit, called as “less thans” or “non-detects”. These low values are contaminants such as trace metals or organic compounds that are known imprecisely[6].

Non-detects is an issue in several disciplines like marine studies, sediment chemistry, and studies of ground water quality within the environmental sciences. There are not many ways of statistical analysis to process the data with non-detects. The methods for extracting information from the datasets that include non-detects are

(i) **Substitution Methods** This method computes statistics for uncensored data by replacing values as functions of the recorded detection limit for every non-detect. Substitution methods are widely used but have no theoretical basis. The non-detect (<DL) can be substituted with DL, DL/2 or 0. The values substituted many times do not carry any relation to the original value therefore it results in huge variation.

(ii) **Maximum Likelihood Estimation** In the late 1950s and early 60s, several papers in statistical journals by A.C. Cohen [2], introduced maximum likelihood estimation (MLE) for determination of the mean and standard deviation of censored data. MLE uses both the uncensored (detected) observations, along with the proportion of data below one of more censoring thresholds (detection limits) to compute statistics for the entire data set. MLE requires that the distribution of the data be known with, normal and lognormal distributions commonly used in environmental work. MLE methods are computed by maximizing the
likelihood function L

\[ L = \prod_A f(x) \prod_B F(x). \]

where \( A \) = Set of complete observations.

\( B \) = Set of left censored observations.

\( f(x) \) = the PDF of the detected observations.

\( F(x) \) = the CDF for the left-censored observations.

The most important consideration for this method is that how well the data can fit the assumed distribution, especially when the data set is small, as it becomes difficult even to estimate the parameter reliably\(^6\).

(iii) **Nonparametric Methods:** Nonparametric methods provide an alternative series of statistical methods that require no or very limited assumptions to be made about the form of the data distribution. Data can be analyzed without assuming any underlying distribution. Therefore these methods are useful if the data has non-detects.

1.1 The Problem Considered in This Thesis

Environmental datasets contain censored observations in many situations. The above mentioned methods are being used to analyze datasets having non-detects. One nonparametric method used in survival analysis is called Kaplan Meier(KM) method\(^8\). The usage of Kaplan Meier method on censored data for estimation of the population mean has been explained by Helsel\(^6\). However when the data is
multivariate, substitution methods are the commonly used methods. The substitution methods on censored multivariate environmental data sets have been investigated in statistical literature, and it was found that substitution with DL/2 gave better results in certain situations[5]. The purpose of this thesis is to develop a Kaplan-Meier type estimate of the covariance matrix $\Sigma$ when the sample has non-detects for a subset of variables. The Kaplan Meier method algorithm for multivariate data is developed, implemented and its usage is demonstrated on various multivariate data sets.

1.2 Thesis Outline

In Chapter 2, we give an overview of the Kaplan-Meier method and how it can be used on multivariate data. In Chapter 3, the algorithm and how the non-detects are handled in the program are explained. In Chapter 4, some simulated examples are presented, the KM estimate for each of these examples is found using the program developed and compared with the estimates obtained other substitution methods followed by Conclusion and some suggestions for future work.
CHAPTER 2

BACKGROUND CONCEPTS

In this chapter, some basic concepts based on which the implementation has been done are explained. In Section 2.1 the concept of Survival Analysis has been introduced and in Section 2.2 Kaplan Meier method has been explained with an example.

2.1 Survival Analysis

Survival analysis is a set of procedures by which time-to-event censored data can be analyzed. Time-to-event data is the data that has the time when an even occurs as endpoint like the time until an electrical component fails, and time of first re- occurrence of a disease (after first treatment). The distinguishing feature of survival analysis is that it incorporates censoring. Censoring occurs when we have some information about the individual survival time but we do not know the time exactly. Most of lifetime datasets contain observations that are censored. For example, we may observe that an individual survives to \( t \) but then he may die due to a different cause than the one we are studying or he may drop out of the study, or the study might end.

This can be better explained by the following scenario. Consider a linear regression
problem. An epidemiologist wishes to determine how the longevity or "survival time" (the dependent variable) depends on the number of cigarettes smoked per day (the independent variable). The experiment lasts 10 years, during which some individuals die and others do not. The survival time of the living individuals is only known to be greater than their age when the experiment ends. These values are therefore "right censored data points".

To explain this better, an example is presented. A researcher in a hospital observes the time of death in days of 8 patients who have been diagnosed with a terminal disease. One of the patients leaves the hospital at day 5 and contact is lost. The event of interest is death and since the death of the patient is only known to be greater than 5 days, day 5 is recorded and is labeled right-censored. At the end of the 40 day study, the day of death of 4 patients were recorded and 3 patients are still alive. How does the researcher record the day of death for the 2 patients that are still alive at the end of the study? What is common is to record day 40 for both and label them as right-censored[1].

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Time of Death (t)</th>
<th>Censor (0=censored, 1=uncensored)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

The statistical analysis of lifetime data is called survival analysis. Let $T$ denote a
nonnegative random variable \( (T>0) \) representing the lifetimes of individuals in some population and if \( T \) is continuous and \( F(t) \) denote the cumulative density function of \( T \) with corresponding probability density function \( f(t) \), where \( f(t) > 0 \) for \( t>0 \).

Then \( F(t) = P(T \leq t) = \int_0^t f(x)dx \)

The probability that an individual survives to time \( t \) is given by the survival function or reliable function.

\[
S(t) = P(T \geq t) = 1 - F(t) = \int_t^\infty f(x)dx
\]

The survival function \( S(t) \) is a monotone decreasing function with \( S(0) = 1 \) and \( S(\infty) = 0 \).

Expected value of \( T \):

The mean or expected value of \( T \) is computed by calculating the area underneath the survival curve\[11\].

\[
E(X) = \int_0^\infty [1 - F(x)]dx
\]

2.2 Kaplan Meier Method

Kaplan-Meier (K-M) method is the standard non parametric method for estimating the mean from censored survival data. The Kaplan-Meier estimator, also known as the product limit estimator\[8\], can be used to calculate values for nonparametric reliability or data sets with censored observations. KM estimates of survival functions are widely used in many clinical studies and are implemented in commercial statistics packages (such as MINITAB, SAS, SPSS) offering routines for survival analysis.
2.2.1 Notations and Definitions

Let $x_1, \ldots, x_n$ be independently identically distributed (iid) variables which represent failure times subject to right censoring times $y_1, \ldots, y_n$. Then the observed data consist of survival times $(t_i, \delta_i) : i = 1, \ldots, n$ where $t_i = \text{Min}(x_i, y_i)$ and $\delta_i = I(t_i = x_i)$. Let $r$ be the number of distinct failure times, which are ordered as $t_1, \ldots, t_r$.

Define $n_j$ to be the number of individuals surviving up to time $t(j)$, i.e., the number of individuals at risk at time $t(j)$, and $d_j$ as the number of failures at time $t(j)$.

The KM estimate the survivor function $S(t) = P(x_i > t)$ is given by:

$$\widehat{S(t)} = \prod_{j \leq t(j)} \frac{n_j - d_j}{n_j}$$

2.2.2 Example

The following example shows how KM method is used to estimate the complementary cumulative density function from a left censored data set.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$x_i$</th>
<th>Censor (0=censored, 1=uncensored)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt; 0.002</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>&lt; 0.002</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.00215</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0.00425</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0.0185</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0.0215</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0.0305</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>&lt; 0.08</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0.35775</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0.8</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>11.5</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2.2 is an environmental data set that contains non-detects (left-censored observations). This data set actually contains two different detection limits 0.002 and...
A censor variable is created that has a value of 1 if the observation is uncensored and 0 if the observation is censored. Kaplan Meier method can only be used of right censored data, so these left censored observations ($X_i's$) are transformed into right censored observations by subtracting each observation from a value $L = (M + E)$, where $M$ is the maximum of all these observations and $E > 0$ is a suitable constant.

In the above example the maximum value ($M$) is 11.5 and let the constant $E=1$. Therefore the right censored observations $Y_i's$ are obtained by $Y_i = L - X_i$. Hence the table with right censored data will be as in Table 2.3

<table>
<thead>
<tr>
<th>i</th>
<th>$y_i$</th>
<th>Censor (0=censored, 1=uncensored)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.498</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>12.498</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>12.49785</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>12.49575</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>12.4815</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>12.4785</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>12.4695</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>12.42</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>12.14225</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>11.7</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The data set presented in Table 2.3 is used to show the implementation of the K-M method. The sample size of the data set is $N = 11$. The $N$ observations ($y_i$) will be put into ascending order so that $0 < y_1 < y_2 < \ldots < y_N$. The measurement scale is divided into chosen intervals, $(0, u_1), (u_1, u_2), \ldots$ The chosen intervals are given in Table 2.4. Let $n_i$, $d_i$, $l_i$ and $S(y)$ be defined as,

$n_i=$ the number of observation at the beginning of the interval.
The number of uncensored observations that have occurred within the interval is denoted by \( d_i \).

The number of censored observations that have occurred within the interval is denoted by \( l_i \).

\( \hat{S}(y) \) is the estimated probability of a measurement being greater than \( y = u_i \).

Given the information above, the implementation of the K-M method to estimate \( S(y) \) is shown in Table 2.4. The estimated survival curve is constructed with the computation of \( \hat{S}(y) \).

<table>
<thead>
<tr>
<th>( i )</th>
<th>( u_i )</th>
<th>( n_i )</th>
<th>( d_i )</th>
<th>( l_i )</th>
<th>( p_i )</th>
<th>( \hat{S}(t) = \prod_{i=1}^{k} p_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>11</td>
<td>10</td>
<td>0</td>
<td>10/11</td>
<td>0.91</td>
</tr>
<tr>
<td>2</td>
<td>11.7</td>
<td>10</td>
<td>9</td>
<td>1</td>
<td>9/10</td>
<td>0.819</td>
</tr>
<tr>
<td>3</td>
<td>12.142</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>8/9</td>
<td>0.7271</td>
</tr>
<tr>
<td>5</td>
<td>12.469</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>6/7</td>
<td>0.62331</td>
</tr>
<tr>
<td>6</td>
<td>12.478</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>5/6</td>
<td>0.5194</td>
</tr>
<tr>
<td>7</td>
<td>12.481</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>4/5</td>
<td>0.41552</td>
</tr>
<tr>
<td>8</td>
<td>12.495</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>3/4</td>
<td>0.31164</td>
</tr>
<tr>
<td>9</td>
<td>12.497</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2/3</td>
<td>0.20776</td>
</tr>
<tr>
<td>10</td>
<td>12.498</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

As mentioned in section 2.1 the expected value is computed by calculating the area underneath the survival curve. As the largest observation is censored, the mean can only be estimated up to \( u_i = 12.497 \) [8].

\[
\mu_y = 1(1) + (0.91)(11.7 - 1) + (0.819)(12.142 - 11.7) + (0.7271)(12.469 - 12.142) + 
(0.62331)(12.478 - 12.469) + (0.5194)(12.481 - 12.478) + (0.41552)(12.495 - 12.481) + 
(0.31164)(12.497 - 12.495)
\]

\[= 11.3504\]
The mean of $X$, $\mu_x$, which is of interest, is computed by taking the mean of $Y$, $\mu_y$, and subtracting it from $L = 12.5$ derived in section 2.2.2.

$$\mu_x = L - \mu_y$$

$$= 12.5 - 11.3504$$

$$= 1.1496$$

2.2.3 Computer Implementation

KM Estimates and Survival Plot can be computed in MINITAB by first inserting the observations of $Y$ in one column and their censored values in another column. Select the following menus from the tool bar:

*Stat → Reliability/Survival → Distribution Analysis (Right Censoring) → Nonparametric Distribution Analysis-Right Censoring.*

The KM is the default Estimation Method, but the column containing the censored values must be indicated along with the value that defines the observation as right-censored (0).

Survival probabilities are usually presented as a survival curve (figure). The "curve" is a step function, with sudden changes in the estimated probability corresponding to times at which an event was observed. The following figures shows the result of finding the K-M estimate and the survival function graph using MINITAB14.
Distribution Analysis: Y

Nonparametric Estimates
Characteristics of Variable

<table>
<thead>
<tr>
<th>Standard 95.0% Normal CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (MTTF)</td>
</tr>
<tr>
<td>11.3409</td>
</tr>
</tbody>
</table>

Median = 12.4815

IQR = 0.2556

Q1 = 12.1423

Q3 = 12.4979

Kaplan-Meier Estimates

<table>
<thead>
<tr>
<th>Time</th>
<th>Number</th>
<th>Survival</th>
<th>Standard 95.0% Normal CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0000</td>
<td>11</td>
<td>0.209021</td>
<td>0.065798 0.729204 1.00000</td>
</tr>
<tr>
<td>11.7000</td>
<td>10</td>
<td>0.818182</td>
<td>0.116291 0.590255 1.00000</td>
</tr>
<tr>
<td>12.1423</td>
<td>9</td>
<td>0.727273</td>
<td>0.124282 0.464086 0.99046</td>
</tr>
<tr>
<td>12.4635</td>
<td>7</td>
<td>0.623377</td>
<td>0.150000 0.229302 0.82727</td>
</tr>
<tr>
<td>12.4785</td>
<td>6</td>
<td>0.519481</td>
<td>0.156909 0.211945 0.82370</td>
</tr>
<tr>
<td>12.4815</td>
<td>5</td>
<td>0.415584</td>
<td>0.156181 0.199475 0.72150</td>
</tr>
<tr>
<td>12.4858</td>
<td>4</td>
<td>0.311668</td>
<td>0.147704 0.022192 0.60118</td>
</tr>
<tr>
<td>12.4979</td>
<td>3</td>
<td>0.207792</td>
<td>0.129971 0.000000 0.46253</td>
</tr>
</tbody>
</table>

Figure 2.1: K-M Estimates using Minitab14

Figure 2.2: Survival Plot using Minitab
METHODOLOGY

Chapter 2 explained how KM method can be used estimate the expected value of the survival time $T$. This chapter explains the usage of KM method in multivariate data set and also the method of estimating mean, variance and covariance matrix when the data is multivariate. In Section 3.2, we discuss some results associated with the development of this Algorithm. Section 3.3 briefly explains the program which has been developed and Section 3.4 the algorithm is presented.

3.1 Motivation

When the dataset is univariate with censored observations then the mean can be estimated using the KM method[8]. Most of the environmental data, life testing data are multivariate but when the data is multivariate, the estimates by substitution over DL/2 are recommended in statistics literature[5]. KM method gives a better estimate for mean than substitution over DL/2 method. This led to the idea of estimating the variance, covariance matrix using the KM method. The theoretical results based on which this algorithm has been developed is presented in the following section[12].
3.2 Concepts

**Estimation of mean** Let $x$ be a random variable of the continuous type with probability density function $f(x)$ which is positive provided $0 < x < \infty$ and is equal to 0 otherwise. Then $E(X) = \int_0^\infty [1 - F(x)] \, dx$.

**Proof.**

$$E(X) = \int_0^\infty [1 - F(x)] \, dx$$

Integrating by parts we get,

$$= (x[1 - F(x)])_0^\infty - \int_0^\infty (-xf(x)) \, dx$$

$$= \int_0^\infty xf(x) \, dx$$

$$= E(X) \quad \square$$

Using the above result, we can derive the following results which leads to an estimator for the covariance matrix.

**Estimation of Variance** Let $Y = X^2$ be a random variable of the continuous type with probability density function $f(y)$ which is positive provided $0 < x < \infty$ and is equal to 0 otherwise. Then $E(Y) = \int_0^\infty [1 - F(y)] \, dx$.

Thus $E(X^2)$ is estimated. Therefore variance of $x$ ($var(x)$) is given by

$$var(X) = E(X^2) - [E(X)]^2.$$  

**Covariance** Let $W = X_i \ast X_j$ be a random variable of the continuous type with probability density function $f(w)$ which is positive provided $0 < x_i < \infty$ and $0 < x_j < \infty$ and is equal to 0 otherwise. Then $E(W) = \int_0^\infty [1 - F(w)] \, dx$.

Thus $E(X_i \ast X_j)$ can be estimated. Hence covariance of $X_i \ast X_j$ ($covar(X_iX_j)$) is given by

$$covar(X_iX_j) = E(X_iX_j) - (E(X_i) \ast E(X_j)).$$
3.3 Approach

The program has been developed using VC++ on Microsoft Visual Studio dot net platform[9, 3]. The multivariate data is given as the input, the program implements the KM method algorithm on the data and estimates the mean and covariance matrix.

3.3.1 Explanation of the Program

The program works both for multivariate and univariate data sets. The raw data will be rows and columns of observations with non-detects. The data observations are organized as given below. Each observation is represented as $X_{ij}$

$i=1, 2, \ldots, n$ and $j=1, 2, \ldots, k$.

$n=$ total number of observations

$k=$ total number of variables ($k = 1$ in case of univariate data)

$X_{ij} =$ $i^{th}$ observation of $j^{th}$ variable.

A censor value ($C_{ij}$) is created for each $X_{ij}$ such that $C_{ij}=1$, if the observation is not censored and $C_{ij}=0$, if the observation is censored. $C_{ij}$ values can be generated using any software like MS-Excel[4] or Minitab 14[10].

The program stores all these observations $X_{ij}$ and $C_{ij}$ in two matrices $Data(ij)$ and $Censor(ij)$ respectively as shown.

\[
Data(ij): \begin{pmatrix}
 item1 & \text{Variable}1 & \text{Variable}2 & \ldots & \text{Variable}k \\
 item2 \\
 \vdots \\
 itemn \\
\end{pmatrix} =
\begin{pmatrix}
 X_{11} & X_{12} & \ldots & X_{1k} \\
 X_{21} & X_{22} & \ldots & X_{21} \\
 \vdots & \vdots & \ddots & \vdots \\
 X_{n1} & X_{n2} & \ldots & X_{nk} \\
\end{pmatrix}
\]
Then the KM method explained in Section 2.2, is applied on each pair of columns for example \( X_{i1} \) and \( C_{i1} \). This gives the KM estimate of the mean \( \hat{\mu}_j \). Thus the estimate of the means \( \hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k \) for the variables \( X_{i1}, X_{i2}, \ldots, X_{ik} \) are estimated.

### 3.3.3 Estimation of \( \hat{\mu}_{ij}'s \)

Next, the columns \( X_{ik}'s \) are multiplied by itself and to other columns and the results are found. The corresponding censor values \( C_{ik}'s \) are also multiplied. For example if \( X_1, X_2, X_3 \) are the variables the their product \( X_1X_1, X_1X_2, X_1X_3, X_2X_2, X_2X_3, X_3X_3 \) are calculated. These results are represented as \( X_{i,(jk)}, i=1, 2, \ldots, n \) and \( jk \) represents the product column.

From the above example,

\[ X_{1,(11)} \rightarrow \text{first observation of the product column } X_1X_1 \]

\[ X_{4,(23)} \rightarrow \text{fourth observation of the product column } X_2X_3 \]

Therefore, \( X_{i,(jk)} \rightarrow i^{th} \text{ observation of the product column } X_jX_k \)

The censor columns \( C_1, C_2, C_3 \) are multiplied as done for the \( X \) column and the product \( C_1C_1, C_1C_2, C_1C_3, C_2C_2, C_2C_3, C_3C_3 \) are calculated. These results are represented as \( C_{i,(jk)}, i=1, 2, \ldots, n \) and \( jk \) represents the product column. If \( X_{11} \) and \( X_{13} \) are censored
then $C_{11} = 0$ and $C_{13} = 0$ then the observation $C_{1,(13)}$ will be 0. If $X_{11}$ is censored and $X_{13}$ is uncensored then $C_{11} = 0$ and $C_{13} = 1$ then the observation $C_{1,(13)}$ will be 0. Only if $X_{11}$ and $X_{13}$ are uncensored then $C_{11} = 1$ and $C_{13} = 1$ then the observation $C_{1,(13)}$ will be 1.

After calculating all possible $X_{i,(jk)}$ and $C_{13}$ KM method is applied on each pair of the $X$, $C$ and the KM estimate means $\hat{\mu}_{ij}$, where $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, k$ are found.

Thus the two columns of estimated means $\hat{\mu}_i$'s and $\hat{\mu}_{ij}$'s are found out and also the covariance matrix $\text{Cov}(ij)$ is calculated as follows

$$\text{Cov}(ij) = (\hat{\mu}_{ij} - (\hat{\mu}_i)^*)(\hat{\mu}_j)$$
3.4 Pseudo Code for the Program

Algorithm 3.4.1 Implementation of Kaplan-Meier method (Constants).

1.01 Constants:
1.02 \( \epsilon : \text{float} \) /*Used while Converting left censored observations to right censored */
1.03 \( DL : \text{float array} \) /* Detection limits */

Variables:
1.04 \( row : \text{integer} \) /* Number of observations corresponding to one variable */
1.05 \( column : \text{integer} \) /* Number of variables in the multivariate dataset */
1.06 \( input - txt : \text{string} \) /* Name of the input file with data and censor values */
1.07 \( nresult : \text{integer} \) /* Number of product columns */
1.08 \( data : \text{float array} [0..row - 1][0..column - 1] \)
   /* all the input data observations \( X_i^t \)'s */
1.09 \( censor : \text{integer array} [0..row - 1][0..column - 1] \)
   /* all the censor observations \( C_i^t \)'s */
1.10 \( product : \text{float array} [0..row - 1][0..nresult - 1] \)
   /* all the data observations in the product columns */
1.11 \( m1 : \text{float array} [0..column - 1] \)
   /* Estimated values of mean vector */
1.12 \( m2 : \text{float array} [0..nresult - 1][0..column - 1] \)
   /* Estimated values of the mean vector for the product terms */
1.13 \( cov : \text{float array} [0..column - 1][0..column - 1] \)
   /* Estimated values of the covariance matrix */
1.14 \( data : \text{float} \) /* Used while Converting left censored */
1.15 \( censor : \text{float} \) /* Used while Converting left censored */
1.16 \( data : \text{float} \) /* Used while Converting left censored */
Algorithm 3.4.2 Implementation of Kaplan-Meier method (Procedures).

Functions:

2.01 GETDATA() :
   Gets input input_{txt}, row, column, DL, ε.
   Using input_{txt} gets the input observations and censor values.
   Stores in data(i, j) and censor(i, j)
   \( i = 0, \ldots, row - 1 \) and \( j = 0, \ldots, column - 1 \)

2.02 SORTDATA() :
   Converts the left censored observations to right censored
   Arranges the observations and corresponding censor values
   in descending order.

2.03 PRINTDATA() :
   Prints the results.

2.04 PROCESS() :
   Computes the product columns of the data and censor columns.
   Stores data and censor in the array product(i, j) and censor(i, j)
   respectively \( i = 0, \ldots, row - 1 \) and \( j = 0, \ldots, nresult - 1 \)

2.05 KPM() :
   Computes the Kaplan-Meier estimates of the column of data
   given as input.
   Gives the output \( m_1(j) \) and \( m_2(i, j) \)
   \( i = 0, \ldots, row - 1 \) and \( j = 0, \ldots, nresult - 1 \)

2.06 COVARIANCE() :
   Takes the KM estimates of data column and the product column
   as input and computes the covariance matrix
   \[ cov(i, j) = m_2(i, j) - m_1(i) \ast m_1(j) \ (i, j = 0, \ldots, nresult - 1) \]
Algorithm 3.4.3 Implementation of Kaplan-Meier method (Program).

Program:

3.01 GETDATA()
3.02 SORTDATA()
3.03 PROCESS()
3.04 KPM()
3.05 COVARIANCE()
3.06 PRINTDATA()
CHAPTER 4

RESULTS

In this chapter the performance of the KM method using the program developed in Chapter 3 is investigated by testing the program with some simulated multivariate datasets as input and the results are compared with the results obtained by some of the existing methods.

4.1 Data Description

In this section, three simulated examples using three different covariance matrices are given. Using the three covariance matrices and the corresponding mean vectors, multivariate random samples of data are generated. For each matrix three random samples of size 20 observations, 50 observations and 100 observations are generated using the programming language R [7]. These generated samples are presented in Appendix.

Then for each sample a detection level (DL) is determined for each such that 20 percent of the observations fall below the DL. These observations are then replaced by respective DL's to obtain the simulated censored data set. These samples are presented in Appendix.
4.2 Example

Original Covariance Matrix $\Sigma_1 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{pmatrix}$

Mean Vector $\mu_1 = \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix}$

Random Sample A random sample of 25 observations is generated using $\Sigma_1$ and $\mu_1$.

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.98250</td>
<td>-2.74732</td>
<td>0.23806</td>
</tr>
<tr>
<td>2.47629</td>
<td>-3.43404</td>
<td>1.36419</td>
</tr>
<tr>
<td>3.46703</td>
<td>1.12805</td>
<td>3.28074</td>
</tr>
<tr>
<td>2.33231</td>
<td>-2.91293</td>
<td>0.19329</td>
</tr>
<tr>
<td>3.97317</td>
<td>-2.38859</td>
<td>2.95039</td>
</tr>
<tr>
<td>1.6389</td>
<td>-3.03085</td>
<td>0.48400</td>
</tr>
<tr>
<td>2.40469</td>
<td>-3.16388</td>
<td>0.53268</td>
</tr>
<tr>
<td>1.96391</td>
<td>-3.63436</td>
<td>0.10889</td>
</tr>
<tr>
<td>2.39136</td>
<td>-0.79621</td>
<td>2.20824</td>
</tr>
<tr>
<td>1.36509</td>
<td>-2.68013</td>
<td>0.74314</td>
</tr>
<tr>
<td>3.62243</td>
<td>-2.25037</td>
<td>2.28205</td>
</tr>
<tr>
<td>1.17900</td>
<td>-2.31988</td>
<td>1.73928</td>
</tr>
<tr>
<td>2.06820</td>
<td>-3.36212</td>
<td>0.47066</td>
</tr>
<tr>
<td>1.03153</td>
<td>-5.56997</td>
<td>-1.1857</td>
</tr>
<tr>
<td>2.60653</td>
<td>-1.53910</td>
<td>3.04318</td>
</tr>
<tr>
<td>4.02005</td>
<td>-1.88061</td>
<td>3.16370</td>
</tr>
<tr>
<td>2.56744</td>
<td>-2.38567</td>
<td>2.16686</td>
</tr>
<tr>
<td>1.43993</td>
<td>-0.64400</td>
<td>1.74997</td>
</tr>
<tr>
<td>2.22368</td>
<td>-3.05128</td>
<td>0.85685</td>
</tr>
<tr>
<td>2.30823</td>
<td>-3.14217</td>
<td>1.45188</td>
</tr>
<tr>
<td>3.48377</td>
<td>-2.55419</td>
<td>1.61262</td>
</tr>
<tr>
<td>2.55082</td>
<td>-4.00030</td>
<td>0.93757</td>
</tr>
<tr>
<td>1.38838</td>
<td>-5.39764</td>
<td>-1.1978</td>
</tr>
<tr>
<td>0.39259</td>
<td>-4.03936</td>
<td>0.19275</td>
</tr>
<tr>
<td>3.91901</td>
<td>-1.19681</td>
<td>4.08749</td>
</tr>
</tbody>
</table>
Sample With Non-Detects In the complete sample from table 4.1 in each column $X_1, X_2, X_3$ the data below the first quartile are made as non-detects, so the first quartile becomes the detection limit (DL). Therefore in each column there are 5 non-detects. The observations below DL are replaced by the '< DL'. In the following table the column $X_1, X_2, X_3$ represents the sample with non-detects. The columns $C_1, C_2, C_3$ represent the censor values of $X_1, X_2, X_3$ respectively. If the observation $X_{ij}$ is an non-detect the corresponding censor value $C_j$ is 0 and $C_j$ is 1 otherwise.

Table 4.2: Sample With Non-detects

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.9825</td>
<td>1</td>
<td>-2.7473</td>
<td>1</td>
<td>&lt; 0.3500</td>
<td>0</td>
</tr>
<tr>
<td>2.4763</td>
<td>1</td>
<td>-3.4000</td>
<td>0</td>
<td>1.3642</td>
<td>1</td>
</tr>
<tr>
<td>3.4670</td>
<td>1</td>
<td>1.1281</td>
<td>1</td>
<td>3.2807</td>
<td>1</td>
</tr>
<tr>
<td>2.3323</td>
<td>1</td>
<td>-2.9129</td>
<td>1</td>
<td>&lt; 0.3500</td>
<td>0</td>
</tr>
<tr>
<td>3.9732</td>
<td>1</td>
<td>-2.3886</td>
<td>1</td>
<td>2.9504</td>
<td>1</td>
</tr>
<tr>
<td>1.6389</td>
<td>1</td>
<td>-4.0309</td>
<td>1</td>
<td>0.4840</td>
<td>1</td>
</tr>
<tr>
<td>2.4047</td>
<td>1</td>
<td>-3.1639</td>
<td>1</td>
<td>0.5327</td>
<td>1</td>
</tr>
<tr>
<td>1.9639</td>
<td>1</td>
<td>&lt; -3.4000</td>
<td>0</td>
<td>&lt; 0.3500</td>
<td>0</td>
</tr>
<tr>
<td>2.3914</td>
<td>1</td>
<td>-0.7962</td>
<td>1</td>
<td>2.2082</td>
<td>1</td>
</tr>
<tr>
<td>&lt; 1.5390</td>
<td>0</td>
<td>-2.6801</td>
<td>1</td>
<td>0.7431</td>
<td>1</td>
</tr>
<tr>
<td>3.6224</td>
<td>1</td>
<td>-2.2504</td>
<td>1</td>
<td>2.2821</td>
<td>1</td>
</tr>
<tr>
<td>&lt; 1.5390</td>
<td>0</td>
<td>-2.3199</td>
<td>1</td>
<td>1.7393</td>
<td>1</td>
</tr>
<tr>
<td>2.0682</td>
<td>1</td>
<td>-3.3621</td>
<td>1</td>
<td>0.4707</td>
<td>1</td>
</tr>
<tr>
<td>&lt; 1.5390</td>
<td>0</td>
<td>&lt; -3.4000</td>
<td>0</td>
<td>&lt; 0.3500</td>
<td>0</td>
</tr>
<tr>
<td>2.6065</td>
<td>1</td>
<td>-1.5391</td>
<td>1</td>
<td>3.0432</td>
<td>1</td>
</tr>
<tr>
<td>4.0201</td>
<td>1</td>
<td>-1.8806</td>
<td>1</td>
<td>3.1637</td>
<td>1</td>
</tr>
<tr>
<td>2.5674</td>
<td>1</td>
<td>-2.3857</td>
<td>1</td>
<td>2.1669</td>
<td>1</td>
</tr>
<tr>
<td>&lt; 1.5390</td>
<td>0</td>
<td>-0.6440</td>
<td>1</td>
<td>1.7500</td>
<td>1</td>
</tr>
<tr>
<td>2.2237</td>
<td>1</td>
<td>-3.0513</td>
<td>1</td>
<td>0.8569</td>
<td>1</td>
</tr>
<tr>
<td>2.3082</td>
<td>1</td>
<td>-3.1422</td>
<td>1</td>
<td>1.4519</td>
<td>1</td>
</tr>
<tr>
<td>3.4838</td>
<td>1</td>
<td>-2.5542</td>
<td>1</td>
<td>1.6126</td>
<td>1</td>
</tr>
<tr>
<td>2.5508</td>
<td>1</td>
<td>&lt; -3.4000</td>
<td>0</td>
<td>0.9376</td>
<td>1</td>
</tr>
<tr>
<td>&lt; 1.5390</td>
<td>0</td>
<td>&lt; -3.4000</td>
<td>0</td>
<td>&lt; 0.3500</td>
<td>0</td>
</tr>
<tr>
<td>&lt; 1.5390</td>
<td>0</td>
<td>&lt; -3.4000</td>
<td>0</td>
<td>&lt; 0.3500</td>
<td>0</td>
</tr>
<tr>
<td>3.9190</td>
<td>1</td>
<td>-1.1968</td>
<td>1</td>
<td>4.0875</td>
<td>1</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
4.3 Comparison of the Results

From the data from table 4.2 the mean and covariance matrix is estimated using the three substitution methods (substitution over DL, substitution over $DL/2$, substitution by zero) and also by the KM-method using the program developed which was explained in Chapter 3.

**Substitution over DL method:** In table 4.2 the columns $X_1, X_2, X_3$ represent data observations with non-detects and the non-detects are actually replaced by the corresponding DL's. The estimates of mean and covariance for this data are obtained using the software Minitab 14 [10]. The estimates are as follows

$$\Sigma_{DL,1} = \begin{pmatrix} 0.699 & 0.387 & 0.738 \\ 0.387 & 1.245 & 0.942 \\ 0.738 & 0.942 & 1.295 \end{pmatrix}$$

$$\mu_{DL,1} = \begin{pmatrix} 2.499 \\ -2.453 \\ 1.489 \end{pmatrix}$$

**Substitution over DL/2 method:** In table 4.2 the columns $X_1, X_2, X_3$ represent data observations with non-detects and now the non-detects are replaced by the corresponding DL/2 values and the mean and covariance for these data observations are estimated using Minitab 14. The estimates are as follows

$$\Sigma_{DL/2,1} = \begin{pmatrix} 1.163 & 0.120 & 0.891 \\ 0.120 & 0.989 & 0.573 \\ 0.891 & 0.573 & 1.400 \end{pmatrix}$$
\[
\mu_{DL/2,1} = \begin{pmatrix}
2.265 & -2.045 & 1.447
\end{pmatrix}
\]

**Substitution over Zero method:** In table 4.2 the columns \(\hat{X}_1, \hat{X}_2, \hat{X}_3\) represent data observations with non-detects and now the non-detects are replaced by 0 and the mean and covariance for these data observations are estimated using Minitab 14. The estimates are as follows
\[
\Sigma_{\text{zero},1} = \begin{pmatrix}
1.850 & -0.317 & 1.628 \\
-0.317 & 1.831 & 0.140 \\
1.628 & 0.140 & 1.518
\end{pmatrix}
\]
\[
\mu_{\text{zero},1} = \begin{pmatrix}
2.080 \\
-1.637 \\
1.405
\end{pmatrix}
\]

**Kaplan Meier Method:** The data from table 4.2 is used as input for the program from chapter 3 This program outputs the Kaplan Meier estimates of mean and covariance.
\[
\Sigma_{KM,1} = \begin{pmatrix}
0.630 & 0.806 & 0.834 \\
0.806 & 1.178 & 1.131 \\
0.834 & 1.131 & 1.18036
\end{pmatrix}
\]
\[
\mu_{KM,1} = \begin{pmatrix}
2.473 \\
-2.444 \\
1.518
\end{pmatrix}
\]

**Estimation of the complete sample:** The estimated covariance and mean for the data from the table 4.1, computed using Minitab 14, is given below.
\[
\Sigma_{CS,1} = \begin{pmatrix}
0.944 & 0.897 & 0.834 \\
0.897 & 2.036 & 1.131 \\
0.972 & 1.157 & 1.831
\end{pmatrix}
\]
\[
\mu_{CS,1} = \begin{pmatrix}
2.352 \\
-2.680 \\
1.339
\end{pmatrix}
\]
Figure 4.1: Comparison Graph for Estimates of $\mu_1$ for 25 Observations

Figure 4.2: Comparison Graph for Estimates of $\sigma_1$ for 25 Observations
From the above results and graphs we can clearly see that the KM method gives a better estimate. Further, the results obtained using all the above mentioned methods are compared with the KM estimate on different data sets, different number of observations and with 20 percent non-detects.

Table 4.3: Estimates of $\Sigma_i$ and $\mu_i$ for 50 observations

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean vector</th>
<th>Covariance matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>$\mu_i = \begin{pmatrix} 2 \ -3 \ 1 \end{pmatrix}$</td>
<td>$\Sigma_i = \begin{pmatrix} 1 &amp; 1 &amp; 1 \ 1 &amp; 3 &amp; 2 \ 1 &amp; 2 &amp; 2 \end{pmatrix}$</td>
</tr>
<tr>
<td>Complete Sample</td>
<td>$\mu_{\text{Comp}} = \begin{pmatrix} 2.080 \ -2.953 \ 1.065 \end{pmatrix}$</td>
<td>$\Sigma_{\text{Comp}} = \begin{pmatrix} 0.929331 &amp; 0.849825 &amp; 0.971362 \ 0.849825 &amp; 2.659954 &amp; 1.897147 \ 0.971362 &amp; 1.897147 &amp; 1.901819 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over DL</td>
<td>$\mu_{\text{DL}} = \begin{pmatrix} 2.249 \ -2.611 \ 1.369 \end{pmatrix}$</td>
<td>$\Sigma_{\text{DL}} = \begin{pmatrix} 0.488350 &amp; 0.331177 &amp; 0.388667 \ 0.331177 &amp; 1.199579 &amp; 0.749972 \ 0.388667 &amp; 0.749972 &amp; 0.767273 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over DL/2</td>
<td>$\mu_{\text{DL}/2} = \begin{pmatrix} 2.061 \ -2.149 \ 1.317 \end{pmatrix}$</td>
<td>$\Sigma_{\text{DL}/2} = \begin{pmatrix} 20075.897 &amp; 1839.321 &amp; 451.002 \ 1839.321 &amp; 248.954 &amp; 50.741 \ 451.002 &amp; 50.741 &amp; 56.181 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over Zero</td>
<td>$\mu_{\text{Zero}} = \begin{pmatrix} 1.871 \ -1.686 \ 1.264 \end{pmatrix}$</td>
<td>$\Sigma_{\text{Zero}} = \begin{pmatrix} 62322.129 &amp; 4225.437 &amp; 1862.226 \ 4225.437 &amp; 708.698 &amp; 156.014 \ 1862.226 &amp; 156.014 &amp; 144.977 \end{pmatrix}$</td>
</tr>
<tr>
<td>Kaplan Meier</td>
<td>$\mu_{\text{KM}} = \begin{pmatrix} 2.25257 \ -2.59297 \ 1.37283 \end{pmatrix}$</td>
<td>$\Sigma_{\text{KM}} = \begin{pmatrix} 0.474801 &amp; 0.706854 &amp; 0.582351 \ 0.706854 &amp; 1.13138 &amp; 0.893163 \ 0.582351 &amp; 0.893163 &amp; 0.745627 \end{pmatrix}$</td>
</tr>
</tbody>
</table>
Figure 4.3: Estimates of $\mu_i$ for 50 Observations

Figure 4.4: Estimates of $\Sigma_1$ for 50 Observations
Table 4.4: Estimates of $\Sigma_i$ and $\mu_i$ for 100 observations

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean vector $\mu_i$</th>
<th>Covariance matrix $\Sigma_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>$\begin{pmatrix} 2 \ -3 \ 1 \end{pmatrix}$</td>
<td>$\begin{pmatrix} 1 &amp; 1 &amp; 1 \ 1 &amp; 3 &amp; 2 \ 1 &amp; 2 &amp; 2 \end{pmatrix}$</td>
</tr>
<tr>
<td>Complete Sample</td>
<td>$\begin{pmatrix} 2.080 \ -2.953 \ 1.065 \end{pmatrix}$</td>
<td>$\begin{pmatrix} 0.944493 &amp; 0.897941 &amp; 0.972800 \ 0.897941 &amp; 3.050185 &amp; 2.094594 \ 0.972800 &amp; 2.094594 &amp; 2.051408 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution</td>
<td>$\begin{pmatrix} 1.8867 \ -2.783 \ 0.969 \end{pmatrix}$</td>
<td>$\begin{pmatrix} 0.681985 &amp; 0.589936 &amp; 0.734905 \ 0.589936 &amp; 1.776597 &amp; 1.300324 \ 0.734905 &amp; 1.300324 &amp; 1.508191 \end{pmatrix}$</td>
</tr>
<tr>
<td>Over DL</td>
<td>$\begin{pmatrix} 1.7603 \ -2.317 \ 1.015 \end{pmatrix}$</td>
<td>$\begin{pmatrix} 0.953956 &amp; 0.429214 &amp; 0.797313 \ 0.429214 &amp; 1.304985 &amp; 0.760127 \ 0.797313 &amp; 0.760127 &amp; 1.391149 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution</td>
<td>$\begin{pmatrix} 1.634 \ -1.787 \ 1.060 \end{pmatrix}$</td>
<td>$\begin{pmatrix} 1.32276 &amp; 0.15780 &amp; 0.84897 \ 0.15780 &amp; 2.36720 &amp; 0.24727 \ 0.84897 &amp; 0.24727 &amp; 1.28679 \end{pmatrix}$</td>
</tr>
<tr>
<td>Over Zero</td>
<td>$\begin{pmatrix} 1.8901 \ -2.8474 \ 0.9725 \end{pmatrix}$</td>
<td>$\begin{pmatrix} 0.66916 &amp; 1.13078 &amp; 0.990441 \ 1.13078 &amp; 1.92719 &amp; 1.67686 \ 0.990441 &amp; 1.67686 &amp; 1.4836 \end{pmatrix}$</td>
</tr>
</tbody>
</table>
Comparison of Estimates of Mean

Figure 4.5: Estimates of $\mu_i$ for 100 Observations

Comparison of Estimates of the Covariance Matrix

Figure 4.6: Estimates of $\Sigma_i$ for 100 Observations
Table 4.5: Estimates of $\Sigma$ and $\mu$ for 25 observations

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean vector</th>
<th>Covariance matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>$\mu = \begin{pmatrix} 526.59 \ 54.69 \ 25.13 \end{pmatrix}$</td>
<td>$\Sigma = \begin{pmatrix} 5691.34 &amp; 600.51 &amp; 217.25 \ 600.51 &amp; 126.05 &amp; 23.37 \ 217.25 &amp; 23.37 &amp; 23.11 \end{pmatrix}$</td>
</tr>
<tr>
<td>Complete Sample</td>
<td>$\mu_{Com} = \begin{pmatrix} 534.5 \ 52.69 \ 24.951 \end{pmatrix}$</td>
<td>$\Sigma_{Com} = \begin{pmatrix} 3901.3111 &amp; 470.2541 &amp; 154.6328 \ 470.2541 &amp; 85.2228 &amp; 20.6265 \ 154.6328 &amp; 20.6265 &amp; 24.6581 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over DL</td>
<td>$\mu_{DL} = \begin{pmatrix} 544.07 \ 54.09 \ 25.801 \end{pmatrix}$</td>
<td>$\Sigma_{DL} = \begin{pmatrix} 2494.805 &amp; 271.123 &amp; 65.827 \ 271.123 &amp; 50.520 &amp; 7.840 \ 65.827 &amp; 7.840 &amp; 12.288 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over DL/2</td>
<td>$\mu_{DL/2} = \begin{pmatrix} 484.4 \ 48.69 \ 23.16 \end{pmatrix}$</td>
<td>$\Sigma_{DL/2} = \begin{pmatrix} 20075.897 &amp; 1839.321 &amp; 451.002 \ 1839.321 &amp; 248.954 &amp; 50.741 \ 451.002 &amp; 50.741 &amp; 56.181 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over Zero</td>
<td>$\mu_{Zero} = \begin{pmatrix} 428.8 \ 43.29 \ 20.52 \end{pmatrix}$</td>
<td>$\Sigma_{Zero} = \begin{pmatrix} 61122.840 &amp; 5066.255 &amp; 1191.531 \ 5066.255 &amp; 639.761 &amp; 125.816 \ 1191.531 &amp; 125.816 &amp; 146.054 \end{pmatrix}$</td>
</tr>
<tr>
<td>Kaplan Meier</td>
<td>$\mu_{KM} = \begin{pmatrix} 544.49 \ 54.98 \ 25.82 \end{pmatrix}$</td>
<td>$\Sigma_{KM} = \begin{pmatrix} 2355.83 &amp; 282.453 &amp; 155.086 \ 282.453 &amp; 34.7154 &amp; 19.255 \ 155.086 &amp; 19.255 &amp; 11.6392 \end{pmatrix}$</td>
</tr>
</tbody>
</table>
Figure 4.7: Estimates of $\mu_i$ for 25 Observations

Figure 4.8: Estimates of $\sigma_{11}$ of $\Sigma_2$ for 25 Observations
Figure 4.9: Estimates of $\sigma_{21}$ and $\sigma_{31}$ of $\Sigma_2$ for 25 Observations

Figure 4.10: Remaining estimates of $\Sigma_2$ for 25 Observations
Table 4.6: Estimates of $\Sigma_2$ and $\mu_2$ for 50 observations

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean vector</th>
<th>Covariance matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>$\mu_2 = \begin{pmatrix} 526.59 \ 54.69 \ 25.13 \end{pmatrix}$</td>
<td>$\Sigma_2 = \begin{pmatrix} 5691.34 &amp; 600.51 &amp; 217.25 \ 600.51 &amp; 126.05 &amp; 23.37 \ 217.25 &amp; 23.37 &amp; 23.11 \end{pmatrix}$</td>
</tr>
<tr>
<td>Complete Sample</td>
<td>$\mu_{Com} = \begin{pmatrix} 522.57 \ 54.11 \ 24.85 \end{pmatrix}$</td>
<td>$\Sigma_{Com} = \begin{pmatrix} 4255.9289 &amp; 559.9779 &amp; 193.9064 \ 559.9779 &amp; 137.2202 &amp; 29.5939 \ 193.9064 &amp; 29.5939 &amp; 26.8204 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over DL</td>
<td>$\mu_{DL} = \begin{pmatrix} 476.30 \ 50.53 \ 22.83 \end{pmatrix}$</td>
<td>$\Sigma_{DL} = \begin{pmatrix} 2321.3877 &amp; 280.748 &amp; 98.751 \ 280.748 &amp; 69.778 &amp; 15.823 \ 98.751 &amp; 15.823 &amp; 16.572 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over DL/2</td>
<td>$\mu_{DL/2} = \begin{pmatrix} 476.3 \ 50.53 \ 22.83 \end{pmatrix}$</td>
<td>$\Sigma_{DL/2} = \begin{pmatrix} 18892.012 &amp; 1382.644 &amp; 574.321 \ 1382.644 &amp; 273.014 &amp; 66.568 \ 574.321 &amp; 66.568 &amp; 63.645 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over Zero</td>
<td>$\mu_{Zero} = \begin{pmatrix} 417.4 \ 44.7 \ 20.10 \end{pmatrix}$</td>
<td>$\Sigma_{Zero} = \begin{pmatrix} 57825.792 &amp; 3484.068 &amp; 1457.464 \ 3484.068 &amp; 695.152 &amp; 147.244 \ 1457.464 &amp; 147.244 &amp; 154.008 \end{pmatrix}$</td>
</tr>
<tr>
<td>Kaplan Meier</td>
<td>$\mu_{KM} = \begin{pmatrix} 535.152 \ 56.442 \ 25.829 \end{pmatrix}$</td>
<td>$\Sigma_{KM} = \begin{pmatrix} 2270.23 &amp; 378.306 &amp; 188.162 \ 378.306 &amp; 67.6276 &amp; 30.0822 \ 188.162 &amp; 30.0822 &amp; 14.0202 \end{pmatrix}$</td>
</tr>
</tbody>
</table>
Figure 4.11: Estimates of $\mu_2$ for 50 Observations

Figure 4.12: Estimates of $\sigma_{11}$ of $\Sigma_2$ for 50 Observations

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 4.13: Estimates of $\sigma_{21}$ and $\sigma_{31}$ of $\Sigma_2$ for 50 Observations

Figure 4.14: Remaining Estimates of $\Sigma_2$ for 50 Observations

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Table 4.7: Estimates of $\Sigma_2$ and $\mu_2$ for 100 observations

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean vector</th>
<th>Covariance matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>$\mu_2 = \begin{pmatrix} 526.59 \ 54.69 \ 25.13 \end{pmatrix}$</td>
<td>$\Sigma_2 = \begin{pmatrix} 5691.34 &amp; 600.51 &amp; 217.25 \ 600.51 &amp; 126.05 &amp; 23.37 \ 217.25 &amp; 23.37 &amp; 23.11 \end{pmatrix}$</td>
</tr>
<tr>
<td>Complete Sample</td>
<td>$\tilde{\mu}_{Com} = \begin{pmatrix} 531.19 \ 54.29 \ 25.469 \end{pmatrix}$</td>
<td>$\tilde{\Sigma}_{Com} = \begin{pmatrix} 5517.1196 &amp; 569.0342 &amp; 239.1887 \ 569.0342 &amp; 122.1796 &amp; 33.0914 \ 239.1887 &amp; 33.0914 &amp; 24.8570 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over DL</td>
<td>$\tilde{\mu}_{DL} = \begin{pmatrix} 539.76 \ 55.88 \ 26.172 \end{pmatrix}$</td>
<td>$\Sigma_{DL} = \begin{pmatrix} 3856.76 &amp; 355.0633 &amp; 156.449 \ 355.0633 &amp; 78.4110 &amp; 19.799 \ 156.449 &amp; 19.799 &amp; 15.7491 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over DL/2</td>
<td>$\tilde{\mu}_{DL/2} = \begin{pmatrix} 481.40 \ 50.13 \ 23.64 \end{pmatrix}$</td>
<td>$\tilde{\Sigma}_{DL/2} = \begin{pmatrix} 22763.04 &amp; 1701.09 &amp; 743.44 \ 1701.09 &amp; 293.36 &amp; 69.37 \ 743.44 &amp; 69.37 &amp; 58.71 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over Zero</td>
<td>$\tilde{\mu}_{Zero} = \begin{pmatrix} 423.00 \ 44.38 \ 21.11 \end{pmatrix}$</td>
<td>$\Sigma_{Zero} = \begin{pmatrix} 62322.129 &amp; 4225.437 &amp; 1862.226 \ 4225.437 &amp; 708.698 &amp; 156.014 \ 1862.226 &amp; 156.014 &amp; 144.977 \end{pmatrix}$</td>
</tr>
<tr>
<td>Kaplan Meier</td>
<td>$\tilde{\mu}_{KM} = \begin{pmatrix} 540.26 \ 54.89 \ 26.2 \end{pmatrix}$</td>
<td>$\tilde{\Sigma}_{KM} = \begin{pmatrix} 3747.74 &amp; 535.248 &amp; 235.609 \ 535.248 &amp; 77.3968 &amp; 33.8503 \ 235.609 &amp; 33.8503 &amp; 15.0432 \end{pmatrix}$</td>
</tr>
</tbody>
</table>
Figure 4.15: Estimates of $\mu_2$ for 100 Observations

Figure 4.16: Estimates of $\sigma_{11}$ of $\Sigma_2$ for 100 Observations
Figure 4.17: Estimates of $\sigma_{21}$ and $\sigma_{31}$ of $\Sigma_2$ for 100 Observations

Figure 4.18: Estimates of $\Sigma_2$ for 100 Observations
Table 4.8: Estimates of $\Sigma_3$ and $\mu_3$ for 25 observations

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean vector</th>
<th>Covariance matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>$\mu_3 = \begin{pmatrix} 368.21 \ 404.63 \ 479.26 \ 502.89 \end{pmatrix}$</td>
<td>$\Sigma_3 = \begin{pmatrix} 2819.29 &amp; 3568.42 &amp; 2943.49 &amp; 2295.35 \ 3568.42 &amp; 7963.14 &amp; 5303.98 &amp; 4065.44 \ 2943.39 &amp; 5303.98 &amp; 6851.32 &amp; 4499.63 \ 2295.35 &amp; 4065.44 &amp; 4499.63 &amp; 4878.99 \end{pmatrix}$</td>
</tr>
<tr>
<td>Complete Sample</td>
<td>$\mu_{com} = \begin{pmatrix} 367.0 \ 396.9 \ 474.1 \ 495.6 \end{pmatrix}$</td>
<td>$\Sigma_{com} = \begin{pmatrix} 3440.21 &amp; 5168.99 &amp; 4689.59 &amp; 3382.68 \ 5168.99 &amp; 10622.35 &amp; 8131.25 &amp; 5887.86 \ 4689.59 &amp; 8131.25 &amp; 10085.41 &amp; 6086.04 \ 3382.68 &amp; 5887.86 &amp; 6086.04 &amp; 5913.69 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over DL</td>
<td>$\mu_{DL} = \begin{pmatrix} 373.77 \ 411.7 \ 484.3 \ 503.6 \end{pmatrix}$</td>
<td>$\Sigma_{DL} = \begin{pmatrix} 2067.34 &amp; 2067.34 &amp; 2683.49 &amp; 2003.24 \ 2067.34 &amp; 10622.35 &amp; 8131.25 &amp; 5887.86 \ 2683.49 &amp; 5249.40 &amp; 7389.00 &amp; 4376.76 \ 2003.24 &amp; 4105.58 &amp; 4376.76 &amp; 4434.10 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over DL/2</td>
<td>$\mu_{DL/2} = \begin{pmatrix} 327.2 \ 374.0 \ 436.9 \ 451.7 \end{pmatrix}$</td>
<td>$\Sigma_{DL/2} = \begin{pmatrix} 11849.2 &amp; 14390.4 &amp; 14390.4 &amp; 10738.4 \ 11884.3 &amp; 19047.5 &amp; 17538.9 &amp; 13989.0 \ 14390.4 &amp; 17538.9 &amp; 23622.5 &amp; 17780.3 \ 10738.4 &amp; 13989.0 &amp; 17780.3 &amp; 20985.9 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over Zero</td>
<td>$\mu_{zero} = \begin{pmatrix} 280.5 \ 336.3 \ 389.5 \ 399.7 \end{pmatrix}$</td>
<td>$\Sigma_{zero} = \begin{pmatrix} 33274.4 &amp; 27895.0 &amp; 37935.5 &amp; 29446.6 \ 27895.0 &amp; 40766.4 &amp; 39027.4 &amp; 31123.7 \ 37935.5 &amp; 39027.4 &amp; 54678.5 &amp; 43869.0 \ 29446.6 &amp; 31123.7 &amp; 43869.0 &amp; 55349.1 \end{pmatrix}$</td>
</tr>
<tr>
<td>Kaplan Meier</td>
<td>$\mu_{KM} = \begin{pmatrix} 374.427 \ 419.674 \ 484.599 \ 503.672 \end{pmatrix}$</td>
<td>$\Sigma_{KM} = \begin{pmatrix} 1932.50 &amp; 3426.16 &amp; 6302.82 &amp; 3124.23 \ 3426.16 &amp; 5069.85 &amp; 5937.69 &amp; 4600.87 \ 6302.82 &amp; 5937.69 &amp; 7047.47 &amp; 5415.18 \ 3124.23 &amp; 4600.87 &amp; 5415.18 &amp; 4251.24 \end{pmatrix}$</td>
</tr>
</tbody>
</table>
Figure 4.19: Estimates of $\mu_3$ for 25 Observations

Figure 4.20: Estimates of $\Sigma_3$ for 100 Observations
<table>
<thead>
<tr>
<th>Method</th>
<th>Mean vector</th>
<th>Covariance matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>$\mu_1 = \begin{pmatrix} 368.21 \ 404.63 \ 479.26 \ 502.89 \end{pmatrix}$</td>
<td>$\Sigma_1 = \begin{pmatrix} 2819.29 &amp; 3568.42 &amp; 2943.49 &amp; 2295.35 \ 3568.42 &amp; 7963.14 &amp; 5303.98 &amp; 4065.44 \ 2943.39 &amp; 5303.98 &amp; 6851.32 &amp; 4499.63 \ 2295.35 &amp; 4065.44 &amp; 4499.63 &amp; 4878.99 \end{pmatrix}$</td>
</tr>
<tr>
<td>Complete Sample</td>
<td>$\mu_{\text{Com}} = \begin{pmatrix} 376.05 \ 427.40 \ 488.70 \ 520.71 \end{pmatrix}$</td>
<td>$\Sigma_{\text{Com}} = \begin{pmatrix} 2555.56 &amp; 3564.37 &amp; 3631.53 &amp; 2553.89 \ 3564.37 &amp; 7456.3 &amp; 6216.94 &amp; 4388.75 \ 3631.53 &amp; 6216.94 &amp; 7870.79 &amp; 5401.16 \ 2553.89 &amp; 4388.75 &amp; 5401.16 &amp; 4921.72 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over DL</td>
<td>$\mu_{DL} = \begin{pmatrix} 383.69 \ 442.00 \ 506.55 \ 528.64 \end{pmatrix}$</td>
<td>$\Sigma_{DL} = \begin{pmatrix} 1395.37 &amp; 1964.59 &amp; 1820.66 &amp; 1514.79 \ 1964.59 &amp; 4403.17 &amp; 3182.75 &amp; 2523.35 \ 1820.66 &amp; 3182.75 &amp; 3953.21 &amp; 3075.00 \ 1514.79 &amp; 2523.35 &amp; 3075.00 &amp; 3482.30 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over DL/2</td>
<td>$\mu_{DL/2} = \begin{pmatrix} 343.30 \ 397.50 \ 453.80 \ 473.60 \end{pmatrix}$</td>
<td>$\Sigma_{DL/2} = \begin{pmatrix} 10533.5 &amp; 10569.1 &amp; 10260.7 &amp; 7800.8 \ 10569.1 &amp; 17258.3 &amp; 14374.3 &amp; 12110.1 \ 10260.7 &amp; 14374.3 &amp; 20132.9 &amp; 16471.4 \ 7800.8 &amp; 12110.1 &amp; 16471.4 &amp; 21113.3 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over Zero</td>
<td>$\mu_{\text{Zero}} = \begin{pmatrix} 302.80 \ 353.00 \ 401.00 \ 418.50 \end{pmatrix}$</td>
<td>$\Sigma_{\text{Zero}} = \begin{pmatrix} 30240.6 &amp; 26981.3 &amp; 26447.6 &amp; 19011.4 \ 26981.3 &amp; 42922.5 &amp; 35760.0 &amp; 30593.6 \ 26447.6 &amp; 35760.0 &amp; 54329.2 &amp; 42480.1 \ 19011.4 &amp; 30593.6 &amp; 42480.1 &amp; 58350.5 \end{pmatrix}$</td>
</tr>
<tr>
<td>Kaplan Meier</td>
<td>$\mu_{KM} = \begin{pmatrix} 383.94 \ 442.411 \ 506.765 \ 529.059 \end{pmatrix}$</td>
<td>$\Sigma_{KM} = \begin{pmatrix} 1344.73 &amp; 2324.03 &amp; 2224.37 &amp; 2088.29 \ 2324.03 &amp; 4256.85 &amp; 3968.65 &amp; 3689.9 \ 2224.37 &amp; 3968.65 &amp; 3845.97 &amp; 3542.12 \ 2088.29 &amp; 3689.9 &amp; 3542.12 &amp; 3354.55 \end{pmatrix}$</td>
</tr>
</tbody>
</table>
Figure 4.21: Estimates of $\mu_s$ for 50 Observations

Figure 4.22: Estimates of $\Sigma_3$ for 50 Observations
Table 4.10: Estimates of $\Sigma_3$ and $\mu_3$ for 100 observations

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>Covariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>$\mu_3 = \begin{pmatrix} 368.21 \ 404.63 \ 479.26 \ 502.89 \end{pmatrix}$</td>
<td>$\Sigma_3 = \begin{pmatrix} 2819.29 &amp; 3568.42 &amp; 2943.49 &amp; 2295.35 \ 3568.42 &amp; 7963.14 &amp; 5303.98 &amp; 4065.44 \ 2943.39 &amp; 5303.98 &amp; 6851.32 &amp; 4499.63 \ 2295.35 &amp; 4065.44 &amp; 4499.63 &amp; 4878.99 \end{pmatrix}$</td>
</tr>
<tr>
<td>Complete Sample</td>
<td>$\mu_{\text{Com}} = \begin{pmatrix} 370.95 \ 404.27 \ 473.97 \ 503.83 \end{pmatrix}$</td>
<td>$\Sigma_{\text{Com}} = \begin{pmatrix} 2556.35 &amp; 2810.10 &amp; 2169.24 &amp; 1645.36 \ 2810.10 &amp; 5890.91 &amp; 4048.98 &amp; 2776.14 \ 2169.24 &amp; 4048.98 &amp; 5118.95 &amp; 3526.98 \ 1645.36 &amp; 2776.14 &amp; 3526.98 &amp; 3822.69 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over DL</td>
<td>$\mu_{\text{DL}} = \begin{pmatrix} 379.85 \ 416.34 \ 482.08 \ 511.17 \end{pmatrix}$</td>
<td>$\Sigma_{\text{DL}} = \begin{pmatrix} 1395.15 &amp; 1537.79 &amp; 1321.01 &amp; 1018.34 \ 1537.79 &amp; 3553.19 &amp; 2554.81 &amp; 1886.34 \ 1321.01 &amp; 2554.81 &amp; 3681.20 &amp; 2571.63 \ 1018.34 &amp; 1886.34 &amp; 2571.63 &amp; 2761.81 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over DL/2</td>
<td>$\mu_{\text{DL}/2} = \begin{pmatrix} 337.50 \ 371.80 \ 429.30 \ 453.50 \end{pmatrix}$</td>
<td>$\Sigma_{\text{DL}/2} = \begin{pmatrix} 10333.6 &amp; 7511.6 &amp; 5365.1 &amp; 5858.5 \ 7511.6 &amp; 14978.3 &amp; 9085.8 &amp; 6553.8 \ 5365.1 &amp; 9085.8 &amp; 18515.9 &amp; 10972.4 \ 5858.5 &amp; 6553.8 &amp; 10972.4 &amp; 18665.3 \end{pmatrix}$</td>
</tr>
<tr>
<td>Substitution Over Zero</td>
<td>$\mu_{\text{Zero}} = \begin{pmatrix} 295.10 \ 327.30 \ 376.60 \ 395.90 \end{pmatrix}$</td>
<td>$\Sigma_{\text{Zero}} = \begin{pmatrix} 30154.7 &amp; 18818.7 &amp; 12841.2 &amp; 15237.0 \ 18818.7 &amp; 38404.9 &amp; 21497.1 &amp; 14329.6 \ 12841.2 &amp; 21497.1 &amp; 50214.7 &amp; 26005.3 \ 15237.0 &amp; 14329.6 &amp; 26005.3 &amp; 54694.0 \end{pmatrix}$</td>
</tr>
<tr>
<td>Kaplan Meier</td>
<td>$\mu_{\text{KM}} = \begin{pmatrix} 380.215 \ 416.375 \ 482.419 \ 511.287 \end{pmatrix}$</td>
<td>$\Sigma_{\text{KM}} = \begin{pmatrix} 1351.88 &amp; 2124.38 &amp; 2177.32 &amp; 1894.76 \ 2124.38 &amp; 3513.32 &amp; 3525.76 &amp; 3043.42 \ 2177.32 &amp; 3525.76 &amp; 3604.22 &amp; 3114.08 \ 1894.76 &amp; 3043.42 &amp; 3114.08 &amp; 2722.80 \end{pmatrix}$</td>
</tr>
</tbody>
</table>
Comparison of Estimates of Mean Vector

Figure 4.23: Estimates of $\mu_i$ for 100 Observations

Comparison of the Estimates of the Covariance matrix

Figure 4.24: Estimates of $\Sigma_i$ for 100 Observations

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
CHAPTER 29

CONCLUSIONS AND FUTURE WORK

In this thesis, we have studied the Kaplan Meier method and the performance of KM method in estimating the covariance matrix and the mean vector on multivariate datasets with censored observations.

The results reported in Chapter 4 demonstrates that the KM estimator of the covariance matrix $\Sigma$ is arguably the best estimator when the multivariate dataset has non-detects.

When multivariate sample has non-detects, the estimator of the covariance matrix developed in this thesis can easily be implemented into a principal component analysis (PCA) procedure which is one of the most common tools used to reduce the dimensionality of the multivariate data. We can also use Monte Carlo simulation to investigate the properties of the KM estimator of the covariance matrix, and compare to the existing estimators, in the presence of non-detects.
BIBLIOGRAPHY


VITA
Graduate College
University of Nevada, Las Vegas

Shanmugapriya Saravanan

Home Address:
969 East Flamingo Road, #167
Las Vegas, NV 89119

Degrees:
Master of Science, Applied Mathematics, 2003
Bharathiar University, India

Thesis Title: Estimation of the Covariance Matrix in Presence of Non-detects

Thesis Examination Committee:
Chairperson, Dr. Ashok Kumar Singh, Ph.D.
Committee Member, Dr. Dieudonne D. Phanord, Ph.D.
Committee Member, Dr. Rohan Dalpatadu, Ph.D.
Graduate Faculty Representative, Dr. Laxmi P. Gewali, Ph.D.