RECONSTRUCTION OF GENE REGULATORY NETWORKS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SİBEL BALCI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
STATISTICS

SEPTEMBER 2014

Approval of the thesis:

RECONSTRUCTION OF GENE REGULATORY NETWORKS

submitted by **SİBEL BALCI** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Statistics Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. İnci Batmaz _____
Head of Department, **Statistics**

Prof. Dr. Ayşen Akkaya _____
Supervisor, **Statistics Dept., METU**

Assoc. Prof. Dr. Tolga Can _____
Co-supervisor, **Computer Engineering Dept., METU**

Examining Committee Members:

Assist. Prof. Dr. Zeynep Kalaylıoğlu _____
Statistics Dept., METU

Prof. Dr. Ayşen Akkaya _____
Statistics Dept., METU

Assist. Prof. Dr. Yeşim Aydın Son _____
Health Informatics Dept., METU

Assist. Prof. Dr. Özlem Türker Bayrak _____
Industrial Engineering Dept., Çankaya University

Assist. Prof. Dr. Ceylan Yozgatlıgil _____
Statistics Dept., METU

Date: 05.09.2014

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Sibel Balcı

Signature :

ABSTRACT

RECONSTRUCTION OF GENE REGULATORY NETWORKS

Balcı, Sibel
Ph.D., Department of Statistics
Supervisor: Prof. Dr. Ayşen Akkaya
Co-supervisor: Assoc. Prof. Dr. Tolga Can

September 2014, 136 pages

With the development of microarray technology, it is now possible to obtain the concentration levels of thousands of genes at a given time or in a given state. By following the changes in the gene expression levels, the responsible genes for cell differentiation or certain diseases can be identified. Gene expression changes are regulated by the interactions between the genes and their products. Gene regulatory networks (GRNs) identify these interactions using the gene expression changes. There are a number of statistical methods to infer GRNs, however, most of them depend on the normality assumption of noises in the data. This thesis considers the multiple linear regression analysis for the reconstruction of GRNs when the error term comes from a Weibull distribution. Since nonnormality complicates the data analysis and results in inefficient estimators, it is proposed to use the modified maximum likelihood (MML) estimation procedure which produces efficient and robust estimators. Also, explanatory variables representing the gene expression levels come from a Weibull distribution. Therefore, they are considered as stochastic and stochastic multiple linear regression analysis is used for inferring GRNs by implementing MML method to estimate the model

parameters. Robustness and power analyses for both stochastic and nonstochastic multiple linear regression model parameters are also given.

Keywords: Gene Regulatory Networks, Weibull Distribution, Multiple Stochastic Linear Regression, Modified Maximum Likelihood Estimation.

ÖZ

GEN DÜZENLEYİCİ AĞLARIN YENİDEN OLUŞTURULMASI

Balcı, Sibel
Doktora, İstatistik Bölümü
Tez Yöneticisi: Prof. Dr. Ayşen Akkaya
Ortak Tez Yöneticisi: Doç. Dr. Tolga Can

Eylül 2014, 136 sayfa

Mikrodizin teknolojisinin geliştirilmesiyle, binlerce genin konsantrasyon düzeylerinin belirli bir zaman ya da belirli bir durum için elde edilmesi artık mümkün. Gen ifade düzeylerindeki değişimlerin takibi ile hücre çeşitliliğine ya da belirli bir hastalığa neden olan genler belirlenebilmektedir. Gen ifadelerindeki değişimler, genler ve gen ürünleri arasındaki etkileşimlerle düzenlenmektedir. Gen düzenleyici ağlar (GDA), gen ifadelerindeki değişimlerini kullanarak bu etkileşimleri ortaya çıkarmaktadır. GDA'ın çıkarımı için kullanılan çok sayıda istatistiksel yöntem mevcuttur ancak birçoğu verideki hataların normallik varsayımına dayanmaktadır. Bu tez, GDA'ın yeniden oluşturulması için hata terimi Weibull dağılımına sahip olan çoklu doğrusal regresyon analizini ele almaktadır. Normal dağılmama durumunun veri analizini zorlaştırmasından ve etkin olmayan tahmincilere yol açmasından dolayı, etkin ve güçlü tahminciler üreten uyarlanmış en çok olabilirlik (UEÇO) tahmin yönteminin kullanılması önerilmektedir. Ayrıca, gen ifade düzeylerini gösteren açıklayıcı değişkenler de Weibull dağılımdan gelmektedir. Bu nedenle, bu değişkenlerin olasılıksal olduğu düşünülmekte ve GDA'ın çıkarımında model parametrelerini tahmin etmek için

UEÇO tahmin yöntemi uygulanarak olasılıksal çoklu doğrusal regresyon analizi kullanılmaktadır. Ek olarak, hem olasılıksal hem de olasılıksal olmayan çoklu doğrusal regresyon model parametreleri için sağlamlık ve güç analizleri verilmiştir.

Anahtar Kelimeler: Gen Düzenleyici Ağlar, Weibull Dağılım, Çoklu Olasılıksal Doğrusal Regresyon, Uyarlanmış En Çok Olabilirlik Tahmini.

To My Parents

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor Prof. Dr. Ayşen Akkaya for her excellent guidance, help, support and patience. She has been much more than an advisor to me. I have learned many professional and personal skills from her.

I am deeply thankful to my co-advisor Assoc. Prof. Dr. Tolga Can for being supportive and helpful throughout the process of this study. He will always be a role model for me.

It is my pleasure to acknowledge my committee members Assist. Prof. Dr. Zeynep Kalaylıoğlu, Assist. Prof. Dr. Yeşim Aydın Son, Assist. Prof. Dr. Özlem Türker Bayrak and Assist. Prof. Dr. Ceylan Yozgatlıgil. They have provided, with kindness, their insight and suggestions, which are very precious to me.

I would like to present my gratitude to The Scientific and Technological Research Council of Turkey (TUBITAK) for providing the scholarship that enabled me to complete my study without any difficulties.

I am also indebted to my dear friends Mert, Gül and Duygu for their generous support and encouragement during my worst moments.

Cem deserves my very special thanks for helping me in every aspect. He tolerates me when even I cannot tolerate myself. All I want is he stands by my side.

I owe my deepest gratitude to my parents. Without their unconditional love, I would not be who I am today. Also, I will always appreciate my beloved brother

and sister for being there for me when I need them. And, I am especially thankful to Öykü, Onur and Defne for giving me renewed hope.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

A	Adenine
AMML	Adaptive Modified Maximum Likelihood
ANOVA	Analysis of Variance
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
BIC	Bayesian Information Criterion
C	Cytosine
cDNA	Complementary Deoxyribonucleic Acid
CGH	Comparative genomic hybridization
cis-eQTL	*cis*-expression quantitative trait loci
DBN	Dynamic Bayesian Network
DNA	Deoxyribonucleic Acid
DPN	Dynamic Probabilistic Networks
FTSS	Fourier Transform for Stable Systems
G	Guanine
GA	Genetic Algorithm
GGM	Gaussian graphical model
GRAM	Genetic Regulatory Modules
GRN	Gene Regulatory Network
LOWESS	Locally Weighted Regression and Smoothing Scatterplots

LS	Least Squares
LSE	Least Squares Estimator
ML	Maximum Likelihood
MML	Modified Maximum Likelihood
mRNA	Messenger RNA
MSE	Mean Squared Error
MVB	Minimum Variance Bound
MWSLE	Minimum Weight Solutions to Linear Equations
NCA	Network Component Analysis
NIR	Network Identification by Multiple Linear Regression
ODE	Ordinary Differential Equation
PBN	Probabilistic Boolean Networks
PCR	Polymerase Chain Reaction
RE	Relative Efficiency
REVEAL	Reverse Engineering Algorithm
RNA	Ribonucleic Acid
RNA-Seq	RNA Sequencing
RSS	Sum of Squares of Residuals
SAGE	Serial Analysis of Gene Expression
SDE	Stochastic Differential Equation
SEM	Structural Equation Model
SML	Sparsity-Aware Maximum Likelihood
SSE	Sum of Squared Errors

T	Thymine
TBN	Temporal Boolean Network
TdGRN	Time-delayed Gene Regulatory Network
TF	Transcription Factor
TSNI	Time series Network Identification
U	Urasil

CHAPTER 1

INTRODUCTION

In all living organisms, there is a hierarchical organization of small building blocks. *Cell* is the smallest unit of this hierarchy. Combination of cells having a special structure and function forms *tissues*. Different kinds of tissues compose an *organ*. Several different organs work together to perform a certain task as an *organ system.* And finally, different organ systems come together and form the organisms.

All genetic information which determines the function of a cell is encoded in the Deoxyribonucleic Acid (DNA) sequence. A gene is a sub segment of DNA and all genes in the genome composes the set of instructions that organism need to survive. The DNA code of genes is converted into Ribonucleic Acid (RNA), which codes for protein products. Amount of the products produced by a particular gene is the expression level of that gene. While all cells in an organism have the same genomic DNA, so the same gene sequences, the expression levels of many genes differ in different kinds of cells and under different conditions. Cell differentiation and cell function are regulated by differential gene expression. If researchers know the conditions under which genes are expressed at high level, they can get hints about the function of those genes.

Gene expression levels are also associated with the disease recurrence. When there is a malfunction at any of the building blocks, organisms cannot perform normally and diseases occur. Most diseases are result from the abnormal activity of genes in the cells. While an organism is leading to diseased state from the

health state, expression levels of the genes related to the disease change. For example, onco genes in the cancer cells are expressed at high levels, but tumor suppressor genes are expressed at lower levels. By comparing the expression levels of the genes in diseased and normal cells, genes responsible for various diseases can be identified and possible therapeutic targets of the drugs can be determined.

Quantification of the gene expression level profiles under different conditions is an important part of the biological and medical research. With the development of high throughput technologies such as DNA microarray and RNA sequencing (RNA-Seq) in molecular biology, researchers now can get the information about the concentration levels of thousands of genes at a given time or in a given state of an organism. However, identification of the responsible genes for cell differentiation or certain diseases by measuring the gene expression levels is not sufficient alone. It is also important to determine how gene products are governed. Cells need every gene product neither at the same time nor in the same amount. In a cell, genes work together by interacting with one another and interdependencies between them determine which, when and how much product is produced by a particular gene, that is, gene expression levels are regulated by these interactions between the genes. Hence, inferring gene regulatory networks (GRNs) becomes a necessity to understand the molecular mechanism of the life.

GRNs identify the interactions between the genes and their products using gene expression data. They describe how the genes are expressed by a cell, which genes are transcribed to RNA and which of them in turn are used for the protein synthesis. By GRNs, the relationships between the genes and their regulators can be visualized mapping the interactions between them onto a graphic.

A large number of studies have been carried out for inferring or reverse-engineering GRNs. Kauffman (1969) has introduced Boolean networks to obtain

GRNs by modeling the gene as a binary device that can realize any one, but only one, of the possible Boolean functions of its K inputs. A generalization of the Boolean networks which is called the Temporal Boolean Network (TBN) has been proposed by Silvescu and Honavar (1997) to examine the dependencies among the activity of genes that span for more than one unit of time. Friedman et al. (1998) have handled the problem of learning dynamic probabilistic networks (DPN) from complete data by the extending Bayesian Information Criterion (BIC) scores and from incomplete data by extending structural equation model (SEM) algorithm. Liang et al. (1998) have investigated the possibility of inferring a complex regulatory network architecture from input/output pattern of its variables and implemented Reverse Engineering Algorithm (REVEAL) using mutual information measures. Muphy and Mian (1999) have showed that the most of the proposed discrete time models in reverse-engineering genetic networks from time series data are all special cases of a general class of models called Dynamic Bayesian Networks (DBNs) and reviewed the used techniques to learn DBNs. Chen et al. (1999) have proposed a differential equation model for gene expression and developed two methods, Minimum Weight Solutions to Linear Equations (MWSLE) and Fourier Transform for Stable Systems (FTSS), to construct model from experimental data. Shmulevich et al. (2002) have introduced the model of Probabilistic Boolean Networks (PBN) which have the rule-based properties of Boolean networks, but are robust in the face of uncertainty. Also, Bar-joseph et al. (2003) have developed an algorithm called Genetic Regulatory Modules (GRAM) which combines information from genome-wide location and expression data sets to explore regulatory networks of gene modules. Kikuchi et al. (2003) have improved the method proposed by Tominaga and Okamoto (1998) for the dynamic modeling of complex biosystems combining a Genetic Algorithm (GA) and the S-system and compared these basic and modified methods. Gardner et al. (2003) have constructed a first-order model of regulatory interactions in a nine-gene subnetwork of the SOS pathway in Escherichia coli and indicated the model to identify correctly the major regulatory

genes and transcriptional targets of mitomycin C activity in the subnetwork. Perrin et al. (2003) have dealt with the identification of GRNs from experimental data using a statistical machine learning approach and proposed a stochastic model of gene interactions capable of handling missing variables. They have estimated the model parameters by a penalized likelihood maximization method. Kao et al. (2004) have handled the complex transcriptional networks and showed the utility of network component analysis (NCA) in determining the multiple transcription factor activities. Ott et al. (2004) have developed an algorithm to obtain the optimal Bayesian networks of considerable size overcoming the uncertainties of heuristic approaches that makes it difficult to draw conclusions from networks estimated by heuristics. Nachman et al. (2004) have presented fine-grained dynamical models of gene transcription and proposed an algorithm based on DBNs to reconstruct these models of GRNs. Laubenbacher and Stigler (2004) have proposed an approach constructing a regulatory network as a time-discrete multi-state dynamical system after they have described some of the existing reverse-engineering methods. Xing and Laan (2005) have described a comprehensive statistical approach to obtain transcriptional regulatory networks using gene expression data, transcription factor binding sites and promoter sequences. Chen et al. (2005) have introduced a stochastic differential equation (SDE) model for the transcriptional regulatory network of the time-course gene expression datasets. They have applied this model to the cell-cycle data of budding yeast Saccharomyces cerevisiae and tried to fit a generalized linear model estimating the transcription pattern of a specific target gene. Boscolo et al. (2005) have addressed the NCA on the basis of some aspects. They have used two-stage least square iterative procedure in NCA and introduced a framework to reconstruct multiple regulatory subnetworks simultaneously. Margolin et al. (2006) have introduced the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) that is a novel-theoretic algorithm for inferring the transcriptional networks from microarray data. Using delayed correlations between genes, Li et al. (2006) have developed a toolbox called Time-delayed

Gene Regulatory Network (TdGRN) to reconstruct regulatory networks from temporal gene expression data. Cho et al. (2006) have given S-tree based genetic programming for both the structural and dynamical modeling of a biological network and estimating the network parameters. Sabatti and James (2006) have introduced a framework which uses DNA sequence information and expression arrays data in concert to analyze the effects of a collection of regulatory proteins on genomic expression levels. Bansal et al. (2006) have presented Time Series Network Identification (TSNI) algorithm which obtain the local network of gene-gene interactions surrounding a gene of interest by perturbing only one of the genes in the network. Bansal et al. (2007) have compared different algorithms used to infer gene networks and showed that these algorithms can correctly reverse-engineer the gene interactions. Cho et al. (2007) have also presented various techniques of reverse-engineering GRNs from gene expression profiles and biological information and arranged systematically these techniques based on the required information. Kaderali and Radde (2008) have handled several approaches given for discovering GRNs and discussed their strengths and weaknesses, also provided information on which models are appropriate under what circumstances and future developments. Faith et al. (2008) have developed context likelihood of relatedness to obtain the transcriptional regulatory relations using transcriptional profiles of an organism across a diverse set of conditions. Scrutinizing several kinds of computational methods used in predicting GRNs in mammalian cells, Lee and Tzou (2009) have showed how the power of different knowledge databases of different types can be used to identify modules and subnetworks. Emmert-Streib et al. (2012) have reviewed the methods available for estimating the GRNs and compared two major approaches with contemporary ones. Sparse structural equation models have been used (Cai et al., 2013) to integrate both gene expression data and *cis*-expression quantitative trait loci (*cis*-eQTL), for modeling gene regulatory networks in accordance with biological evidence about genes regulating or being regulated by a small number of genes. A

systematic inference method named sparsity-aware maximum likelihood (SML) has been also developed for SEM estimation.

1.1 Motivation of the Study

The motivation of this dissertation comes from the work of Gardner et al. (2003). In their study, they develop an algorithm called Network Identification by Multiple Linear Regression (NIR) in which a model of the connections and functional relations between elements in a network is inferred from measurements of system dynamics by applying multiple linear regression analysis.

They use the method of least squares (LS) to estimate the parameters of the multiple linear regression model. While constructing the model, they assume that the noise term and explanatory variables representing the expression levels of genes in the model are normally distributed. However, when the real data used in their study is examined, it is seen that the residuals obtained by using LS estimators (LSEs) fit Weibull distribution better. Similarly, conditional distributions of explanatory variables are obtained as Weibull distribution, too. Since the explanatory variables are stochastic, LS estimators of the model parameters are not same with the maximum likelihood (ML) estimators anymore. In addition, the relations between the explanatory variables are not taken into consideration in their study.

In our study, it is proposed to use a stochastic multiple linear regression model when error term and explanatory variables come from a Weibull distribution considering the dependency between the explanatory variables to construct GRNs.

1.2 Aim and Contribution of the Study

The main focus of this dissertation is to obtain a statistical computational method that can be used for inferring gene regulatory networks from gene expression data by researchers. It is aimed to improve the NIR algorithm by dealing with the statistical assumptions needed to apply NIR algorithm. This study makes the following contributions:

- It is known that non-normality complicates the data analysis and results in inefficient estimators. Therefore, it is very important to improve statistical procedures which are efficient and robust to deviations from an assumed distribution. This study provides a robust estimation technique for the multiple linear regression analysis when the noise has a Weibull distribution by estimating the model parameters using the method of modified maximum likelihood (MML) estimation.

- As mentioned at Section 1.1, explanatory variables represent the gene expression levels and they are subject to the measurement errors. Therefore, they are stochastic and they have a distribution. The parameter estimators obtained by using NIR algorithm are not the ML estimators anymore when the explanatory variables are stochastic, which means that the estimators of the model parameters obtained by using NIR algorithm lost their good properties. This study handle this problem using stochastic multiple linear regression analysis.

- Lastly, Gardner et al. (2003) ignore the relationships between the explanatory variables in their study. However, some explanatory variables are collinear since the gene expression levels are regulated by the interactions between genes. Therefore, this study takes into account the relationships between the explanatory variables and estimate the partial correlation coefficients between the explanatory variables by

implementing method of MML in the stochastic multiple linear regression model.

1.3 Organization of the Study

This thesis consists of six chapters. In Chapter 1, a brief introduction to gene regulatory networks is given by emphasizing the importance of them in molecular biology. Also, the publications related to the GRNs are presented comprehensively and the motivation of the thesis is described. Furthermore, the aim and the contributions of the study are stated.

In Chapter 2, a biological background of gene regulation needed to understand the rest of the thesis is provided. It explains the gene expression and mentions some high throughput techniques used to measure the gene expression levels. It also reviews the existing methods used for inferring gene regulatory networks. Especially, NIR algorithm is examined in detail since it is the motivation of this study.

Chapter 3 gives the theoretical explanation of the multiple linear regression models. Since the expression data used in the regression model fit a Weibull distribution, Weibull distribution and its properties are also described in this chapter. Then, MML and LS estimators of the parameters in the multiple linear regression model with nonstochastic covariates are derived and the test statistics based on LS and MML estimators are obtained to test significance of model parameters.

In Chapter 4, stochastic linear regression model is used to infer GRNs and model parameters are estimated by using MML and LS estimation methods considering the relationships between the explanatory variables. Test statistics based on LS and MML estimators are also obtained for stochastic linear regression model.

In Chapter 5, MML and LS estimation techniques are compared by examining the efficiency, robustness and power properties of them through a comprehensive simulation study. In addition, a real life application is given in this chapter.

Finally, the last chapter of the thesis concludes the work that has been done, suggests some ideas about the gene regulatory networks and gives the related future work.

CHAPTER 2

BIOLOGICAL AND HISTORICAL BACKGROUND

This chapter presents a biological background to elucidate the gene expression and the gene regulation comprehensively by giving the definitions of some genetic materials such as cell, genes and DNA etc. Also, it describes the high throughput techniques used to measure gene expression levels and explains the microarray technology in detail. Furthermore, some commonly used methods for inferring gene regulatory networks are reviewed in this chapter.

2.1 Biological Background

Cells are the minimal units of all living organisms that contain a multitude of specific chemical transformations providing the energy needed by the cells and coordinating all of the events (Lee, 2004). The regulation of gene expression levels is maybe the most important task of cells to meet their needs and to adopt the environmental changes.

Macromolecules such as DNA, RNA and proteins define the structure of cells and govern most of the activities of life (Lee, 2004) and especially play the main roles in the process of the expression of the genetic information.

DNA is a double-stranded and helical molecule composed of four nucleotides: adenine (A), guanine (G), thymine (T) and cytosine (C). The sequence of these four nucleotides encodes the genetic information stored in DNA and hence, gives the genetic instructions for the development and the proper functioning of the

organisms. Since each strand of the DNA molecule is the complementary of the other, the double helix structure of the DNA molecules adds nothing to the information contained in a single strand. In DNA, A pairs with T, and C with G.

Genes are segments of DNA and contain specific instructions which allow a cell to produce a specific product. Although every cell of an individual organism contains the same DNA, carrying the same information, different kinds of cells are available. As mentioned in Chapter 1, this differentiation is resulted from that all the genes are not expressed in the same way in all cells (Draghici, 2003).

Cells need the products of some kind of genes called *housekeeping genes* at all time. It is assumed that these genes are expressed at constant levels in different cell types. However, expression levels of particular genes are affected from their circumstances and changes in their expression levels determine the distinct biological characteristics and hence cause organismal complexity and diversity.

Differentiation between cells is given by different patterns of gene activations which in turn control the production of proteins. A gene is active, or expressed, if the cell produces the protein encoded by the gene. If a lot of protein is produced, the gene is said to be highly expressed. If no protein is produced, the gene is not expressed or unexpressed. The objective of researchers is to detect and quantify gene expression levels under particular circumstances (Draghici, 2003).

Gene expression is the most fundamental level at which genotype gives rise to the phenotype. It is the entire process that takes the information contained in genes on DNA and turns that information into proteins. Gene expression occurs in three major stages: *Replication, Transcription* and *Protein Synthesis* (or *Translation*). In the replication process, a double-stranded DNA molecule is duplicated to give identical copies. RNA, a single-stranded molecule which uses a nucleotide called uracil (U) instead of thymine present in DNA, is transcribed from DNA by

enzymes called RNA polymerases and is generally further processed by other enzymes. This process is called transcription. In the process of protein synthesis, RNA sequence is translated into a sequence of amino acids. A combination of 20 different amino acids forms the proteins. Proteins are the complex organic compounds consisting of the immediate expression of the genetic information stored in DNA and attending various tasks essential for survival of the cell. These three stages are all together called the central dogma of molecular biology (Watson and Crick, 1958; Crick, 1970) and presented in Figure 2.1.

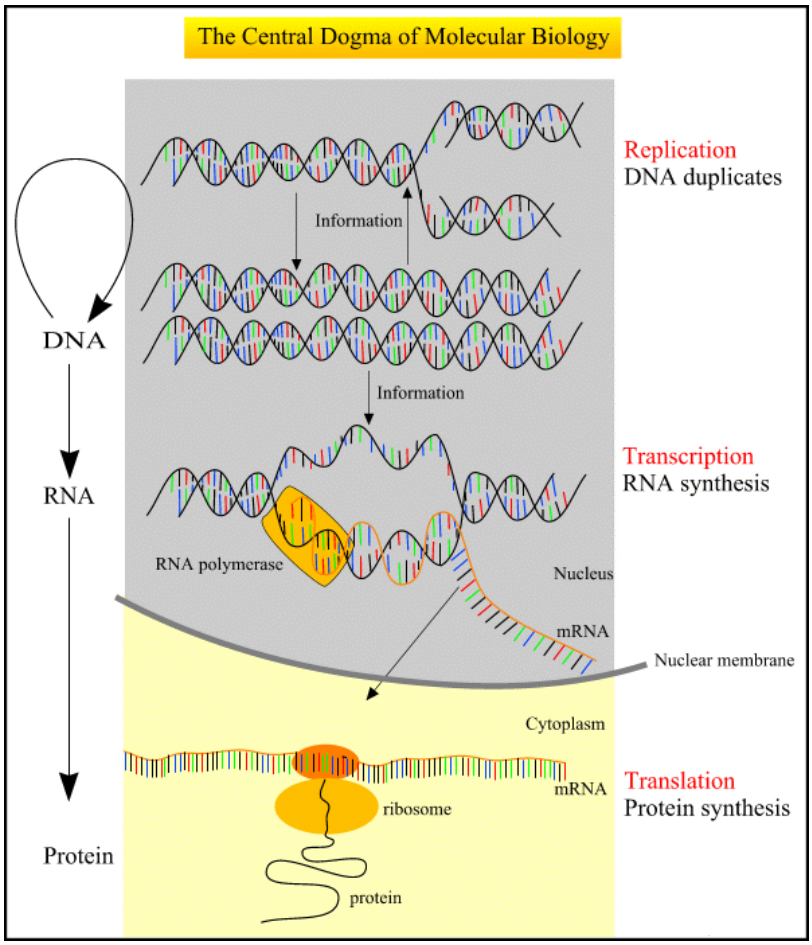


Figure 2.1: Central dogma of molecular biology. *Figure is adapted from* "http://users.ugent.be/~avierstr/principles/centraldogma.html".

2.2 Gene Regulatory Networks

Regulation of gene expression controls the amount and timing of a functional gene product. In the process of gene expression, transcription factors (TFs), which are the specialized proteins, bind to promoter region of DNA and intervene the rate of protein synthesis. An increase in the rate of protein synthesis is called as the activation or up-regulation of the gene and a decrease in the rate of protein synthesis is called as the inhibition or down-regulation of the gene (Panse and Kshirsagar, 2013).

Genes regulate the expression levels by interacting each other through gene products. A gene regulatory network is a collection of genes and gene products (RNAs and proteins) and describes the regulatory relationships between genes, proteins and other cellular components. Gene regulation can be visualized by graphs in which nodes show genes or gene products and directed edges connecting nodes show the dependency between them. In Figure 2.2, the graphical representation of a simple gene regulatory network is illustrated.

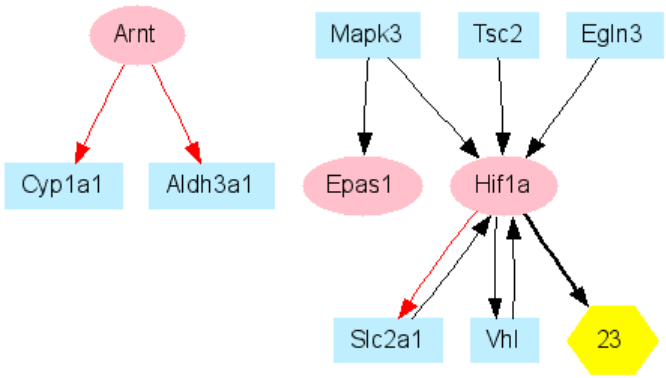


Figure 2.2: An example of gene regulatory network: ellipses are TFs; boxes are genes; hexagons are the clustered genes. *Figure is adapted from* "http://rulai.cshl.edu/TRED/GRN/HIF.htm".

Reconstruction of gene regulatory networks holds great importance especially in the field of system biology. Accurate prediction of GRNs provides an opportunity to study the dynamics of specific gene under particular diseased or experiment conditions. It also helps to study diseases that are caused by dysregulated genes. Hence, GRNs enables to develop new treatment methods for illnesses and to analyze the effects of drugs on genes (Karlebach and Shamir, 2008; Panse and Kshirsagar, 2013).

2.3 Microarray Technology

With the development of high throughput technologies, expression levels of thousands of genes can be measured simultaneously. Gene expression data allow researchers to infer gene regulatory networks by observing changes in gene expression profiles under various experiment conditions and under different cell cycle stages. Hence, behaviors of genes can be analyzed.

There are several kinds of molecular biology techniques, as listed below, to quantify the gene expression and microarrays and next generation RNA sequencing are the most current high-throughput techniques:

- Comparative Genomic Hybridization (CGH)
- Serial Analysis of Gene Expression (SAGE)
- RNA Sequencing
- Real Time- Polymerase Chain Reaction (PCR)
- Microarrays

In this subsection, only microarray analysis of gene expression is handled since it is the most widely used technique.

A DNA microarray, also known as DNA chip, has been introduced by Schena et al. (1995) for the first time and become a very popular technique to identify the gene expression changes resulted from a variety conditions such as development, aging and diseases or drugs (Alizadeh et al ., 2000; Bilban et al., 2000; Tanaka et al., 2000; Young, 2000). It consists of a small membrane or glass slide containing samples of many genes arranged in a regular pattern. The surface of a microarray is spotted with oligonucleotides that are the small parts of DNA molecules up to 25 nucleotides, complementary DNA (cDNA) or small fragments of polymerase chain reaction products that represent specific gene coding regions. There are thousands of microscopic spot known as probes on a microarray and each probe corresponds to a particular gene (Amaratunga and Cabrera, 2004).

Microarrays can be classified as single-channel (one-color) and two-channel (two-color) arrays and both types of microarray are used for the hybridization experiments. One-color microarray measures the intensity of only one hybridized biological sample while two-color microarray measures expression ratios between two hybridized samples.

DNA microarray technology depends on the parallel hybridization of labeled target to immobilized probes (Schena et al., 1995). Differential gene expression is determined by using a two-color scheme. Firstly, the messenger RNA (mRNA) is isolated from the experimental samples such as healthy or tumor tissue sample and reverse-transcribed into more stable complementary DNA. Then cDNA samples are labeled by a fluorescent dye (generally healthy cDNA is labeled green and tumor cDNA is labeled red) and combined sample is hybridized to the microarray chip. Target sample binding to a probe generates a signal and its strength depends upon the amount of target sample binding to that probe. Then, fluorescent intensity on each probe is measured and converted into the raw data by using a special scanner.

Active genes produce many mRNA molecules, hence, many labeled cDNA samples, and generate a very bright fluorescent spots. Genes that are less active produce fewer mRNA molecules, thus, less labeled cDNA samples, and generate dimmer fluorescent spots. If there is no fluorescence, none of the messenger molecules have hybridized to the DNA which indicates that the gene is inactive.

Application areas of microarrays can be summarized as follows:

- *Gene discovery*: Microarray technology is used to identify genes and to determine their function and expression levels at the particular condition (Cho et al., 1998; Chu et al., 1998; Tao et al., 1999; Laub et al., 2000; Wei et al., 2001; Chan et al., 2003).
- *Gene regulation studies:* Microarray technology is used to infer gene regulatory networks describing the regulatory relationships between genes and gene products (de Saizieu et al., 2000; Gross et al., 2000; Arfin et al., 2000; Kuhn et al., 2001; Britton et al., 2002).
- *Disease diagnosis:* Microarray technology is used to determine disease by the identification of changes in the expression levels of particular genes (Gingeras et al., 1998; van't Veer et al., 2002; Macoska 2002).
- *Drug discovery and toxicology:* Microarray technology is used to develop treatments for illnesses by studying the therapeutic responses to drugs. It is also used to search the impacts of toxins on the cells (Wilson et al., 1999; Bammert and Fostel, 2000; Clarke et al., 2001)

2.4 Data Analysis Preparation

As mentioned in Section 2.3, microarray technology measures the labelled fluorescent intensities which represent the amount of mRNA molecules isolated from the experimental samples and converts these intensities to the gene expression data. However, microarray experiments are generally subject to some

sources of variations and these variations mask the biological signals of the actual interest, which means that fluorescent intensity changes may not always show the actual expression changes.

Measurement errors that affect the expression data can be classified into two categories: systematic error and random error. Systematic error is a bias resulting from array spotting, scanning, labelling, hybridization etc., and it reflects the accuracy of experiment measurements (Claverie, 1999; Schuchhardt et al., 2000; Lou et al., 2001; Tseng et al., 2001; Yue et al., 2001). For example, a well-known systematic error is fluorescent dye bias. When two identical samples are labelled with red and green colors and hybridized to same slide, it is expected that green intensities and red intensities are at the same level since there is no differential expression. However, red intensities generally tend to be lower than green intensities (Smyth et al., 2003). Once systematic errors are identified and removed, it is considered that the remaining measurement errors are random. Random error is a measure of uncertainty in the measurements and reflects the precision of expression data. It constitutes a noise which prevents the changes in biological signals to be determined correctly. Changes in the expression levels can be distinguished from random noise by using some statistical tests.

Systematic errors can be removed or controlled by using strict experimental procedures and employing normalization methods. After all background corrections are carried out, normalization of microarray data have to be performed to make observations comparable each other. In microarray data analysis, it is generally aimed to identify the differentially expressed genes by comparing the expression levels of genes under different conditions, for this reason, gene expressions are represented as the ratio of two florescent intensities (Parmigiani et al., 2003). Although ratios provide an intuitive measure of expression changes, they treat up-regulated and down-regulated genes differently. The intensity ratios usually have a skewed distribution since the ratios of down-regulated genes take

the values in the interval $(0,1)$ whereas the ratios of up-regulated genes take the values in the interval $(1,\infty)$. For example, genes up-regulated by a factor of 2 have an expression ratio of 2 whereas those down-regulated by the same factor have an expression level of -0.5. To overcome this disadvantage of the ratios, expression data need to be transformed before the normalization. The most commonly used transformation is the logarithm base 2 transformation. It provides the genes up-regulated and down-regulated by a factor of 2 to have a $log_2(ratio)$ of 1 and -1, respectively (Quackenbush, 2002).

A number of normalization approaches have been introduced to remove the systematic errors. The most well-known approaches can be listed as

- Global Normalization
- Total RNA Normalization
- Self-Normalization
- Housekeeping Gene Normalization
- Locally Weighted Regression and Smoothing Scatterplots (LOWESS)

Global normalization is the first proposed method for the normalization of the microarray data. This approach relates the red intensity to the green intensity by a multiplicative constant and shifts the center of the distribution of transformed expression ratios to zero (Yang et al., 2002). Total RNA normalization assumes that amount of total RNA carried by each cell does not change over time (Fang et al., 2003). The self-normalization method removes the systematic error by applying a subtract operation to the data since the error on log scale is assumed to be additive. The housekeeping gene normalization evaluates the labelling and sample hybridization by spotting a set of housekeeping genes on the array. In this approach, it is assumed that the housekeeping genes are expressed at a constant level under different experimental conditions (Yang et al., 2002). Transformed expression ratios have some intensity-dependent variations and LOWESS

normalization removes these variations by applying a smoothing adjustment (Cleveland, 1979; Quackenbush, 2002).

Unlike systematic errors, random errors cannot be removed entirely but they can be estimated from the observed data. By the replication of the experiment, random errors can be minimized since it is expected that replicates give same results under the same condition except for random error (Nadon and Shoemaker, 2002).

2.5 Historical Background

Since the modelling of gene regulatory networks has become a very useful tool for the analysis of the gene interactions, numerous methods have been proposed to construct gene regulatory networks in the literature. These methods can be classified as *physical approach* and *influence approach.* In the physical approach, the proteins regulating the transcription and DNA motifs to which they bind are identified, and hence, true molecular interactions are determined. However, the influence approach does not seek true physical interactions, instead, it describes the regulatory influences between RNA transcripts by observing the changes in the transcription levels. For the modelling of gene regulatory networks, the influence approach is more preferable to the physical approach since the physical approach needs more prior knowledge and specific data. Models used for inferring gene regulatory networks are also divided into two groups as dynamic and static. Dynamic models which contain a time-component are used when the dynamic behavior of the network is required (Hecker, 2007).

In this study, the most widely used methods are reviewed. These are listed as follows:

- Boolean Networks
- Gaussian Graphical Models
- Bayesian Networks
- Ordinary Differential Equations
- Network Identification by Multiple Linear Regression

2.5.1 Boolean Networks

Boolean networks have been used firstly by Kauffman (1969) to model behavior of the large nets of randomly interconnected binary genes. In the modelling of gene regulation by Boolean networks, each node represents a gene and directed edge represents biological interaction between two genes.

A Boolean network can be explained by the definition given below:

Definition 1 (Boolean Network): A Boolean network is a tuple $G=(X,B)$ where $X=(x_1,x_2,...,x_n)\in\{0,1\}^n$ is a vector of Boolean variables, and B is a set of Boolean functions $B=\{f_1,f_2,...,f_n\}$, $f_i:\{0,1\}^n\rightarrow\{0,1\}$ (Kaderali and Radde, 2008).

In the gene regulatory networks, x_i represents the state of gene i and f_i represents the interactions between them. It is assumed that each gene can be in either state "on" or "off". State "on" means that gene is expressed above some threshold while state "off" means that gene is expressed below that threshold. If the gene i is in the state "on", x_i takes the value of 1 and if the gene i is in the state "off", x_i takes the value of 0.

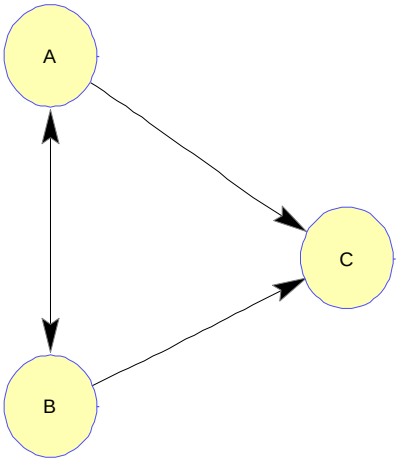
By using Boolean functions, the states of all genes are updated at a discrete time step:

$$x_i(t+1) = f_i(x_1(t), x_2(t), \ldots, x_n(t)) \tag{2.1}$$

Here it is assumed that genes update their states simultaneously. At any given time *t*, the expression or state of the network are represented by the states of all nodes as follows:

$$x(t) = (x_1(t), x_2(t), \ldots, x_n(t)) \tag{2.2}$$

An example of the Boolean networks is shown in Figure 2.3. Here, pointed array shows an activation. For example, if gene B is active, then it will activate gene A. If gene A is active, then it will activate gene B. Also if either gene A or gene B is active, then gene C will be activated.



Input			*Output*		
A	*B*	*C*	*A'*	*B'*	*C'*
0	0	0	0	0	0
0	0	1	0	0	0
0	1	0	1	0	1
0	1	1	1	0	1
1	0	0	0	1	1
1	0	1	0	1	1
1	1	0	1	1	1
1	1	1	1	1	1

Figure 2.3: A representation of sample Boolean network.

Because of the dynamic properties of Boolean networks, this method is quite popular for the reconstruction of gene regulatory networks. However, forming an accurate network is not an easy issue since it is impossible to determine the values

of 2^n states in a Boolean network with *n* nodes. Instead, Kauffman (1969) have introduced *NK* Boolean networks which studies a class of Boolean networks of *n* nodes. In this approach, each node has a randomly selected *k* inputs from *n* nodes and has $n!/(n-k)!$ possible combination of k inputs. In addition, there are 2^{2^k} possible functions for each node. Hence, the number of possible networks is obtained as follows:

$$\left(2^{2^k}\frac{n!}{(n-k)!}\right)^n \tag{2.3}$$

As the number of nodes *n* increases, the number of the possible networks grows exponentially. To solve this problem, the number of edges directed into a node is bounded by a constant.

2.5.2 Gaussian Graphical Models

Gaussian graphical model (GGM) is a very popular approach to the reconstruction of gene regulatory networks (Dobra et al., 2004). In this approach, it is assumed that the available data come from a multivariate Gaussian distribution (Whittaker, 1990). Hence, the aim is to determine the conditional independencies among genes by deriving the partial correlations in the joint probability distribution of expression data.

Gaussian graphical models give the direct association between genes but indirect associations can also be obtained easily (Wang et al., 2013). GGM is described by a graph $G=(V,E)$ where $V=\{1,2,...,p\}$ corresponds to the node sets representing the variables and $E=(e_{ij})$ corresponds to the edge set representing conditional independencies between nodes. If there is no edge between two nodes

($e_{ij} = 0$), then these two nodes are conditionally independent given all other nodes.

In the GGM, $X = (X_1, X_2, ..., X_p)$ represents the real valued states of nodes and follow a multivariate Gaussian distribution with mean 0 and covariance matrix Σ. Hence, partial correlations can be obtained by the inverse covariance matrix $\Omega = \Sigma^{-1} = \{ w_{ij} \}$. They are obtained as follows:

$$\rho_{ij} = -\frac{w_{ij}}{\sqrt{w_{ii} w_{jj}}} \tag{2.4}$$

where ρ_{ij} is the partial correlation between gene *i* and gene *j* given all other genes. If e_{ij} is 0, then w_{ij} becomes 0 and 0 valued elements in the inverse covariance matrix give the conditionally independent genes in the network.

2.5.3 Bayesian Networks

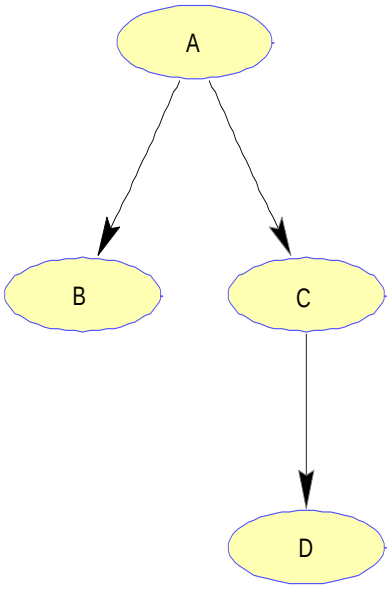
Bayesian network model has been proposed by Friedman et al. (2000) and Hartemink et al. (2001) for inferring gene regulatory networks. It determines the probabilistic relationships between the nodes of the network by establishing a directed acyclic graph. Directed acyclic graph is denoted by $G = (X, A)$ where the nodes $X = (X_1, X_2, ..., X_n)$ correspond to the random variables representing the expressions of genes and the directed edges A represent the probabilistic dependencies between the random variables. An edge from X_j to X_i shows the dependency of X_i on X_j. In this case, X_j is called a parent of X_i. Therefore, X_i has a conditional probability distribution denoted by $p(x_i \mid parents(x_i))$ where $parents(x_i)$ is the set of parents for X_i. If X_i does not have a parent, then it is

unconditional probability distribution $p(x_i)$. In a Bayesian network, it is assumed that each random variable is independent of its non-descendants. Hence, the joint probability distribution function of $X_1, X_2, ..., X_n$ can be written as follows:

$$p(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p(x_i \mid parents(x_i)) \tag{2.5}$$

Figure 2.4 gives a simple Bayesian network consisting of four nodes A, B, C and D with discrete states "$on = 1$" and "$off = 0$". It is seen that A is the parent of both B and C, and C is the parent of D. D is assumed to be conditionally independent from A given C. By the given probabilities, joint probabilities can be computed from the graph. For example;

$$P(A=1, B=1, C=0, D=1) = P(A=1)P(B=1 \mid A=1)P(C=0 \mid A=1) \tag{2.6}$$
$$P(D=1 \mid C=0) = 0.60 \times 0.70 \times 0.35 \times 0.30$$



A	$P(A)$
1	0.60
0	0.40

A	$P(B=1 \mid A)$	$P(B=0 \mid A)$
1	0.70	0.30
0	0.50	0.50

A	$P(C=1 \mid A)$	$P(C=0 \mid A)$
1	0.75	0.35
0	0.95	0.05

C	$P(D=1 \mid C)$	$P(D=0 \mid C)$
1	0.80	0.20
0	0.30	0.70

Figure 2.4: An example for Bayesian networks.

Inferring Bayesian networks consists of two parts. In the first part, the best graph G is found given the observed data D. In the second part, the best conditional probabilities are obtained given the graph G and observed data D.

Model structure G is sampled from the posterior probability of a network topology given by

$$p(G \mid D) = \frac{p(D \mid G)p(G)}{p(D)} \tag{2.7}$$

where $P(G)$ is the prior distribution over network structures. Here conditional distribution $p(D \mid G)$ can be computed as follows:

$$p(D \mid G) = \int p(D \mid q, G)p(q \mid G)dq \tag{2.8}$$

in which q is the parameter vector for the conditional distributions p, $p(D \mid q,G)$ is the likelihood function and $p(q \mid G)$ is the prior distribution of the parameters.

If the structure of graph G and observed data D are assumed to be given, then the details of the conditional distributions can be enhanced by obtaining the values of parameters of the conditional distributions. The posterior distribution of the parameters q is given by

$$p(q \mid D,G) = \frac{p(D \mid q,G)p(q \mid G)}{p(D \mid G)}. \tag{2.9}$$

Bayesian network modelling is a fascinating method to construct gene regulatory networks since they are stochastic and thus they can deal with the noisy measurements.

2.5.4 Ordinary Differential Equations

Unlike Bayesian networks, ordinary differential equations (ODEs) provide a deterministic aspect in the reconstruction of gene regulatory networks. Using ordinary differential equations, concentrations of RNAs, proteins and other cellular molecules can be modelled by a discrete or continuous time-dependent variable.

Changes in the expression level of a gene at a particular time is explained by a rate equation which has the mathematical form

$$\frac{dx_i}{dt} = f_i(x_1, x_2, ..., x_n, p, u) \tag{2.10}$$

in which x_i $(1 \le i \le n)$ is the expression level of gene *i* at time *t*, *n* is the number of genes, *p* is the parameter set of the network and *u* is the external perturbation to the network. The function f_i can be linear, piecewise linear or nonlinear.

ODEs have been used firstly by Chen et al. (1999) to model gene regulation. They form the gene network by a simple linear function

$$f(x(t)) = Ax(t) \tag{2.11}$$

where *A* is $n \times n$ matrix of elements a_{ij} defining the regulatory relation between gene *i* and gene *j*.

The most widely used class of the ODEs is S-systems. They have been used to reconstruct gene regulatory networks by Kikuchi et al. (2003). S-systems are described as follows:

$$\frac{dx_i(t)}{dt} = \alpha_i \prod x_j(t)^{g_{ij}} - \beta_i \prod x_j(t)^{h_{ij}} \tag{2.12}$$

Here g_{ij} and h_{ij} are the kinetic orders and α_i and β_i are the rate constants. The first and second terms at the right hand side describe the effects of activators and inhibitors, respectively.

In Figure 2.5, a linear additive model is given as an example of the linear ordinary equations. Nodes represent the expression levels of genes, dashed lines represent the inhibiting relations, full lines represent the activating relations and q_{ij} represents the strength of effect of gene *i* on gene *j*. For example, the expression level of gene *A* at time *t+1* can be obtained by the equation

$$x_A(t+1) = x_A(t) + q_{CA}x_C(t) + q_{DA}x_D(t). \qquad (2.13)$$

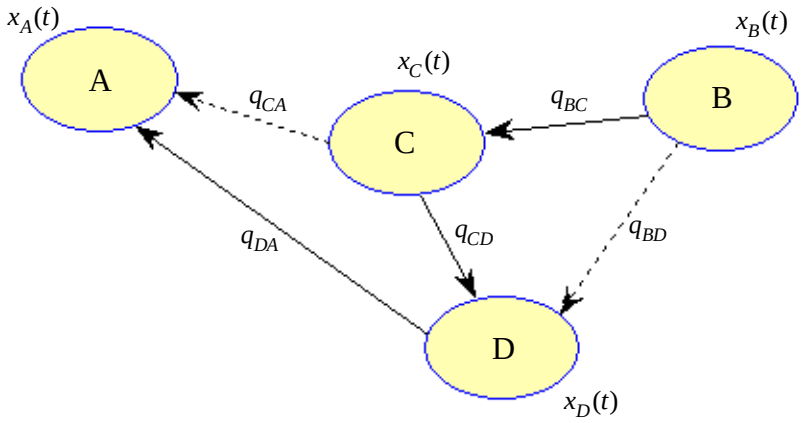


Figure 2.5: An example for a linear additive model.

2.5.5 Network Identification by Multiple Linear Regression

Gardner et al. (2003) have developed an algorithm called Network Identification by Multiple Linear Regression (NIR) to infer functional relationships between the genes. This algorithm models the behavior of a gene regulatory network by first-order linear equations describing the rate of accumulation of each network species resulting from a transcriptional perturbation:

$$d\mathbf{x}/dt = \mathbf{A}\mathbf{x} + \mathbf{u} \qquad (2.14)$$

where $\mathbf{x}$ represents the mRNA concentrations of genes, $d\mathbf{x}/dt$ represents the rate of accumulation of mRNA concentrations, $\mathbf{A}$ represents the network model describing regulatory relations between mRNA concentrations and $\mathbf{u}$ represents the set of external perturbations.

For each gene in the network, Equation (2.14) can be written in the following form:

$$\frac{dx_{il}}{dt} = \sum_{j=1}^{N} a_{ij} x_{jl} + u_{il}, \quad i = 1,...,N, \; l = 1,...,M, \qquad (2.15)$$

$$= \underline{a}_i^T \underline{x}_l + u_{il}$$

where x_{il} is the mRNA concentration of gene *i* following the perturbation in experiment *l*; a_{ij} represents the influence of gene *j* on gene *i;* and u_{il} is an external perturbation to the expression of gene *i* in experiment *l*. By using matrix notation, the rate of accumulation for all *N* genes in the network is given by

$$\frac{d\underline{x}_l}{dt} = \mathbf{A}.\underline{x}_l + \underline{u}_l, \quad l = 1,...,M, \qquad (2.16)$$

where $\underline{x}_l$ is an $N \times 1$ vector of mRNA concentrations of the *N* genes in experiment *l*, **A** is an $N \times N$ connectivity matrix, composed of elements a_{ij}, and $\underline{u}_l$ is an $N \times 1$ vector of the perturbations applied to each of the *N* genes in experiment *l*.

Near a steady-state point which means that gene expression does not change substantially over time ($\frac{d\underline{x}_l}{dt} = 0$), the following equation is obtained:

$$\mathbf{A}.\mathbf{X} = -\mathbf{U}, \qquad (2.17)$$

where **X** is an $N \times M$ matrix composed of columns $\underline{x}_l$; **U** is an $N \times M$ with each column, $\underline{u}_l$.

Equation (2.17) can be solved only if $M \geq N$. To satisfy this condition, number of experiments can be increased, but then, **A** will be extremely sensitive to noise in the perturbations and be unreliable. To overcome this problem, it is assumed that the maximum number of regulators of each gene, *k*, is less than *M*, i.e., the network is not fully connected.

Since the data in both **X** and **U** are noisy, two noise terms are added to Equation (2.17) and the following multiple linear regression model is obtained for each possible combination of k out of N weights:

$$\underline{y}_i^T = \underline{b}_i^T . \mathbf{Z} + \underline{e}_i^T \tag{2.18}$$

where $\underline{y}_i$ is an $M \times 1$ vector of measurements of $y_{il} = -u_{il}$, $\underline{b}_i$ is a $k \times 1$ vector representing one of (*N* choose *k*) possible combinations of the elements of $\underline{a}_i$, **Z** is a $K \times M$ matrix, where each column is the vector $\underline{z}_l$ for one of the M experiments ($\underline{z}_l = \underline{x}_l + \underline{\gamma}_l$, $\underline{\gamma}_l$ represents normally distributed measurement noise on the mRNA concentrations in experiment *l*); and $\underline{e}_i$ is an $M \times 1$ vector of noise ($e_{il} = \varepsilon_{il} - \underline{b}_i^T \underline{\gamma}_l$, ε_{il} represents normally distributed measurement noise on perturbations of gene *i* in experiment *l*).

By applying method of LS, NIR algorithm obtains the estimator of $\underline{b}_i$ in the multiple linear regression model given by Equation (2.18) as follows:

$$\underline{\tilde{b}}_i = (\mathbf{Z}.\mathbf{Z}^{\mathbf{T}})^{-1}.\mathbf{Z}.\underline{y}_i \tag{2.19}$$

To obtain the best estimate for $\underline{b}_i$, $\underline{\tilde{b}}_i$ is calculated for each of the (*N* choose *k*) combinations of weights for gene *i* and the estimate $\underline{\tilde{b}}_i$ with the smallest sum of

squared errors is selected as the best approximation of a_i in Equation (2.15). Sum of squared errors (SSE) lost function is defined by

$$SSE_i^k = \sum_{l=1}^{M} (y_{il} - \tilde{y}_{il})^2 = \sum_{l=1}^{M} (y_{il} - \underline{\tilde{b}}_i^T . z_l)^2 . \qquad (2.20)$$

From Equation (2.19), the predictor for $\underline{y}_i$ given the data matrix **Z** is

$$\underline{\tilde{y}}_i^T == \underline{\tilde{b}}_i^T . \mathbf{Z} . \qquad (2.21)$$

Gardner et al. (2003) assumes that the noise term in the model given by Equation (2.18) is normally distributed and the least square estimators are the most efficient for normal data. They also assume that the regressors are uncorrelated and propose to use ridge regression when some of the regressors are collinear. Furthermore, they states that $\underline{\tilde{b}}_i$ is not the maximum likelihood estimator of the model parameter $\underline{b}_i$.

CHAPTER 3

METHODOLOGY FOR GENE REGULATORY NETWORKS BY MULTIPLE LINEAR REGRESSION ANALYSIS UNDER WEIBULL DISTRIBUTION

In this chapter, the algorithm of network identification by multiple linear regression developed by Gardner et al. (2003) is considered and it is aimed to improve this algorithm by handling the normality assumption needed to apply the algorithm.

For the reconstruction of gene regulatory networks, NIR algorithm forms a multiple linear regression model for each gene in the network as follows:

$$y_i = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + ... + \gamma_q x_{iq} + e_i,\ i = 1,..,n \qquad (3.1)$$

in which $y_i = -u_i$ represents the external perturbation to the expression level of a particular gene in experiment i, x_{ij} $(i = 1,2,..,n;\ j = 1,2,...,q)$ represents the expression level (with noise) of gene j following the perturbation in experiment i, γ_j $(j = 1,2,...,q)$ represents the effect of gene j on a particular gene and e_i represents the error term.

In the study of Gardner et al. (2003), NIR algorithm is applied to a nine transcript subnetwork of the SOS pathway regulating the cell survival and repairing after DNA damage in Escherichia coli. This subnetwork is chosen to include the genes called *lexA*, *recA*, *ssb*, *recF*, *din*, *umuDC*, *rpoD*, *rpoH* and *rpoS*. It is known that

the genes *lexA* and *recA* regulate more than 30 genes directly and they are the principle mediators of the SOS response. Also, the genes *ssb*, *recF*, *din* and *umuDC* are the other regulatory genes with known involvement in SOS response. However, regulatory role of the genes *rpoD, rpoH* and *rpoS* in SOS response are not completely known.

They apply nine external perturbations to the network by overexpressing different one of the genes in each perturbation. Then they form a multiple linear regression model for each of nine genes and obtain the regulatory connections in the network assuming that the networks are not fully connected.

Table 3.1 gives the expression data for gene *i* in all perturbation experiments. Expression levels are the RNA expression ratios between perturbed and control groups. They are obtained by the formula $x_i = [RNA_i]^{pert} / [RNA_i]^{cont} - 1$ where $[RNA_i]^{pert}$ and $[RNA_i]^{cont}$ represents the expression level of perturbed and control groups for gene *i*, respectively. Table 3.2 gives the external perturbation for each gene. Relative magnitude of a perturbation to gene *i* is computed by the formula $u_i = [RNA_i]^{vec} / [RNA_i]^{cont}$ where $[RNA_i]^{vec}$ is the the concentration of gene *i* RNA synthesized from the overexpression vector pBADX53 in each training experiment.

Table 3.1: Expression data

	Genes								
Training Perturbations	*recA*	*lexA*	*ssb*	*recF*	*dinI*	*umuDC*	*rpoD*	*rpoH*	*rpoS*
recA	0.906	0.212	0.018	0.104	0.119	0.076	-0.122	0.178	0.072
lexA	-0.132	0.383	-0.107	-0.050	-0.097	-0.189	-0.047	-0.183	-0.128
ssb	-0.139	-0.117	10.524	-0.273	0.056	-0.124	-0.102	0.036	0.073
recF	0.187	0.064	0.061	0.139	0.315	0.250	-0.107	-0.070	0.081
dinI	0.291	0.169	0.080	0.180	2.147	0.347	-0.011	-0.034	0.305
umuDC	-0.061	-0.087	0.013	0.146	0.142	2.017	0.104	-0.155	0.051
rpoD	-0.077	0.039	0.064	0.069	-0.068	-0.067	3.068	0.008	-0.061
rpoH	-0.017	0.125	0.089	-0.004	0.135	-0.172	0.365	26.633	0.274
rpoS	-0.025	0.084	-0.07	0.275	0.113	-0.022	0.217	0.087	0.672

Table 3.2: Gene perturbed in training perturbations

recA	*lexA*	*ssb*	*recF*	*dinI*	*umuDC*	*rpoD*	*rpoH*	*rpoS*
0.6529	1.1711	13.412	1.6705	4.5415	2.3555	4.7083	12.8658	4.1089

NIR algorithm assumes that the error term in the model (3.1) is normally distributed with zero mean and hence the method of least squares gives the most efficient estimators. However, the random errors may not have a normal distribution and under non-normality LS estimators are known to be neither efficient nor robust. Therefore, it is very important to develop statistical procedures which are efficient and robust to deviations from an assumed distribution.

In the literature, there are several robust estimation methods such as Huber's and Tukey's estimation and MML estimation. To provide robustness against non-normality, here it is proposed to use the MML estimation procedure in multiple linear regression analysis for inferring GRN when the errors have a Weibull distribution. The first reason for preferring the MML estimators is that they have explicit forms and they are easily computed. The second one is that Huber's and Tukey's estimators are robust only for the long-tailed distributions.

In this study, instead of assuming that errors come from a normal distribution, the distribution of errors is determined. To decide their distribution, the best multiple linear regression model for each gene using the expression data given in Table 3.1 is obtained and the residuals are computed by using the method of least squares. Then, their skewness and kurtosis are matched with the theoretical values of different distributions. Also, the Q-Q plots of the obtained residuals for different distributions are examined and it is seen that the distribution of residuals fits a Weibull distribution better.

To show that the use of MML estimation procedure in multiple linear regression analysis leads to an improvement to NIR algorithm, parameters of multiple linear regression model are also estimated by using LS estimation method in this study.

3.1 Least Squares Estimation under Weibull Distribution

A family of Weibull distributions with shape parameter *p* is given by

$$f(x;p,\sigma)=\frac{p}{\sigma^{p}}x^{p-1}exp\left\{-\left(\frac{x}{\sigma}\right)^{p}\right\},\quad 0<x<\infty. \tag{3.2}$$

When $p=1$, the pdf becomes

$$f(x;\sigma)=\frac{1}{\sigma}exp\left(-\frac{x}{\sigma}\right),\quad 0<x<\infty \tag{3.3}$$

which is the pdf of exponential distribution with parameter σ.

The mean and variance, respectively, are

$$E(x)=\Gamma(1+1/p)\sigma, \tag{3.4}$$

$$Var(x)=(\Gamma(1+2/p)-\Gamma^{2}(1+1/p))\sigma^{2}\ .$$

The cumulative distribution function is given by

$$F(x;p,\sigma)=1-exp\left(-\left(\frac{x}{\sigma}\right)^{p}\right),\quad 0<x<\infty. \tag{3.5}$$

The Weibull family (3.2) represents a wide variety of skewed distributions, both with kurtosis greater than as well as less than 3.

To get the information about the nature of the Weibull distribution, its skewness and kurtosis values for particular shape parameters are given by Table 3.3.

Table 3.3: Skewness and kurtosis values of Weibull distribution

p	1.5	2	2.5	3	4	6	8
Skewness	1.064	0.631	0.358	0.168	-0.087	-0.373	-0.534
Kurtosis	4.365	3.246	2.858	2.705	2.752	3.036	3.328

In this thesis, for the multiple linear regression model given in Equation (3.1), it is assumed that e_i have a Weibull distribution with shape parameter p (>0). LS estimators of γ_j $(j=1,2,...,q)$ are same as those in the ordinary least squares approach. However, LS estimators of γ_0 and σ need to be bias corrected under Weibull distribution. Bias corrected LS estimators of γ_0 and σ are given by

$$\tilde{\sigma} = \frac{\sqrt{\sum_{i=1}^{n}\left(Y_i - \sum_{\mathrm{j}=1}^{q}\tilde{\gamma}_j X_{\mathrm{ij}}\right)^2}}{\sqrt{(n-q-1)(\Gamma(1+2/p)-\Gamma^2(1+1/p))}}\ , \qquad (3.6)$$

$$\tilde{\gamma}_0 = \bar{y} - \sum_{i=1}^{q}\tilde{\gamma}_j \bar{x}_j - \Gamma(1+1/p)\tilde{\sigma}$$

where

$$X_{ij} = x_{ij} - \bar{x}_j, \quad \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad \text{and}$$

$$Y_i = y_i - \bar{y}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i .$$

Hence, the estimated variance-covariance matrix of LS estimator $\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_1, \tilde{\gamma}_2, ..., \tilde{\gamma}_q)$ is obtained as follows:

$$Cov(\tilde{\boldsymbol{\gamma}}) = (\mathbf{X}'\mathbf{X})^{-1}\tilde{\sigma}^2, \quad \mathbf{X}_{(n\times q)} = (x_{ij}) \tag{3.7}$$

To test the equality of model parameters, the hypothesis

$$H_0 : \gamma_1 = \gamma_2 = ... = \gamma_q \quad versus \tag{3.8}$$
$$H_1 : At\ least\ one\ of\ them\ different\ from\ others,$$

is used.

The test statistic F based on LS estimators is defined as follows:

$$F = \frac{\tilde{\boldsymbol{\gamma}}'\mathbf{X}'\mathbf{y}}{q\tilde{\sigma}^2}, \quad \mathbf{X}_{(n\times q)} = (x_{ij}), \quad \mathbf{y}_{(n\times 1)} = (y_i). \tag{3.9}$$

Under the null hypothesis, F has an F-distribution with degrees of freedom q and $(n-q-1)$. If the null hypothesis is rejected, then individual parameters are tested:

$$H_0 : \gamma_j = 0 \quad (j = 1,2,..,q) \quad versus \tag{3.10}$$
$$H_1 : \gamma_j \neq 0$$

For the hypotheses given above, the test statistic based on the LS estimators is given by

$$T_j = \frac{\tilde{\gamma}_j}{S(\tilde{\gamma}_j)} \tag{3.11}$$

where $S(\tilde{\gamma}_j)$ is the standard error of $\tilde{\gamma}_j$. For $n \le 20$, T_j has a *t*-distribution with degrees of freedom $(n-q-1)$. However, for $n > 20$, the null distribution of T_j is *N(0 ,1)* and large values of T_j lead to the rejection of the null hypothesis (Islam et al., 2001).

The variances of LS estimators of model parameters in multiple linear regression analysis are very sensitive to the location and scale of the explanatory variables and to data anomalies (outliers). To rectify this problem, the re-parameterized model given by (Akkaya and Tiku, 2008) is also considered in this study:

$$y_i = \gamma_0 + \sum_{j=1}^{q} \gamma_j u_{ij} + e_i, \;\; i = 1,..,n, \;\; j = 1,...,q \tag{3.12}$$

where

$$u_{ij} = \frac{(x_{ij} - \bar{x}_j)}{s_j}, \quad \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad \text{and} \quad s_j^2 = \frac{1}{n}\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 .$$

For the re-parameterized model, LS estimators of γ_j $(1 \le j \le q)$ are given by

$$\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_j) = (\mathbf{U}'\mathbf{U})^{-1}(\mathbf{U}'\mathbf{y}), \;\; \mathbf{U}_{(n\times q)} = (u_{ij}), \;\; \mathbf{y}_{(n\times 1)} = (y_i). \tag{3.13}$$

Bias corrected LS estimators of γ_0 and σ are given by

$$\tilde{\sigma} = \frac{\sqrt{\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{q}\tilde{\gamma}_j U_{ij}\right)^2}}{\sqrt{(n-q-1)(\Gamma(1+2/p)-\Gamma^2(1+1/p))}}\ , \qquad (3.14)$$

$$\tilde{\gamma}_0 = \bar{y} - \sum_{i=1}^{q}\tilde{\gamma}_j \bar{u}_j - \Gamma(1+1/p)\tilde{\sigma}$$

where

$$U_{ij} = u_{ij} - \bar{u}_j\,, \quad \bar{u}_j = \frac{1}{n}\sum_{i=1}^{n} u_{ij} \quad \text{and}$$

$$Y_i = y_i - \bar{y}\,, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,.$$

3.2 Modified Maximum Likelihood Estimation under Weibull Distribution

Under the assumption of Weibull distribution for error terms e_i, the Fisher likelihood function for each gene is

$$L = \left(\frac{p}{\sigma}\right)^n \prod_{l=1}^{n} z_i^{p-1}\, exp(-\sum_{i=1}^{n} z_i^p)\,, \quad i = 1,..,n \qquad (3.15)$$

where

$$z_i = \frac{e_i}{\sigma} = \frac{y_i - \gamma_0 - \gamma_1 x_{i1} - \gamma_2 x_{i2} - ... - \gamma_q x_{iq}}{\sigma}.$$

The likelihood equations for estimating γ_0, $\gamma_j, (j = 1,2,..,q)$ and σ are

$$\frac{\partial \ln L}{\partial \gamma_0} = -\frac{p-1}{\sigma}\sum_{i=1}^{n} g_1(z_i) + \frac{p}{\sigma}\sum_{i=1}^{n} g_2(z_i) = 0, \tag{3.16}$$

$$\frac{\partial \ln L}{\partial \gamma_j} = -\frac{p-1}{\sigma}\sum_{i=1}^{n} g_1(z_i)x_{ij} + \frac{p}{\sigma}\sum_{i=1}^{n} g_2(z_i)x_{ij} = 0, \quad j = 1,2,..,q,$$

and

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} - \frac{p-1}{\sigma}\sum_{i=1}^{n} g_1(z_i)z_i + \frac{p}{\sigma}\sum_{i=1}^{n} g_2(z_i)z_i = 0$$

where $g_1(z_i) = z_i^{-1}$ and $g_2(z_i) = z_i^{p-1}$.

These equations have no explicit solutions since they include non-linear functions $g_1(z_i)$ and $g_2(z_i)$. They can be solved by using some iterative methods, however, it is enormously problematic to obtain the solutions by iteration since the iterations may never converge or converge to wrong values (Puthenpura and Sinha, 1986; Akkaya and Tiku, 2008a; Islam and Tiku, 2004). Moreover, there are too many equations to iterate simultaneously which is formidable task. Also, it is difficult to make any analytical study of the resulting maximum likelihood estimators, especially for small samples. Therefore, the method of modified maximum likelihood developed by Tiku (1967) is proposed to obtain the explicit solutions for the non-linear equations.

MML estimation procedure has very good statistical properties which are listed below:

1. MML estimators are the explicit functions of sample observations. Thus, they are computed very easily and it is simple to determine their properties (Vaughan and Tiku, 2000).

2. MML estimators are asymptotically equivalent to the ML estimators when some regulatory conditions hold. This means that MML estimators are asymptotically fully efficient, i.e., they are unbiased and their variances are equal to the minimum variance bound (MVB) (Bhattacharyya, 1985; Tiku et al., 1986 and Vaughan and Tiku, 2000).

3. MML estimators are almost fully efficient which means they have no or negligible bias and their variances are only marginally bigger than the MVBs even for small samples (Smith et al., 1973; Lee et al., 1980; Tan, 1985 and Tiku et al., 1986).

4. MML method is essentially self-censoring, since it assigns small weights to extremes.

In this method, likelihood equations given in Equation (3.16) are written in terms of the order statistics since the complete sums are invariant to the ordering. Let $z_{(1)} \le z_{(2)} \le ... \le z_{(n)}$ be the order statistics for $z_i \ \ (i = 1,2,...,n)$. Then, the ordered $z_i \ \ (i = 1,2,...,n)$ variates are obtained as follows:

$$z_{(i)} = \frac{e_{(i)}}{\sigma} = \frac{y_{[i]} - \gamma_0 - \gamma_1 x_{[i]1} - \gamma_2 x_{[i]2} - ... - \gamma_q x_{[i]q}}{\sigma} \tag{3.17}$$

in which $(y_{[i]}, x_{[i]j})$ are concomitants of $z_{(i)}$.

The resulting likelihood equations are given by

$$\frac{\partial \ln L}{\partial \gamma_0} = -\frac{p-1}{\sigma}\sum_{i=1}^{n} g_1\left(z_{(i)}\right) + \frac{p}{\sigma}\sum_{i=1}^{n} g_2\left(z_{(i)}\right) = 0, \tag{3.18}$$

$$\frac{\partial \ln L}{\partial \gamma_j} = -\frac{p-1}{\sigma}\sum_{i=1}^{n} g_1\left(z_{(i)}\right)x_{[i]j} + \frac{p}{\sigma}\sum_{i=1}^{n} g_2\left(z_{(i)}\right)x_{[i]j} = 0, \quad j = 1,2,..,q$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} - \frac{p-1}{\sigma}\sum_{i=1}^{n} g_1\left(z_{(i)}\right)z_{(i)} + \frac{p}{\sigma}\sum_{i=1}^{n} g_2\left(z_{(i)}\right)z_{(i)} = 0 .$$

Nonlinear terms in the likelihood equations given in (3.18) can be linearized by using the first two terms of Taylor series expansion of $g_1(z_i)$ and $g_2(z_i)$ around $t_{(i)} = E(z_{(i)})$ as follows:

$$g_1(z_{(i)}) \cong \alpha_{1i} - \beta_{1i} z_{(i)} \quad i = 1,2,...,n , \tag{3.19}$$

$$g_2(z_{(i)}) \cong \alpha_{2i} + \beta_{2i} z_{(i)}$$

where

$$\alpha_{1i} = 2t_{(i)}^{-1} \quad \text{and} \quad \alpha_{2i} = (2-p)t_{(i)}^{p-1},$$

$$\beta_{1\mathrm{i}} = t_{(i)}^{-2} \quad \text{and} \quad \beta_{2\mathrm{i}} = (p-1)t_{(i)}^{p-2} .$$

The exact values of $t_{(i)}$ are given by Harter (1964). However, for $\mathrm{n} \geq 10$, the approximated values of $t_{(i)}$ are given by

$$\int_0^{t_{(i)}} pz^{p-1}\exp(-z^p)dz = \frac{i}{n+1}, \quad i = 1,2,...,n ,$$

$$t_{(i)} = \left[-ln\left\{1 - \frac{i}{n+1}\right\}\right]^{1/p} . \tag{3.20}$$

Replacing nonlinear terms in likelihood equations by their linear approximations, the modified likelihood equations are obtained:

$$\frac{\partial ln\ L}{\partial \gamma_0} \cong \frac{\partial ln\, L^*}{\partial \gamma_0} = -\frac{p-1}{\sigma}\sum_{i=1}^{n}(\alpha_{1i} - \beta_{1i} z_{(i)}) + \frac{p}{\sigma}\sum_{i=1}^{n}(\alpha_{2i} + \beta_{2i} z_{(i)}) = 0\,, \qquad (3.21)$$

$$\frac{\partial ln\ L}{\partial \gamma_j} \cong \frac{\partial ln\, L^*}{\partial \gamma_j} = -\frac{p-1}{\sigma}\sum_{i=1}^{n}(\alpha_{1i} - \beta_{1i} z_{(i)}) x_{[i]j} + \frac{p}{\sigma}\sum_{i=1}^{n}(\alpha_{2i} + \beta_{2i} z_{(i)}) x_{[i]j} = 0\,,$$

$$\frac{\partial ln\ L}{\partial \sigma} \cong \frac{\partial ln\, L^*}{\partial \sigma} = -\frac{n}{\sigma} - \frac{p-1}{\sigma}\sum_{i=1}^{n}(\alpha_{1i} - \beta_{1i} z_{(i)}) z_{(i)} + \frac{p}{\sigma}\sum_{i=1}^{n}(\alpha_{2i} + \beta_{2i} z_{(i)}) z_{(i)} = 0\,.$$

The solutions of these equations give the MML estimators of γ_0, $\gamma_j, (j = 1,2,..,q)$ and σ:

$$\hat{\gamma}_0 = \bar{y}_{[.]} - \hat{\gamma}_1 \bar{x}_{[.]1} - ... - \hat{\gamma}_q \bar{x}_{[.]q} - \frac{\Delta}{m}\hat{\sigma}\,, \qquad (3.22)$$

$$\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_j) = \mathbf{K} - \mathbf{D}\hat{\sigma} \quad \text{and}$$

$$\hat{\sigma} = \frac{-B + \sqrt{B^2 + 4nC}}{2\sqrt{n(n-q-1)}}$$

where

$$\mathbf{K}_{q\times 1} = (K_j) = (\mathbf{X}'\boldsymbol{\Gamma}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Gamma}\mathbf{Y})\,,$$

$$\mathbf{D}_{q\times 1} = (D_j) = (\mathbf{X}'\boldsymbol{\Gamma}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Delta}\mathbf{1}_n)\,,$$

$$\mathbf{Y} = \begin{bmatrix} Y_{[1]} \\ Y_{[2]} \\ \vdots \\ Y_{[n]} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} X_{[1]1} & X_{[1]2} & \dots & X_{[1]q} \\ X_{[2]1} & X_{[2]2} & \dots & X_{[2]q} \\ \vdots & \vdots & \vdots & \vdots \\ X_{[n]1} & X_{[n]n} & \dots & X_{[n]q} \end{bmatrix},$$

$$\mathbf{\Gamma} = \begin{bmatrix} \delta_1 & 0 & \dots & 0 \\ 0 & \delta_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \delta_n \end{bmatrix}, \quad \mathbf{\Delta} = \begin{bmatrix} \Delta_1 & 0 & \dots & 0 \\ 0 & \Delta_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \Delta_n \end{bmatrix}, \quad \mathbf{1}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

$$\delta_i = (p-1)\beta_{1i} + p\beta_{2i} \ , \quad \Delta_i = (p-1)\alpha_{1i} - p\alpha_{2i},$$

$$Y_{[i]} = y_{[i]} - \bar{y}_{[.]}, \quad X_{[i]j} = x_{[i]j} - \bar{x}_{[.]j}$$

$$\bar{y}_{[.]} = \frac{1}{m}\sum_{i=1}^{n} \delta_i y_{[i]} \ , \quad \bar{x}_{[.]j} = \frac{1}{m}\sum_{i=1}^{n} \delta_i x_{[i]j} \ , \quad m = \sum_{i=1}^{n} \delta_i \ , \quad \Delta = \sum_{i=1}^{n} \Delta_i$$

$$B = \sum_{i=1}^{n} \Delta_i (Y_{[i]} - K_1 X_{[i]1} - ... - K_q X_{[i]q}),$$

$$C = \sum_{i=1}^{n} \delta_i (Y_{[i]} - K_1 X_{[i]1} - ... - K_q X_{[i]q})^2.$$

As mentioned above, MML estimators are fully efficient, i.e., they are unbiased and their variances are equal to MVB (Bhattacharyya, 1985; Tiku et al., 1986 and Vaughan and Tiku, 2000). Since the modified maximum likelihood equations are asymptotically equivalent to the maximum likelihood equations, asymptotic variance-covariance matrix of MML estimators are given by the inverse of Fisher information matrix. Fisher information matrix is the negative expectation of the second derivatives of log-likelihood with respect to the parameters. Hence, for the

multiple linear regression model when the errors follow a Weibull distribution, it is obtained as follows:

$$I_{(q+2)\times(q+2)} = \left(\frac{np^2}{\sigma}P\right)\times \begin{bmatrix} 1 & \bar{x}_{.1} & \cdots & \bar{x}_{.q} & Q/P \\ \bar{x}_{.1} & \sum_{i=1}^{n} x_{i1}^2/n & \cdots & \sum_{i=1}^{n} x_{i1}x_{iq}/n & \bar{x}_{.1}(Q/P) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{x}_{.q} & \sum_{i=1}^{n} x_{iq}x_{i1}/n & \cdots & \sum_{i=1}^{n} x_{iq}^2/n & \bar{x}_{.q}(Q/P) \\ Q/P & \bar{x}_{.1}(Q/P) & \cdots & \bar{x}_{.q}(Q/P) & 1/P \end{bmatrix} \quad (3.23)$$

in which

$$P = \left(1-\frac{1}{p}\right)^2 \Gamma\left(1-\frac{2}{p}\right),$$

$$Q = \Gamma\left(2-\frac{1}{p}\right) \quad \text{and}$$

$$\bar{x}_{.j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad (j = 1,2,..,q).$$

To test the hypotheses given in Equation (3.8), F^* statistic based on MML estimators is given by

$$F^* = \frac{\hat{\boldsymbol{\gamma}}'\mathbf{X}'\boldsymbol{\Gamma}\mathbf{Y}}{q\hat{\sigma}^2} \ . \quad (3.24)$$

Under the null hypothesis, F^* is referred to the F-distribution with degrees of freedom q and $(n-q-1)$. If the null hypothesis is rejected, then individual parameters are tested:

$$H_0 : \gamma_j = 0 \quad (j = 1,2,..,q) \quad versus \qquad (3.25)$$
$$H_1 : \gamma_j \neq 0$$

For the hypotheses given above, the test statistic based on the MML estimators is given as follows:

$$T_j^* = \frac{\hat{\gamma}_j}{S(\hat{\gamma}_j)} \qquad (3.26)$$

where $S(\hat{\gamma}_j)$ is the standard error of $\hat{\gamma}_j$. For $n \leq 20$, T_j^* has a t-distribution with degrees of freedom $(n-q-1)$. However, for $n > 20$, the null distribution of T_j^* is $N(0,1)$ and large values of T_j^* lead to the rejection of the null hypothesis (Islam et al., 2001).

For the re-parameterized model given in Equation (3.12), MML estimators of γ_0, $\gamma_j, (j = 1,2,..,q)$ and σ are:

$$\hat{\gamma}_0 = \bar{y}_{[.]} - \hat{\gamma}_1 \bar{u}_{[.]1} - ... - \hat{\gamma}_q \bar{u}_{[.]q} - \frac{\Delta}{m}\hat{\sigma}\,, \qquad (3.27)$$

$$\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_j) = \mathbf{K} - \mathbf{D}\hat{\sigma} \quad \text{and}$$

$$\hat{\sigma} = \frac{-B + \sqrt{B^2 + 4nC}}{2\sqrt{n(n-q-1)}}$$

where

$$\mathbf{K}_{q\times 1} = (K_j) = (\mathbf{U}'\mathbf{\Gamma}\mathbf{U})^{-1}(\mathbf{U}'\mathbf{\Gamma}\mathbf{Y}),$$

$$\mathbf{D}_{q\times 1} = (D_j) = (\mathbf{U}'\mathbf{\Gamma}\mathbf{U})^{-1}(\mathbf{U}'\mathbf{\Delta}\mathbf{1}_n),$$

$$\mathbf{Y} = \begin{bmatrix} Y_{[1]} \\ Y_{[2]} \\ \vdots \\ Y_{[n]} \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} U_{[1]1} & U_{[1]2} & \dots & U_{[1]q} \\ U_{[2]1} & U_{[2]2} & \dots & U_{[2]q} \\ \vdots & \vdots & \vdots & \vdots \\ U_{[n]1} & U_{[n]n} & \dots & U_{[n]q} \end{bmatrix},$$

$$\mathbf{\Gamma} = \begin{bmatrix} \delta_1 & 0 & \dots & 0 \\ 0 & \delta_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \delta_n \end{bmatrix}, \quad \mathbf{\Delta} = \begin{bmatrix} \Delta_1 & 0 & \dots & 0 \\ 0 & \Delta_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \Delta_n \end{bmatrix}, \quad \mathbf{1}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

$$\delta_i = (p-1)\beta_{1i} + p\beta_{2i} \ , \quad \Delta_i = (p-1)\alpha_{1i} - p\alpha_{2i},$$

$$Y_{[i]} = y_{[i]} - \bar{y}_{[.]}, \quad U_{[i]j} = u_{[i]j} - \bar{u}_{[.]j}$$

$$\bar{y}_{[.]} = \frac{1}{m}\sum_{i=1}^{n}\delta_i y_{[i]}, \quad \bar{u}_{[.]j} = \frac{1}{m}\sum_{i=1}^{n}\delta_i u_{[i]j}, \quad m = \sum_{i=1}^{n}\delta_i, \quad \Delta = \sum_{i=1}^{n}\Delta_i$$

$$B = \sum_{i=1}^{n}\Delta_i(Y_{[i]} - K_1U_{[i]1} - \ldots - K_qU_{[i]q}),$$

$$C = \sum_{i=1}^{n}\delta_i(Y_{[i]} - K_1U_{[i]1} - \ldots - K_qU_{[i]q})^2.$$

CHAPTER 4

METHODOLOGY FOR GENE REGULATORY NETWORKS BY STOCHASTIC MULTIPLE LINEAR REGRESSION ANALYSIS UNDER WEIBULL DISTRIBUTION

In the multiple linear regression model constructed to infer gene regulatory networks, explanatory variables represent the gene expression levels and hence they have a distribution since they are subject to the measurement errors. When the distribution of the real expression data given in the study of Gardner et al. (2003) is examined, it is seen that the expression levels denoted by the explanatory variables in the multiple linear regression model have also a Weibull distribution.

There are numerous real-life situations in which the explanatory variables are stochastic. Vaughan and Tiku (2000) studied a simple stochastic model assuming that the explanatory variable has an extreme-value distribution while the error terms have a normal distribution. They obtained the MML estimators of model parameters based on both complete likelihood and conditional likelihood functions and showed that the estimators based on the conditional likelihood function are more efficient. Also, Sazak et al. (2006) considered a simple regression model when both the explanatory variable and errors come from a generalized logistic distribution. In addition, Oral (2006) dealt with a binary regression with a covariate having a generalized logistic distribution. Furthermore, Islam and Tiku (2010) studied the multiple linear regression model when the errors have a Student's t and covariates have a generalized logistic distribution. Morever, Tiku and Akkaya (2010) worked on the quadratic

regression with stochastic variates assuming that the errors and standardized covariate have long-tailed symmetric distribution.

Treating the explanatory variables in the multiple linear regression model as nonstochastic when, in fact, they are stochastic can yield biased and inefficient estimators (Islam and Tiku, 2010).

In this study, the explanatory variables in the multiple linear regression model are considered as stochastic since they have a Weibull distribution. To overcome the problems mentioned in the preceding paragraph, it is proposed to use stochastic multiple linear regression analysis for the reconstruction of GRNs.

As stated in Chapter 3, the variances of LS estimators of model parameters in multiple linear regression analysis are very sensitive to the location and scale of the explanatory variables and to design anomalies (outliers). To rectify this situation, the re-parameterized model given by (Akkaya and Tiku, 2008) is considered:

$$y_i = \gamma_0 + \sum_{j=1}^{q} \gamma_j u_{ij} + e_i, \quad u_{ij} = \frac{(x_{ij} - \mu_j)}{\sigma_j} \quad i = 1,..,n, \quad j = 1,...,q \tag{4.1}$$

In a multiple linear regression analysis, it is assumed that the explanatory variables are uncorrelated. However, it is known that there are regulatory relationships between some of them since they represent the gene expression levels. Therefore, the relationships between the explanatory variables are taken into account and the partial correlation coefficients between the explanatory variables are estimated by implementing method of MML in the stochastic multiple linear regression model in this study. To be able to make efficiency and robustness comparisons of parameters, the LS estimators for the stochastic multiple linear regression model are also given.

The joint pdf of the random variables $(X_1, X_2, ...X_q)$ can be written as follows:

$$f(x_1, x_2, ..., x_q) = f_1(x_1) f_2(x_2 \mid x_1) ... f_q(x_q \mid x_1, ..., x_{q-1}), \tag{4.2}$$

where $f(x_1)$ is the marginal distribution of X_1 and $f_j(x_j \mid x_1, x_2, ..., x_{j-1})$ is the conditional distribution of X_j given $X_l = x_l \ (1 \le l \le j-1)$.

For a multivariate normal distribution, the marginal and conditional distributions in Equation (4.2) are all normal and have the following means and variances (Islam and Tiku, 2010):

$$E(X_1) = \mu_1, \tag{4.3}$$

$$Var(X_1) = \sigma_1^2,$$

$$E(X_j \mid X_1 = x_1, ..., X_{j-1} = x_{j-1}) = \mu_{j.I_j} + \sum_{l=1}^{j-1} \theta_{jl} x_l \ ,$$

$$Var(X_j \mid X_1 = x_1, ..., X_{j-1} = x_{j-1}) = \sigma_j^2 \prod_{l=1}^{j-1} (1 - \rho_{jl.I_l}^2) = \sigma_{j.I_j}^2$$

where

$$\mu_{1.I_1} = \mu_1, \quad \sigma_{1.I_1}^2 = \sigma_1^2,$$

$$\mu_{j.I_j} = \mu_j - \sum_{l=1}^{j-1} \theta_{jl} \mu_l$$

$$\theta_{jl} = \frac{\sigma_j}{\sigma_l}\rho_{jl.I_l} \quad (1 \le l \le j-1).$$

Here, μ_j and σ_j are the location and scale parameter in the distribution of X_j, respectively, I_a represents the set of integers {1, 2, …, $a-1, a+1, \ldots$, j-1} and $\rho_{jl.I_l}$ is the partial correlation coefficient between X_j and X_l.

A multitude of non-normal distributions can be obtained by changing some or all distributions in Equation (4.2), specifically with the same means and variances.

In the re-parameterized model given by Equation (4.1), it is assumed that e_i have a Weibull distribution with shape parameter p and w_{ij} have a Weibull distribution with shape parameter p_j where

$$w_{ij} = \frac{1}{\sigma_{j.I_j}}\left(x_{ij} - \mu_{j.I_j} - \sum_{l=1}^{j-1}\theta_{jl}x_{il}\right), \quad i = 1,..,n; \; j = 1,...,q.$$

4.1 Least Squares Estimation for Stochastic Multiple Linear Regression

Least squares estimators of parameters for the re-parameterized multiple linear regression model with stochastic covariates are obtained by minimizing

$$\sum_{i=1}^{n}\left(e_i - E(e_i)\right)^2 \quad \text{and} \quad \sum_{i=1}^{n}\left(e_{ij}^* - E(e_{ij}^*)\right)^2 \qquad (4.4)$$

where

$$e_i = (y_i - \gamma_0 - \sum_{j=1}^{q}\gamma_j u_{ij}) \quad \text{and} \quad e_{ij}^* = x_{ij} - \mu_{j.I_j} - \sum_{l=1}^{j-1}\theta_{jl}x_{il}.$$

Since it is assumed that w_{ij} comes from a Weibull distribution, the least squares estimators of μ_j, σ_{j} and $\hat{\rho}_{\mathrm{jl.I_l}}$ have to be corrected for bias. The bias corrected LS estimators are

$$\tilde{\mu}_{\mathrm{j}} = \tilde{\mu}_{j.I_j} + \sum_{l=1}^{j-1} \tilde{\theta}_{jl} \tilde{\mu}_{l.I_l}, \quad \tilde{\mu}_1 = \tilde{\mu}_{1.I_1}, \tag{4.5}$$

$$\tilde{\sigma}_{\mathrm{j}} = \sqrt{\tilde{\sigma}^2_{j.I_j} + \sum_{l=1}^{j-1} \tilde{\theta}^2_{jl} \tilde{\sigma}^2_{l.I_l}}, \quad \tilde{\sigma}_1 = \tilde{\sigma}_{1.I_1},$$

$$\tilde{\rho}_{\mathrm{jl.I_l}} = \tilde{\theta}_{jl} \frac{\tilde{\sigma}_l}{\tilde{\sigma}_j}, \quad \tilde{\rho}_{21.I_1} = \tilde{\rho}_{21},$$

where

$$\tilde{\mu}_{j.I_j} = \bar{x}_j - \sum_{l=1}^{j-1} \tilde{\theta}_{jl} \bar{x}_l - \Gamma(1+1/p_j)\tilde{\sigma}_{j.I_j},$$

$$\tilde{\sigma}_{j.I_j} = \frac{\sqrt{\sum_{i=1}^{n} \left(X_{ij} - \sum_{l=1}^{j-1} \tilde{\theta}_{jl} X_{il} \right)^2}}{\sqrt{(n-j)\Gamma(1+2/p_j) - \Gamma^2(1+1/p_j))}},$$

$\tilde{\boldsymbol{\theta}}_{j-1} = (\mathbf{X}'_{j-1}\mathbf{X}_{j-1})^{-1}(\mathbf{X}'_{j-1}\mathbf{x}_j) = (\tilde{\theta}_{jl})$ and

$$\mathbf{X}_{j-1} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1,j-1} \\ X_{21} & X_{22} & \dots & X_{2,j-1} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{n,j-1} \end{bmatrix}, \quad \mathbf{x}_j = \begin{bmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{nj} \end{bmatrix},$$

$$X_{ij} = x_{ij} - \bar{x}_j, \quad \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}.$$

Also error term in the model is assumed to have a Weibull distribution. Bias corrected LS estimators of γ_0, γ_j $(1 \le j \le q)$ and σ are given by

$$\tilde{\gamma}_0 = \bar{y} - \sum_{j=1}^{q} \tilde{\gamma}_j \bar{u}_j - \Gamma(1+1/p)\tilde{\sigma}, \tag{4.6}$$

$$\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_j) = (\mathbf{U}'\mathbf{U})^{-1}(\mathbf{U}'\mathbf{Y}),$$

$$\tilde{\sigma} = \frac{\sqrt{\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{q} \tilde{\gamma}_j \tilde{U}_{ij}\right)^2}}{\sqrt{(n-q-1)(\Gamma(1+2/p) - \Gamma^2(1+1/p))}},$$

$$X_{ij} = x_{ij} - \bar{x}_j, \quad \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},$$

$$\bar{u}_j = \frac{1}{n}\sum_{i=1}^{n} \tilde{u}_{ij}, \quad \tilde{u}_{ij} = \frac{(x_{ij} - \tilde{\mu}_j)}{\tilde{\sigma}_j},$$

$$\mathbf{U}_{(n\times q)} = (\tilde{U}_{ij}), \quad \tilde{U}_{ij} = \tilde{u}_{ij} - \bar{u}_j,$$

$$\mathbf{Y}_{(n\times 1)} = (Y_i), \quad Y_i = y_i - \bar{y}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

4.2 Modified Maximum Likelihood Estimation for Stochastic Multiple Linear Regression

Since some of the explanatory variables are correlated in the reconstruction of gene regulatory networks by multiple linear regression model, the Fisher likelihood function should be written as in the following form:

$$L = L(z)L(x_1)L(x_2 \mid x_1)...L(x_q \mid x_1, x_2, ..., x_{q-1}) \tag{4.7}$$

$$\propto \left[\left(\frac{1}{\sigma} \right)^n \prod_{i=1}^{n} z_i^{p-1} e^{-z_i^p} \right] . \left[\prod_{j=1}^{q} \left(\left(\frac{1}{\sigma_{j.I_j}} \right)^n \prod_{i=1}^{n} w_{ij}^{p_j - 1} e^{-w_{ij}^{p_j}} \right) \right]$$

where $z_i = \frac{e_i}{\sigma} = \frac{1}{\sigma}(y_i - \gamma_0 - \sum_{j=1}^{q} \gamma_j u_{ij})$ and $w_{ij} = \frac{1}{\sigma_{j.I_j}} \left(x_{ij} - \mu_{j.I_j} - \sum_{l=1}^{j-1} \theta_{jl} x_{il} \right)$.

Likelihood equations are obtained as follows:

$$\frac{\partial \ln L}{\partial \mu_{j.I_j}} = (p_j - 1) \sum_{i=1}^{n} g_1(w_{ij}) - p_j \sum_{i=1}^{n} g_2(w_{ij}) = 0, \tag{4.8}$$

$$\frac{\partial \ln L}{\partial \theta_{jl}} = (p_j - 1) \sum_{i=1}^{n} g_1(w_{ij}) X_{[i]l} - p_j \sum_{i=1}^{n} g_2(w_{ij}) X_{il} = 0,$$

$$\frac{\partial \ln L}{\partial \sigma_{j.I_j}} = n + (p_j - 1) \sum_{i=1}^{n} g_1(w_{ij}) w_{ij} - p_j \sum_{i=1}^{n} g_2(w_{ij}) w_{ij} = 0,$$

$$\frac{\partial \ln L}{\partial \gamma_0} = (p - 1) \sum_{i=1}^{n} h_1(z_i) - p \sum_{i=1}^{n} h_2(z_i) = 0,$$

$$\frac{\partial \ln L}{\partial \gamma_j} = (p-1)\sum_{i=1}^{n} h_1(z_i)U_{ij} - p\sum_{i=1}^{n} h_2(z_i)U_{ij} = 0,\ j = 1,...,q \quad \text{and}$$

$$\frac{\partial \ln L}{\partial \sigma} = n + (p-1)\sum_{i=1}^{n} h_1(z_i)z_i - p\sum_{i=1}^{n} h_2(z_i)z_i = 0$$

where

$$g_1(w_{ij}) = w_{ij}^{-1} \text{ and } g_2(w_{ij}) = w_{ij}^{p_j - 1},$$

$$h_1(z_i) = z_i^{-1} \text{ and } h_2(z_i) = z_i^{p-1},$$

These equations have no explicit solutions since they include non-linear functions $g_1(w_{ij})$, $g_2(w_{ij})$, $h_1(z_i)$ and $h_2(z_i)$. Therefore, the method of modified maximum likelihood is used to obtain the explicit solutions for the non-linear equations given in Equation (4.8).

Likelihood equations are written in terms of the order statistics since the complete sums are invariant to the ordering. Let $z_{(1)} \le z_{(2)} \le ... \le z_{(n)}$ and $w_{(1)j} \le w_{(2)j} \le ... \le w_{(n)j}$ be the order statistics for z_i $(i = 1,2,...,n)$ and w_{ij} $(i = 1,2,...,n;\ j = 1,2,...,q)$, respectively.

The ordered z_i variates are obtained as follows:

$$z_{(i)} = \frac{e_{(i)}}{\sigma} = \frac{1}{\sigma}\left(y_{[i]} - \gamma_0 - \sum \gamma_j \frac{(x_{[i]j} - \mu_j)}{\sigma_j} \right), \qquad (4.9)$$

where $(y_{[i]}, x_{[i]1}, ..., x_{[i]q})$ are the concomitant observations associated with the $e_{(i)}$. Similarly, the ordered w_{ij} variates are obtained as follows:

$$w_{(i)j} = \frac{e^*_{(i)j}}{\sigma_{j.I_j}} = \frac{1}{\sigma_{j.I_j}}\left(x_{[i]j} - \mu_{j.I_j} - \sum_{l=1}^{j-1} \theta_{jl} x_{[i]l} \right) \tag{4.10}$$

where $(x_{[i]1}, ..., x_{[i]q})$ are the concomitant observations associated with the $e^*_{(i)j}$.

Likelihood equations in terms of ordered variates are given by

$$\frac{\partial \ln L}{\partial \mu_{j.I_j}} = (p_j - 1)\sum_{i=1}^{n} g_1(w_{(i)j}) - p_j \sum_{i=1}^{n} g_2(w_{(i)j}) = 0, \tag{4.11}$$

$$\frac{\partial \ln L}{\partial \theta_{jl}} = (p_j - 1)\sum_{i=1}^{n} g_1(w_{(i)j}) X_{[i]l} - p_j \sum_{i=1}^{n} g_2(w_{(i)j}) X_{[i]l} = 0,$$

$$\frac{\partial \ln L}{\partial \sigma_{j.I_j}} = n + (p_j - 1)\sum_{i=1}^{n} g_1(w_{(i)j}) w_{(i)j} - p_j \sum_{i=1}^{n} g_2(w_{(i)j}) w_{(i)j} = 0,$$

$$\frac{\partial \ln L}{\partial \gamma_0} = (p - 1)\sum_{i=1}^{n} h_1(z_{(i)}) - p\sum_{i=1}^{n} h_2(z_{(i)}) = 0,$$

$$\frac{\partial \ln L}{\partial \gamma_j} = (p - 1)\sum_{i=1}^{n} h_1(z_{(i)}) U_{[i]j} - p\sum_{i=1}^{n} h_2(z_{(i)}) U_{[i]j} = 0, \ j = 1, ..., q \ \text{ and}$$

$$\frac{\partial \ln L}{\partial \sigma} = n + (p - 1)\sum_{i=1}^{n} h_1(z_{(i)}) z_{(i)} - p\sum_{i=1}^{n} h_2(z_{(i)}) z_{(i)} = 0$$

where

$g_1(w_{(i)j}) = w_{(i)j}^{-1}$ and $g_2(w_{(i)j}) = w_{(i)j}^{p_j-1}$ $(j = 1,...,q)$,

$h_1(z_{(i)}) = z_{(i)}^{-1}$ and $h_2(z_{(i)}) = z_{(i)}^{p-1}$.

The nonlinear terms in the likelihood equations are linearized by using the first two terms of Taylor series expansion of $g_1(w_{(i)j})$, $g_2(w_{(i)j})$, $h_1(z_{(i)})$ and $h_2(z_{(i)})$:

$$g_1(w_{(i)j}) \cong \alpha_{1ij} - \beta_{1ij} w_{(i)j} \quad i = 1,...,n; \; j = 1,...,q \tag{4.12}$$

$g_2(w_{(i)j}) \cong \alpha_{2ij} + \beta_{2ij} w_{(i)j}$,

$h_1(z_{(i)}) \cong \alpha_{1i} - \beta_{1i} z_{(i)}$ and

$h_2(z_{(i)}) \cong \alpha_{2i} + \beta_{2i} z_{(i)}$

where

$\alpha_{1ij} = 2t_{(i)j}^{-1}$, $\alpha_{2ij} = (2 - p_j) t_{(i)j}^{p_j-1}$,

$\beta_{1ij} = t_{(i)j}^{-2}$, $\beta_{2ij} = (p_j - 1) t_{(i)j}^{p_j-2}$,

$\alpha_{1i} = 2t_{(i)}^{-1}$, $\alpha_{2i} = (2 - p) t_{(i)}^{p-1}$ and

$\beta_{1i} = t_{(i)}^{-2}$, $\beta_{2i} = (p - 1) t_{(i)}^{p-2}$.

For $n \geq 10$, the approximated values of $t_{(i)}$ are given by

$$\int_0^{t_{(i)}} pz^{p-1} \exp(-z^p) dz = \frac{i}{n+1}, \quad i = 1, 2, ..., n,$$

$$t_{(i)} = \left[-ln\left\{1 - \frac{i}{n+1}\right\}\right]^{1/p}, \quad i = 1, 2, ..., n. \tag{4.13}$$

Similarly, the approximated values of $t_{(i)j}$ can be obtained from

$$t_{(i)j} = \left[-\ln\left\{1 - \frac{i}{n+1}\right\}\right]^{1/p_j}, \quad i = 1, 2, ..., n;\ j = 1, 2, ..., q. \tag{4.14}$$

Replacing nonlinear terms in the likelihood equations by their linear approximations, the modified likelihood equations are obtained:

$$\frac{\partial ln\ L}{\partial \mu_{j.I_j}} = (p_j - 1)\sum_{i=1}^{n}(\alpha_{1ij} - \beta_{1ij} w_{(i)j}) - p_j \sum_{i=1}^{n}(\alpha_{2ij} + \beta_{2ij} w_{(i)j}) = 0, \tag{4.15}$$

$$\frac{\partial ln\ L}{\partial \theta_{jl}} = (p_j - 1)\sum_{i=1}^{n}(\alpha_{1ij} - \beta_{1ij} w_{(i)j}) X_{[i]l} - p_j \sum_{i=1}^{n}(\alpha_{2ij} + \beta_{2ij} w_{(i)j}) X_{[i]l} = 0,$$

$$\frac{\partial ln\ L}{\partial \sigma_{j.I_j}} = n + (p_j - 1)\sum_{i=1}^{n}(\alpha_{1ij} - \beta_{1ij} w_{(i)j}) w_{(i)j} - p_j \sum_{i=1}^{n}(\alpha_{2ij} + \beta_{2ij} w_{(i)j}) w_{(i)j} = 0,$$

$$\frac{\partial ln\ L}{\partial \gamma_0} = (p - 1)\sum_{i=1}^{n}(\alpha_{1i} - \beta_{1i} z_{(i)}) - p\sum_{i=1}^{n}(\alpha_{2i} + \beta_{2i} z_{(i)}) = 0,$$

$$\frac{\partial \ln L}{\partial \gamma_j} = (p-1)\sum_{i=1}^{n}(\alpha_{1i} - \beta_{1i} z_{(i)})U_{[i]j} - p\sum_{i=1}^{n}(\alpha_{2i} + \beta_{2i} z_{(i)})U_{[i]j} = 0,$$

$$\frac{\partial \ln L}{\partial \sigma} = n + (p-1)\sum_{i=1}^{n}(\alpha_{1i} - \beta_{1i} z_{(i)})z_{(i)} - p\sum_{i=1}^{n}(\alpha_{2i} + \beta_{2i} z_{(i)})z_{(i)} = 0.$$

The solutions of the above equations give the MML estimators of μ_j, σ_{j} and $\hat{\rho}_{\mathrm{jl.I_l}}$:

$$\hat{\mu}_{\mathrm{j}} = \hat{\mu}_{j.I_j} + \sum_{l=1}^{j-1}\hat{\theta}_{jl}\hat{\mu}_l, \quad \hat{\mu}_1 = \hat{\mu}_{1.I_1}, \tag{4.16}$$

$$\hat{\sigma}_{\mathrm{j}} = \sqrt{\hat{\sigma}^2_{j.I_j} + \sum_{l=1}^{j-1}\hat{\theta}^2_{jl}\hat{\sigma}^2_{l.I_l}}, \ \hat{\sigma}_1 = \hat{\sigma}_{1.I_1},$$

$$\hat{\rho}_{\mathrm{jl.I_l}} = \hat{\theta}_{jl}\frac{\hat{\sigma}_l}{\hat{\sigma}_j}, \ \hat{\rho}_{21.I_1} = \hat{\rho}_{21},$$

where

$$\hat{\mu}_{j.I_j} = \bar{x}_{jj[.]} - \sum_{l=1}^{j-1}\hat{\theta}_{jl}\bar{x}_{lj[.]} - \frac{\Delta_j}{m_j}\hat{\sigma}_{j.I_j},$$

$$\Delta_{\mathrm{j}} = \sum_{i=1}^{n}\delta_{ij}, \quad \delta_{\mathrm{ij}} = (p_j - 1)\alpha_{1ij} - p_j\alpha_{2ij},$$

$$m_{\mathrm{j}} = \sum_{i=1}^{n}m_{ij}, \quad m_{\mathrm{ij}} = (p_j - 1)\beta_{1ij} + p_j\beta_{2ij},$$

$$\hat{\sigma}_{j.I_j} = \frac{-B_j + \sqrt{B_j^2 + 4nC_j}}{2n} \ ,$$

$$B_j = \sum_{i=1}^{n} \delta_{ji} \left(X_{[i]j} - \sum_{l=1}^{j-1} K_{jl} X_{[i]l} \right) \ ,$$

$$C_j = \sum_{i=1}^{n} m_{ij} \left(X_{[i]j} - \sum_{l=1}^{j-1} K_{jl} X_{[i]l} \right)^2 ,$$

$$X_{r[i]} = x_{[i]r} - \bar{x}_{rj[.]} \ \ (1 \le r \le j) \, .$$

In Equation (4.16), the expressions for $\bar{x}_{rj[.]}$, K_{jl} and $\hat{\theta}_{jl}$ $(1 \le r \le j, 1 \le l \le j-1)$ are as follows:

$$\bar{x}_{rj[.]} = \frac{1}{m_j} \sum_{i=1}^{n} m_{ij} x_{[i]r} \ , \ \ \hat{\boldsymbol{\theta}}_{j-1} = \mathbf{K}_{j-1} - \mathbf{D}_{j-1} \hat{\sigma}_{j.I_j} = \hat{\theta}_{jl} \, , \qquad (4.17)$$

$$\mathbf{K}_{j-1} = (\mathbf{X}'_{j-1} \mathbf{M}_j \mathbf{X}_{j-1})^{-1} (\mathbf{X}'_{j-1} \mathbf{M}_j \mathbf{x}_j) = (K_{jl}) ,$$

$$\mathbf{D}_{j-1} = (\mathbf{X}'_{j-1} \mathbf{M}_j \mathbf{X}_{j-1})^{-1} (\mathbf{X}'_{j-1} \boldsymbol{\delta}_j \mathbf{1}_n) = (D_{jl})$$

where

$$\mathbf{X}_{j-1} = \begin{bmatrix} X_{[1]1} & X_{[1]2} & \cdots & X_{[1]j-1} \\ X_{[2]1} & X_{[2]2} & \cdots & X_{[2]j-1} \\ \vdots & \vdots & \vdots & \vdots \\ X_{[n]1} & X_{[n]2} & \cdots & X_{[n]j-1} \end{bmatrix}, \quad \mathbf{x}_j = \begin{bmatrix} X_{[1]j} \\ X_{[2]j} \\ \vdots \\ X_{[n]j} \end{bmatrix},$$

$$\boldsymbol{\delta}_j = \begin{bmatrix} \delta_{1j} & 0 & \dots & 0 \\ 0 & \delta_{2j} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \delta_{nj} \end{bmatrix}, \quad \mathbf{M}_j = \begin{bmatrix} m_{1j} & 0 & \dots & 0 \\ 0 & m_{2j} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & m_{nj} \end{bmatrix},$$

$$\hat{\boldsymbol{\theta}}_{j-1} = \begin{bmatrix} \hat{\theta}_{j1} \\ \hat{\theta}_{j2} \\ \vdots \\ \hat{\theta}_{jj-1} \end{bmatrix}, \quad \mathbf{1}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

The MML estimators of γ_0 and γ_j $(1 \le j \le q)$ are

$$\hat{\gamma}_0 = \bar{y}_{[.]} - \sum_{j=1}^{q} \hat{\gamma}_j \bar{u}_{[.]j} - \frac{\Delta}{m}\hat{\sigma}, \tag{4.18}$$

$$\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_j) = \mathbf{K} - \mathbf{D}\hat{\sigma}$$

where

$$\Delta = \sum_{i=1}^{n} \delta_i, \quad \delta_i = (p-1)\alpha_{1i} - p\alpha_{2i},$$

$$m = \sum_{i=1}^{n} m_i, \quad m_i = (p-1)\beta_{1i} + p\beta_{2i},$$

$$\bar{u}_{[.]j} = \frac{1}{m}\sum_{i=1}^{n} m_i \hat{u}_{[i]j}, \quad \hat{u}_{[i]j} = \frac{(x_{[i]j} - \hat{\mu}_j)}{\hat{\sigma}_j},$$

$$\mathbf{K}_{q\times 1} = (\mathbf{U}'\mathbf{M}\mathbf{U})^{-1}(\mathbf{U}'\mathbf{M}\mathbf{Y}) = (K_j),$$

$$\mathbf{D}_{q\times 1} = (\mathbf{U}'\mathbf{M}\mathbf{U})^{-1}(\mathbf{U}'\boldsymbol{\delta}\mathbf{1}_n) = (D_j),$$

$$\mathbf{U}_{(n\times q)} = (\hat{U}_{[i]j}),\ \hat{U}_{[i]j} = \hat{u}_{[i]j} - \bar{u}_{[.]j},$$

$$\mathbf{Y}_{(n\times 1)} = (Y_{[i]}),\quad Y_{[i]} = y_{[i]} - \bar{y}_{[.]},\quad \bar{y}_{[.]} = \frac{1}{m}\sum_{i=1}^{n} m_i y_{[i]},$$

$$\mathbf{M}_{(n\times n)} = diag(m_i),\ \boldsymbol{\delta}_{(n\times n)} = diag(\delta_i),\ \mathbf{1}_{(n\times 1)} = (1).$$

The MML estimator of σ is

$$\hat{\sigma} = \frac{-B + \sqrt{B^2 + 4nC}}{2n}\ , \tag{4.19}$$

where

$$B = \sum_{i=1}^{n} \delta_i \left(Y_{[i]} - \sum_{j=1}^{q} K_j \hat{U}_{j[i]} \right) \text{ and}$$

$$C = \sum_{i=1}^{n} m_i \left(Y_{[i]} - \sum_{j=1}^{q} K_j \hat{U}_{j[i]} \right)^2 .$$

4.3 Hypothesis Testing for Stochastic Multiple Linear Regression

To test the hypotheses given in Equation (3.8), F and F^* statistics based on LS and MML estimators, respectively are given by

$$F = \sum_{j=1}^{q} \tilde{\gamma}_j \sum_{i=1}^{n} \frac{u_{ij} y_i}{q\tilde{\sigma}^2} \quad \text{and} \quad F^* = \sum_{j=1}^{q} \hat{\gamma}_j \sum_{i=1}^{n} \frac{m_{ij} u_{ij} y_i}{q\hat{\sigma}^2}\,. \tag{4.20}$$

True distribution of F and F^* are intractable at the present time but they have F distribution, approximately (Islam et al., 2010). Therefore, the null distributions of F and F^* are referred to the F-distribution with degrees of freedom q and $(n-q-1)$. Large values of F and F^* lead to the rejection of the null hypothesis.

In the hypothesis testing procedure for stochastic multiple linear regression analysis, the test statistics F and F^* are obtained at two stages. At the fist stage, x_{ij} is treated as nonstochastic since its distribution complicates obtaining the test statistics. At the second stage, μ_j and σ_j are estimated by using the distribution of x_{ij} and the stochasticity of x_{ij} is taken into consideration by this way.

4.4 Asymptotic Covariance Matrix for Stochastic Multiple Linear Regression

As mentioned before, modified maximum likelihood equations are asymptotically equivalent to the maximum likelihood equations. Therefore, the variance-covariance matrix of MML estimators can be obtained by taking the inverse of the Fisher information matrix. However it is very difficult to derive the elements of Fisher information matrix for $q \geq 2$. Instead, sample information matrix $\hat{I}$ can be used.

The elements of sample information are given by negative of the second derivative of log-likelihood computed at the MML estimates of the model parameters.

Inverse of $\hat{I}$ gives the approximated values of variances and covariances of MML estimators. Since it takes too much space to present the elements of $\hat{I}$ here, just some elements of $\hat{I}$ are given for $\mu_1, \sigma_1, \mu_2, \sigma_2, \rho_{12}, \gamma_0, \gamma_1, \gamma_2$ and σ when $q = 2$:

$$\hat{I}_{\mu_1\mu_1} = \frac{\hat{\gamma}_1^2}{\hat{\sigma}^2\hat{\sigma}_1^2}\left[(p-1)\sum_{i=1}^{n} h_1'(\hat{z}_i) - p\sum_{i=1}^{n} h_2'(\hat{z}_i)\right] + \tag{4.21}$$

$$\frac{1}{\hat{\sigma}_1^2}\left[(p_1-1)\sum_{i=1}^{n} g_1'(\hat{w}_{i1}) - p_1\sum_{i=1}^{n} g_2'(\hat{w}_{i1})\right] +$$

$$\frac{\hat{\rho}_{21}^2}{(1-\hat{\rho}_{21}^2)\hat{\sigma}_1^2}\left[(p_2-1)\sum_{i=1}^{n} g_1'(\hat{w}_{i2}) - p_2\sum_{i=1}^{n} g_2'(\hat{w}_{i2})\right],$$

$$\hat{I}_{\mu_1\sigma_1} = \frac{\hat{\gamma}_1^2}{\hat{\sigma}^2\hat{\sigma}_1^2}\left[(p-1)\sum_{i=1}^{n} h_1'(\hat{z}_i)\hat{u}_{i1} - p\sum_{i=1}^{n} h_2'(\hat{z}_i)\hat{u}_{i1}\right] +$$

$$\frac{1}{\hat{\sigma}_1^2}\left[(p_1-1)\sum_{i=1}^{n} g_1'(\hat{w}_{i1})\hat{u}_{i1} - p_1\sum_{i=1}^{n} g_2'(\hat{w}_{i1})\hat{u}_{i1}\right] +$$

$$\frac{\hat{\rho}_{21}^2}{(1-\hat{\rho}_{21}^2)\hat{\sigma}_1^2}\left[(p_2-1)\sum_{i=1}^{n} g_1'(\hat{w}_{i2})\hat{u}_{i1} - p_2\sum_{i=1}^{n} g_2'(\hat{w}_{i2})\hat{u}_{i1}\right],$$

$$\hat{I}_{\mu_1\mu_2} = \frac{\hat{\gamma}_1\hat{\gamma}_2}{\hat{\sigma}^2\hat{\sigma}_1\hat{\sigma}_2}\left[(p-1)\sum_{i=1}^{n} h_1'(\hat{z}_i) - p\sum_{i=1}^{n} h_2'(\hat{z}_i)\right] -$$

$$\frac{\hat{\rho}_{21}}{\hat{\sigma}_1\hat{\sigma}_2(1-\hat{\rho}_{21}^2)}\left[(p_2-1)\sum_{i=1}^{n} g_1'(\hat{w}_{i2}) - p_2\sum_{i=1}^{n} g_2'(\hat{w}_{i2})\right],$$

$$\hat{I}_{\mu_1\sigma_2} = \frac{\hat{\gamma}_1\hat{\gamma}_2}{\hat{\sigma}^2\hat{\sigma}_1\sigma_2}\left[(p-1)\sum_{i=1}^{n} h_1'(\hat{z}_i)\hat{u}_{i2} - p\sum_{i=1}^{n} h_2'(\hat{z}_i)\hat{u}_{i2}\right] -$$

$$\frac{\hat{\rho}_{21}}{\hat{\sigma}_1\hat{\sigma}_2(1-\hat{\rho}_{21}^2)}\left[(p_2-1)\sum_{i=1}^{n} g_1'(\hat{w}_{i2})\hat{u}_{i2} - p_2\sum_{i=1}^{n} g_2'(\hat{w}_{i2})\hat{u}_{i2}\right],$$

$$\hat{I}_{\mu_1\rho_{21}} = \frac{1}{\hat{\sigma}_1(1-\hat{\rho}_{21}^2)^{3/2}}\left[(p_2-1)\sum_{i=1}^{n} g_1(\hat{w}_{i2}) - p_2\sum_{i=1}^{n} g_2(\hat{w}_{i2})\right] +$$

$$\frac{\hat{\rho}_{21}^2}{\hat{\sigma}_1(1-\hat{\rho}_{21}^2)^2}\left[(p_2-1)\sum_{i=1}^{n} g_1'(\hat{w}_{i2})\hat{u}_{i2} - p_2\sum_{i=1}^{n} g_2'(\hat{w}_{i2})\hat{u}_{i2}\right],$$

$$\hat{I}_{\mu_1\gamma_0} = -\frac{\hat{\gamma}_1}{\hat{\sigma}^2\hat{\sigma}_1}\left[(p-1)\sum_{i=1}^{n} h_1'(\hat{z}_i) - p\sum_{i=1}^{n} h_2'(\hat{z}_i)\right],$$

$$\hat{I}_{\mu_1\gamma_1} = -\frac{\hat{\gamma}_1}{\hat{\sigma}^2\hat{\sigma}_1}\left[(p-1)\sum_{i=1}^{n} h_1'(\hat{z}_i)\hat{u}_{i1} - p\sum_{i=1}^{n} h_2'(\hat{z}_i)\hat{u}_{i1}\right],$$

$$\hat{I}_{\mu_1\gamma_2} = -\frac{\hat{\gamma}_1}{\hat{\sigma}^2\hat{\sigma}_1}\left[(p-1)\sum_{i=1}^{n} h_1'(\hat{z}_i)\hat{u}_{i2} - p\sum_{i=1}^{n} h_2'(\hat{z}_i)\hat{u}_{i2}\right],$$

$$\hat{I}_{\mu_1\sigma} = -\frac{\hat{\gamma}_1}{\hat{\sigma}^2\hat{\sigma}_1}\left[(p-1)\sum_{i=1}^{n} h_1'(\hat{z}_i)\hat{z}_i - p\sum_{i=1}^{n} h_2'(\hat{z}_i)\hat{z}_i\right],$$

and so on. Here,

$$\hat{u}_{ij} = \frac{x_{i1} - \hat{\mu}_j}{\hat{\sigma}_j}, \quad i = 1,2,...,n;\ j = 1,2,$$

$$\hat{z}_i = \frac{y_i - \hat{\gamma}_0 - \hat{\gamma}_1\hat{u}_{i1} - \hat{\gamma}_2\hat{u}_{i2}}{\hat{\sigma}},$$

$$\hat{w}_{i1} = \hat{u}_{i1},$$

$$\hat{w}_{2i} = \frac{y_i - \hat{\mu}_2 - \hat{\sigma}_2\hat{\rho}_{21}\hat{u}_{i1}}{\hat{\sigma}_2(1-\hat{\rho}_{21}^2)^{1/2}},$$

$$g_1'(w_{ij}) = -w_{ij}^{-2}, \quad i = 1,2,...,n;\ j = 1,2,$$

$$g_2'(w_{ij}) = (p_j - 1)w_{ij}^{p_j - 2},$$

$$h_1'(z_i) = -z_i^{-2} \quad \text{and} \quad h_2'(z_i) = (p-1)z_i^{p-2}.$$

CHAPTER 5

SIMULATION STUDY AND APPLICATION

An estimator is said to be a good estimator if it is unbiased, efficient and robust to the deviations from an asummed distribution and data anamolies. With the estimators having these properties, more reliable results can be obtained in the hypothesis testing. Therefore, a comprehensive simulation study is conducted to examine these properties of the LS and MML estimators for multiple linear regression model with both stochastic and nonstochastic covariates in this chapter. For multiple linear regression analysis with nonstochastic covariates, the error term is assumed to be Weibull. Similarly, for multiple linear regression analysis with stochastic covariates, it is assumed that the error term and explanatory variables come from a Weibull distribution. Also, power comparisons of the test statistics based on LS and MML estimators are presented for the proposed regression models to show that which one of the obtained F and F^* statistics are more powerfull. All simulation results given in this study are based on $100000/n$ Monte Carlo runs. Lastly, applications of these two proposed models are given through the expression data presented in Table 3.1.

5.1 Bias and Efficiency Comparisons

To explore whether LS and MML estimators are unbiased or not, the means of these estimators are obtained and compared with the true values of the parameters. Since it is known that both LS and MML estimators are unbiased, the variances of these estimators are obtained and the relative efficiencies (REs) of LS estimators are computed to decide which one is better. REs of LS estimator is obtained by

$$RE(LS) = 100 \times \frac{Variance\ of\ MML}{Variance\ of\ LS}. \quad (5.1)$$

Also, the mean squared error (MSE) is used as the performance characterization of the estimators. MSE of an estimator $\hat{\theta}$ is given by

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + (Bias(\hat{\theta}))^2. \quad (5.2)$$

Monte Carlo averages, variances and MSEs of LS and MML estimators and REs of LS estimators are given in Table 5.1 for the multiple linear regression model when the explanatory variables are nonstochastic and errors have a Weibull distribution with the shape parameter $p = 2, 4, 6$ and 8 and the number of explanatory variables q is 3. Without loss of generality, it is assumed that $\sigma = 1$, $\gamma_0 = 0$ and $\gamma_j = 1$ $(j = 1, 2, .., q)$. Here, *n* is taken as 10 since the sample size in the real life applications is generally small. Table 5.1 indicates that both LS and MML estimators have a negligible bias and variances of MML estimators are smaller than those of LS estimators, that is, MML estimators are more efficient. Also, the MSEs of MML estimators are smaller than those of LS estimators. Therefore, it can be said that MML estimators have better properties in terms of unbiasedness and efficiency.

The properties stated above are also examined for the re-parameterized multiple linear regression model with nonstochastic covariates. Obtained Monte Carlo results are represented in Table 5.2 for $n = 10$ and $q = 3$. When Table 5.1 and Table 5.2 are compared with each other, it is seen that re-parameterization reduces the variances of the estimators.

Table 5.1: Monte Carlo averages, variances, MSEs and REs for multiple linear regression with nonstochastic covariates; $n = 10$, $q = 3$, $\sigma = 1$, $\gamma_0 = 0$ and $\gamma_j = 1$ $(j = 1, 2, .., q)$.

	$p = 2$					$p = 4$				
	γ_0	γ_1	γ_2	γ_3	σ	γ_0	γ_1	γ_2	γ_3	σ
Mean (LS)	0.044	1.001	0.995	0.995	0.956	0.032	1.002	1.001	1.000	0.964
Mean (MML)	0.078	1.002	0.994	0.996	0.945	0.052	1.002	1.001	1.000	0.934
n × var (LS)	4.117	4.521	4.529	4.665	0.839	1.723	1.337	1.383	1.379	0.742
n × var (MML)	3.623	4.197	4.132	4.281	0.602	1.571	1.208	1.255	1.251	0.567
MSE (LSE)	4.119	4.521	4.529	4.665	0.841	1.724	1.337	1.383	1.379	0.743
MSE (MML)	3.629	4.197	4.132	4.281	0.605	1.574	1.208	1.255	1.251	0.571
RE(LS)	88	93	91	92	72	91	90	91	91	76
	$p = 6$					$p = 8$				
Mean (LS)	0.038	1.000	0.997	1.003	0.959	0.036	1.001	1.001	0.999	0.961
Mean (MML)	0.039	0.999	0.998	1.002	0.941	0.032	1.001	1.001	1.000	0.947
n × var (LS)	1.306	0.695	0.688	0.684	0.792	1.167	0.420	0.430	0.417	0.850
n × var (MML)	1.223	0.631	0.631	0.613	0.697	1.106	0.385	0.388	0.383	0.768
MSE (LSE)	1.307	0.695	0.688	0.684	0.794	1.168	0.420	0.430	0.417	0.852
MSE (MML)	1.225	0.631	0.631	0.613	0.700	1.107	0.385	0.388	0.383	0.771
RE(LS)	94	91	92	90	88	95	92	90	92	90

Table 5.2: Monte Carlo averages, variances, MSEs and REs for re-parameterized multiple linear regression with nonstochastic covariates; $n=10$, $q=3$, $\sigma=1$, $\gamma_0=0$ and $\gamma_j=1$ $(j=1,2,..,q)$.

	$p=2$					$p=4$				
	γ_0	γ_1	γ_2	γ_3	σ	γ_0	γ_1	γ_2	γ_3	σ
Mean (LS)	0.042	1.001	0.997	1.002	0.951	0.039	0.999	1.001	1.000	0.956
Mean (MML)	0.172	1.000	0.998	1.002	0.895	0.061	0.999	1.001	1.000	0.926
n × var (LS)	0.606	0.305	0.301	0.306	0.818	0.707	0.092	0.091	0.092	0.750
n × var (MML)	0.426	0.285	0.281	0.287	0.688	0.667	0.083	0.082	0.083	0.694
MSE (LSE)	0.608	0.305	0.301	0.306	0.821	0.709	0.092	0.091	0.092	0.752
MSE (MML)	0.456	0.285	0.281	0.287	0.699	0.671	0.083	0.082	0.083	0.700
RE(LS)	70	93	93	94	84	94	90	90	90	93
	$p=6$					$p=8$				
Mean (LS)	0.043	1.000	1.000	0.999	0.953	0.039	1.000	1.000	1.001	0.958
Mean (MML)	0.044	1.000	1.000	0.999	0.935	0.036	1.000	1.000	1.001	0.945
n × var (LS)	0.782	0.044	0.044	0.045	0.802	0.870	0.028	0.029	0.028	0.878
n × var (MML)	0.764	0.038	0.038	0.039	0.755	0.833	0.022	0.022	0.022	0.808
MSE (LSE)	0.784	0.044	0.044	0.045	0.805	0.871	0.028	0.029	0.028	0.879
MSE (MML)	0.766	0.038	0.038	0.039	0.759	0.834	0.022	0.022	0.022	0.811
RE(LS)	98	87	87	87	94	96	77	79	78	92

Given in Table 5.3 are the Monte Carlo averages, variances and MSEs of LS and MML estimators and REs of LS estimators for stochastic multiple linear regression analysis for $n=10$ and $q=3$. It is seen that both LS and MML estimators have negligible bias and MML estimators are more efficient than LS estimators since they have smaller variances. Also, MML estimators have smaller MSEs.

Table 5.3 Monte Carlo averages, variances, MSEs and REs for stochastic multiple linear regression; $n = 10$, $q = 3$.

		Mean		$n \times$ *Variance*		*MSE*		
Parameter	*True Value*	*LS*	*MML*	*LS*	*MML*	*LS*	*MML*	*RE*(*LS*)
				$p = 4,\ p_1 = 2,\ p_2 = 4,\ p_3 = 6$				
μ_1	0	0.025	0.029	0.409	0.401	0.410	0.402	98
μ_2	0	0.017	0.018	0.497	0.481	0.497	0.481	97
μ_3	0	0.044	0.041	0.658	0.638	0.660	0.640	97
σ_1	1	0.972	0.969	0.583	0.555	0.584	0.556	95
σ_2	1	0.930	0.929	0.312	0.303	0.317	0.308	97
σ_3	1	0.849	0.852	0.206	0.202	0.229	0.224	98
ρ_{21}	0.5	0.511	0.521	1.183	1.169	1.183	1.169	99
$\rho_{31.2}$	0.5	0.571	0.568	0.292	0.280	0.297	0.285	96
$\rho_{32.1}$	0.5	0.434	0.434	0.362	0.358	0.366	0.362	99
γ_0	0	0.126	0.118	6.981	6.139	6.997	6.153	88
γ_1	1	0.971	0.968	5.162	5.107	5.163	5.108	99
γ_2	1	0.935	0.928	3.621	3.554	3.625	3.559	98
γ_3	1	0.844	0.857	8.462	8.369	8.486	8.389	99
σ	1	0.959	0.962	0.742	0.720	0.744	0.721	97
				$p = 8,\ p_1 = 2,\ p_2 = 4,\ p_3 = 6$				
μ_1	0	0.024	0.030	0.421	0.390	0.422	0.391	93
μ_2	0	0.019	0.018	0.496	0.477	0.496	0.477	96
μ_3	0	0.040	0.041	0.671	0.649	0.673	0.651	97
σ_1	1	0.974	0.965	0.580	0.554	0.581	0.555	96
σ_2	1	0.859	0.879	0.310	0.280	0.330	0.295	90
σ_3	1	0.953	0.949	0.211	0.200	0.213	0.203	95
ρ_{21}	0.5	0.461	0.478	1.163	1.147	1.165	1.147	99
$\rho_{31.2}$	0.5	0.530	0.525	0.295	0.271	0.296	0.272	92
$\rho_{32.1}$	0.5	0.432	0.437	0.359	0.326	0.364	0.330	91
γ_0	0	0.128	0.134	4.445	4.412	4.461	4.430	99
γ_1	1	0.979	0.968	1.901	1.860	1.901	1.861	98
γ_2	1	0.930	0.926	1.276	1.218	1.281	1.223	95
γ_3	1	0.898	0.897	2.591	2.415	2.601	2.426	93
σ	1	0.954	0.957	0.858	0.827	0.860	0.829	96

MML estimators are expected to give better results for larger sample sizes. Therefore, simulation results of Table 5.1 and Table 5.3 are also obtained for $n = 30$ and $n = 50$ and represented by Table A.1 and Table A.2, respectively in Appendix A.

5.2 Robustness Comparisons of Estimators

In statistical analyses, it is expected that the obtained estimators have an optimal properties with respect to the assumed distributions. There are several procedures such as graph-plotting methods and goodness of fit tests to get information about the underlying distributions, however, these procedures might be unsuccessful in determining the shape parameters. Misspecified parameters, contaminations and data anomalies (outliers, inliers, etc.) cause the deviations from an assumed distribution and this situation brings the issue of robustness in focus.

An estimator is called *robust* if it is fully efficient (or nearly so) for an assumed distribution but maintains high efficiency for plausible alternatives. To examine the robustness properties of LS and MML estimators in multiple linear regression model with nonstochastic covariates, here it is assumed that the errors come from a Weibull distribution with parameter $p = 8$ and the random samples are generated from the following plausible models;

i) Misspecification of the distribution

   (1) $Weibull(\ p = 7, \sigma\ )$

   (2) $Weibull(\ p = 9, \sigma\ )$

   (3) $Beta(\ 4, 2.5\ )$

ii) Contamination model

   (4) $0.90\, Weibull(\ p = 8, \sigma\ ) + 0.10\, Weibull(\ p = 7, \sigma\ )$

   (5) $0.90\, Weibull(\ p = 8, \sigma\ ) + 0.10\, \mathrm{Beta}(4, 2.5)$

Table 5.4 gives the Monte Carlo averages and variances of LS and MML estimators and REs of LS estimators for the plausible alternatives given above when $n = 30$. It indicates that the MML estimators are remarkably efficient and robust than the LS estimators and the efficiencies of LS estimators decreases as the sample size increases.

Table 5.4: Robustness comparisons for multiple linear regression model with nonstochastic covariates, $n = 30$, $q = 3$, $\sigma = 1$, $\gamma_0 = 0$ and $\gamma_j = 1\ (j = 1,2,..,q)$.

	Model1					*Model2*				
	γ_0	γ_1	γ_2	γ_3	σ	γ_0	γ_1	γ_2	γ_3	σ
Mean (LS)	-0.093	0.996	0.998	1.002	0.990	0.064	1.002	0.996	1.000	0.989
Mean (MML)	-0.097	0.996	0.998	1.002	0.989	0.066	1.002	0.996	1.000	0.990
n × var (LS)	0.792	0.255	0.268	0.258	0.599	0.868	0.270	0.263	0.264	0.653
n × var (MML)	0.746	0.234	0.241	0.233	0.557	0.770	0.235	0.229	0.230	0.565
RE(LS)	94	92	90	90	93	89	87	87	87	87
	Model3					*Model4*				
	γ_0	γ_1	γ_2	γ_3	σ	γ_0	γ_1	γ_2	γ_3	σ
Mean (LS)	-0.084	1.000	1.000	1.000	0.992	0.001	1.001	0.999	1.003	0.985
Mean (MML)	-0.086	1.001	1.000	1.001	0.997	0.003	1.000	0.999	1.001	0.978
n × var (LS)	0.668	0.268	0.267	0.274	0.461	0.812	0.276	0.264	0.266	0.603
n × var (MML)	0.656	0.244	0.238	0.251	0.447	0.744	0.244	0.238	0.239	0.544
RE(LS)	98	91	89	92	97	92	88	90	90	90

Table 5.4 (Continued)

Model5

	γ_0	γ_1	γ_2	γ_3	σ
Mean (LS)	-0.094	1.001	0.997	1.004	1.118
Mean (MML)	-0.071	1.001	0.998	1.003	1.106
n × var (LS)	1.275	0.397	0.396	0.405	0.963
n × var (MML)	0.975	0.298	0.296	0.304	0.707
RE(LS)	76	75	75	75	73

To examine the robustness properties of LS and MML estimators in stochastic multiple linear regression model, it is assumed that e_i comes from a Weibull distribution with shape parameter $p = 8$. It is also assumed that, for $q = 3$, w_{i1}, w_{i2} and w_{i3} come from a Weibull distribution with parameters $p_1 = 2$, $p_2 = 4$ and $p_3 = 6$, respectively. Under these assumptions, the random samples are generated from the following plausible models;

i) Misspecification of the distribution

(1) $Weibull(\ p = 7, \sigma)$

(2) $Beta(4, 2.5)$

ii) Contamination model

(3) $0.90\, Weibull(\ p = 8, \sigma) + 0.10\, Weibull(\ p = 7, \sigma)$

(4) $0.90\, Weibull(\ p = 8, \sigma) + 0.10\, Beta(4, 2.5)$

Given in Table 5.5 are the Monte Carlo averages and variances of LS and MML estimators and REs of LS estimators under the given plausible alternatives for

stochastic multiple linear regression model when $n = 30$. It is seen that MML estimators are also remarkably efficient and robust than the LS estimators when the regression model has stochastic covariates. Also, the table indicates that the efficiencies of LS estimators decreases as the sample size increases.

For $n = 50$, simulations of Table 5.4 and Table 5.5 are given in Table A.3 and Table A.4, respectively in Appendix A.

Table 5.5: Robustness comparisons for stochastic multiple linear regression model; $n = 30$, $q = 3$.

		Model 1					*Model 2*				
		Mean		$n \times$ *Variance*			*Mean*		$n \times$ *Variance*		
Parameter	*True Value*	*LS*	*MML*	*LS*	*MML*	*RE(LS)*	*LS*	*MML*	*LS*	*MML*	*RE(LS)*
μ_1	0	0.010	0.009	0.374	0.347	93	0.011	0.005	0.398	0.378	95
μ_2	0	0.009	0.008	0.493	0.454	92	0.014	0.006	0.498	0.468	94
μ_3	0	0.009	0.009	0.580	0.545	94	0.014	0.006	0.568	0.545	96
σ_1	1	0.989	0.989	0.551	0.491	89	0.992	0.993	0.570	0.502	88
σ_2	1	0.970	0.971	0.547	0.487	89	0.937	0.959	0.576	0.501	87
σ_3	1	0.945	0.944	0.568	0.500	88	0.944	0.953	0.574	0.511	89
ρ_{21}	0.5	0.548	0.549	0.255	0.237	93	0.520	0.510	0.253	0.233	92
$\rho_{31.2}$	0.5	0.524	0.524	0.283	0.255	90	0.527	0.522	0.269	0.244	91
ρ_{321}	0.5	0.517	0.537	0.422	0.397	94	0.531	0.530	0.437	0.411	94
γ_0	0	-0.090	-0.076	4.643	4.272	92	-0.088	-0.111	4.182	3.890	93
γ_1	1	0.986	0.988	1.553	1.491	96	0.988	0.995	1.438	1.279	89
γ_2	1	0.901	0.900	1.470	1.381	94	0.979	0.982	1.234	1.074	87
γ_3	1	0.947	0.949	1.668	1.551	93	0.949	0.952	1.398	1.231	88
σ	1	0.990	0.992	1.023	0.910	89	0.993	0.994	0.448	0.439	98

Table 5.5 (Continued)

Parameter	True Value	Model 3					Model 4				
		Mean		$n \times$ Variance			Mean		$n \times$ Variance		
		LS	MML	LS	MML	RE(LS)	LS	MML	LS	MML	RE(LS)
μ_1	0	0.011	0.004	0.412	0.383	93	0.009	0.005	0.397	0.373	94
μ_2	0	0.009	0.007	0.508	0.458	90	0.010	0.008	0.497	0.462	93
μ_3	0	0.014	0.006	0.549	0.505	92	0.014	0.008	0.585	0.538	92
σ_1	1	0.989	0.996	0.583	0.519	89	0.990	0.993	0.562	0.506	90
σ_2	1	0.970	0.999	0.589	0.512	87	0.959	0.968	0.541	0.493	91
σ_3	1	0.942	0.942	0.566	0.498	88	0.943	0.951	0.535	0.476	89
ρ_{21}	0.5	0.525	0.524	0.262	0.234	89	0.527	0.525	0.266	0.250	94
$\rho_{31.2}$	0.5	0.532	0.530	0.269	0.243	90	0.522	0.519	0.271	0.252	93
$\rho_{32.1}$	0.5	0.526	0.525	0.424	0.373	88	0.523	0.521	0.431	0.410	95
γ_0	0	0.033	0.010	4.544	4.226	93	-0.084	-0.082	5.282	4.806	91
γ_1	1	0.991	0.999	1.406	1.265	90	0.988	0.998	1.759	1.548	88
γ_2	1	0.970	0.975	1.284	1.181	92	0.960	0.973	1.678	1.460	87
γ_3	1	0.939	0.941	1.822	1.658	91	0.958	0.962	1.901	1.616	85
σ	1	0.991	0.995	0.642	0.597	93	1.120	1.102	0.986	0.868	88

5.3 Power Comparisons of Test Statistics

As mentioned in Chapter 3 and 4, the distributions of the test statistics F and F^* under the null hypothesis are referred to central F-distribution, $F(q,n-q-1)$. The power function of F and F^* tests are given by

$$P(F \geq F_{1-\alpha}(q,n-q-1)\,|\,H_1) \quad \text{and} \quad P(F^* \geq F_{1-\alpha}(q,n-q-1)\,|\,H_1), \quad (5.3)$$

respectively, where α denotes the Type I error. Type I error is the probability of the rejection of null hypothesis when it is true. In hypothesis testing procedure, it is desired that the power of the test is low under the null hypothesis and high under the alternative hypothesis.

Since the MML estimators of the model parameters have smaller variances, the test statistic F^* is expected to be more powerful than the test statistic F.

Table 5.6 gives the Monte Carlo power values of F and F^* of the true model and plausible alternatives given in Section 5.2 for the multiple linear regression analysis with nonstochastic covariates when $q=2$. In the true model, error terms is assumed to be Weibull with the shape parameter $p=8$. For $\gamma_1=0$, the power reduces to Type I error which is assumed as 0.05 in this study. This table indicates that Type I errors based on MML estimators are close to preassumed $\alpha=0.05$ while the test statistic based on LS estimators does not provide the preassumed Type I error.

Table 5.6: Power of F and F^* tests for multiple linear regression model with nonstochastic covariates; true model $Wei(8,\sigma)$, $\alpha = 0.05$, $n = 10$, $q = 2$, $\gamma_0 = 0$, $\gamma_2 = 0$ and $\sigma = 1$.

Model	*True*		*(1)*		*(2)*		*(3)*		*(4)*		*(5)*	
γ_1	F	F^*	F	F^*	F	F^*	F	F^*	F	F^*	F	F^*
0.0	0.004	0.052	0.004	0.059	0.005	0.058	0.002	0.059	0.003	0.054	0.003	0.044
0.2	0.020	0.154	0.013	0.145	0.021	0.149	0.011	0.146	0.017	0.154	0.008	0.097
0.4	0.063	0.443	0.054	0.440	0.070	0.445	0.038	0.429	0.062	0.443	0.025	0.312
0.6	0.156	0.764	0.124	0.750	0.170	0.758	0.104	0.749	0.152	0.762	0.066	0.617
0.8	0.293	0.923	0.253	0.919	0.322	0.928	0.218	0.927	0.295	0.921	0.136	0.843
1.0	0.452	0.980	0.410	0.977	0.475	0.977	0.366	0.979	0.446	0.976	0.243	0.943
1.2	0.606	0.994	0.567	0.992	0.631	0.992	0.547	0.994	0.594	0.992	0.361	0.979
1.4	0.733	0.998	0.708	0.998	0.756	0.998	0.691	0.998	0.731	0.998	0.488	0.994

To provide a precise comparison of powers of F and F^* statistics, the 5% points of their distributions are determined exactly by simulation. They are given in Table 5.7.

Table 5.7: The exact 5% points of the distributions of F and F^* for multiple linear regression model with nonstochastic covariates.

	True Model	*Model 1*	*Model 2*	*Model 3*	*Model 4*	*Model 5*
F	1.60	1.45	1.59	1.35	1.56	1.26
F^*	4.83	4.83	4.85	4.89	4.90	4.39

Given in Table 5.8 are the Monte Carlo power values of F and F^* obtained by using the simulated critical values. This table indicates that Type I errors based on both LS and MML estimators provide the preassumed $\alpha = 0.05$. However, power values are higher and converge to 1.0 faster when the MML estimators are used in testing procedure. For larger sample sizes, it is expected that the rate of convergence will be higher. Figure 5.1 gives the graphs of the power curves for various values of parameter γ_1.

Table 5.8: Power of F and F^* obtained by using simulated critical values for multiple linear regression model with nonstochastic covariates; true model $Wei(8,\sigma)$, $n=10$, $q=2$, $\gamma_0=0$, $\gamma_2=0$ and $\sigma=1$.

Model	*True*		*(1)*		*(2)*		*(3)*		*(4)*		*(5)*	
γ_1	F	F^*	F	F^*	F	F^*	F	F^*	F	F^*	F	F^*
0.0	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
0.2	0.143	0.169	0.135	0.166	0.142	0.193	0.141	0.174	0.133	0.180	0.114	0.144
0.4	0.396	0.435	0.397	0.435	0.418	0.434	0.398	0.421	0.403	0.436	0.313	0.349
0.6	0.651	0.753	0.653	0.748	0.672	0.757	0.670	0.733	0.651	0.750	0.551	0.645
0.8	0.831	0.922	0.839	0.913	0.845	0.921	0.865	0.917	0.838	0.918	0.735	0.862
1.0	0.930	0.976	0.942	0.979	0.937	0.978	0.953	0.977	0.935	0.977	0.870	0.952
1.2	0.977	0.993	0.979	0.991	0.975	0.992	0.988	0.993	0.975	0.994	0.938	0.982
1.4	0.991	0.998	0.993	0.997	0.992	0.997	0.997	0.998	0.993	0.999	0.975	0.994

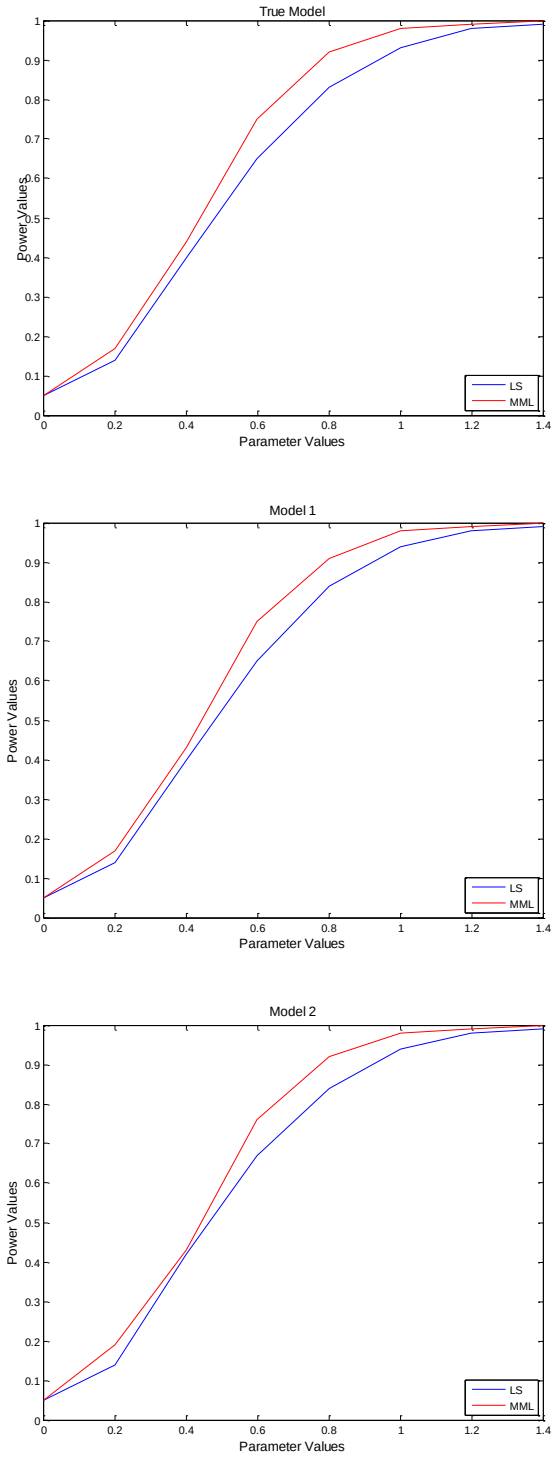
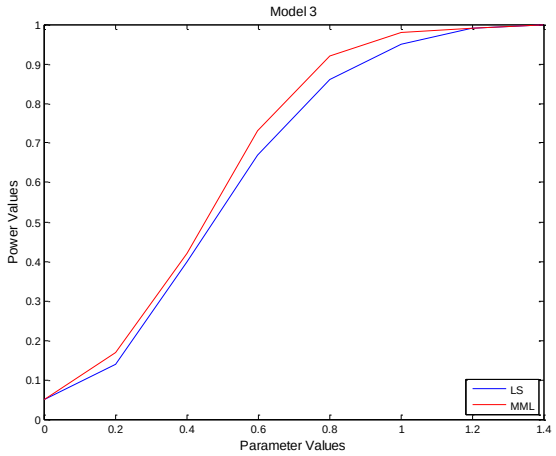
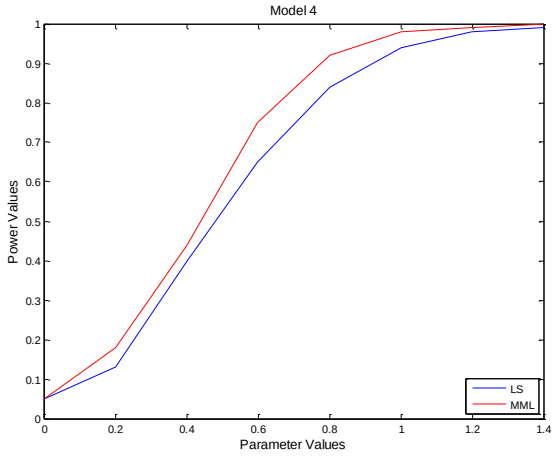


Figure 5.1: Power graphs of the tests for multiple linear regression model with nonstochastic covariates; $n=10$, $\gamma_0=0$, $\gamma_2=0$ and $\sigma=1$.





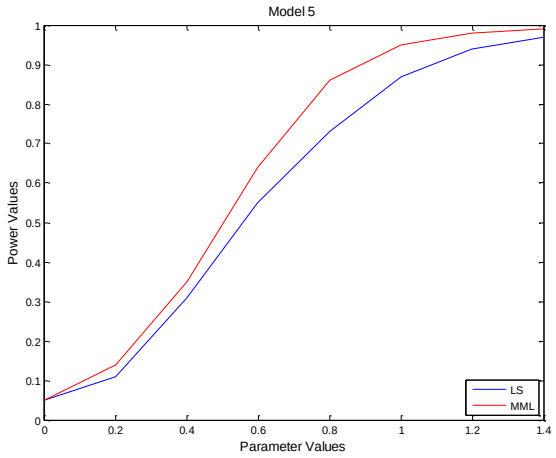


Figure 5.1 (Continued)

Given in Table 5.9 are the Monte Carlo power values of F and F^* of the true model and plausible alternatives given in Section 5.2 for the stochastic multiple linear regression analysis for $q = 2$. In the true model, error terms is assumed to be Weibull with the shape parameter $p = 8$. For $\gamma_1 = 0$, the power reduces to Type I error which is assumed as 0.05 in this study. This table indicates that Type I errors based on MML estimators are close to $\alpha = 0.05$ while the test statistic based on LS estimators does not provide the preassumed Type I error.

Table 5.9: Power of F and F^* tests for multiple linear regression model with stochastic covariates; true model $Wei(8,\sigma)$, $q=2$, $p_1=2$, $p_2=4$, $n=10$, $\gamma_0=0$, $\gamma_2=0$ and $\sigma=1$.

Model	*True*		*(1)*		*(2)*		*(3)*		*(4)*	
γ_1	F	F^*	F	F^*	F	F^*	F	F^*	F	F^*
0.0	0.012	0.052	0.007	0.052	0.007	0.053	0.011	0.051	0.005	0.040
0.2	0.091	0.164	0.051	0.312	0.042	0.309	0.062	0.318	0.028	0.234
0.4	0.227	0.767	0.193	0.765	0.165	0.758	0.222	0.774	0.110	0.668
0.6	0.461	0.941	0.432	0.943	0.400	0.939	0.465	0.943	0.275	0.898
0.8	0.691	0.988	0.661	0.983	0.654	0.986	0.683	0.985	0.458	0.967
1.0	0.851	0.995	0.826	0.997	0.827	0.997	0.839	0.996	0.649	0.991
1.2	0.922	0.999	0.917	0.999	0.924	0.999	0.924	0.999	0.783	0.997
1.4	0.966	1.000	0.964	1.000	0.968	1.000	0.968	1.000	0.881	0.999

To provide a precise comparison of powers of F and F^* statistics, the 5% points of their distributions are determined exactly by simulation. They are given in Table 5.10.

Table 5.10: The exact 5% points of the distributions of F and F^* for stochastic multiple linear regression model.

	True Model	*Model 1*	*Model 2*	*Model 3*	*Model 4*
F	2.40	2.10	1.90	2.40	1.90
F^*	4.75	4.75	4.72	4.72	4.30

Monte Carlo power values of F and F^* obtained by using the simulated critical values are given in Table 5.11. It is seen that Type I errors based on both LS and MML estimators provide the preassumed $\alpha = 0.05$. However, power values are higher and converge to 1.0 faster when the MML estimators are used in testing procedure. For larger sample sizes, it is expected that the rate of convergence will be higher. Figure 5.2 gives the graphs of the power curves for various values of parameter γ_1.

Table 5.11: Power of F and F^* tests obtained by using simulated critical values for multiple linear regression model with stochastic covariates; true model $Wei(8,\sigma)$, $q=2$, $p_1=2$, $p_2=4$, $n=10$, $\gamma_0=0$, $\gamma_2=0$ and $\sigma=1$.

Model	*True*		*(1)*		*(2)*		*(3)*		*(4)*	
γ_1	F	F^*	F	F^*	F	F^*	F	F^*	F	F^*
0.0	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
0.2	0.222	0.325	0.228	0.317	0.242	0.302	0.220	0.320	0.181	0.259
0.4	0.516	0.765	0.546	0.771	0.580	0.759	0.515	0.769	0.436	0.699
0.6	0.779	0.944	0.805	0.940	0.834	0.941	0.785	0.944	0.697	0.921
0.8	0.918	0.987	0.933	0.982	0.951	0.988	0.915	0.987	0.865	0.976
1.0	0.969	0.996	0.979	0.996	0.987	0.997	0.971	0.997	0.939	0.994
1.2	0.991	0.999	0.992	1.000	0.995	0.999	0.990	0.999	0.975	0.998
1.4	0.998	1.000	0.998	1.000	0.999	1.000	0.996	0.999	0.992	0.999

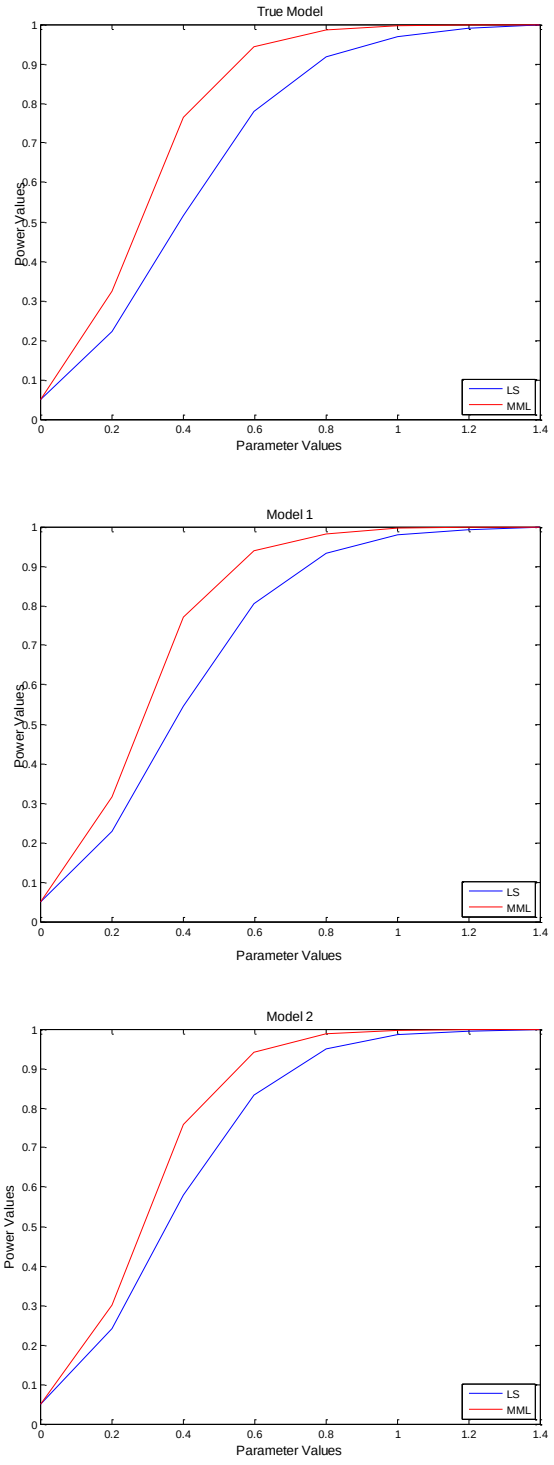
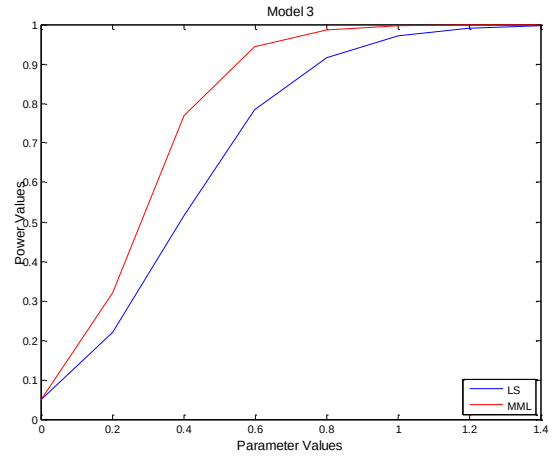


Figure 5.2: Power graphs of the tests for stochastic multiple linear regression model; $n=10$, $\gamma_0=0$, $\gamma_2=0$ and $\sigma=1$.



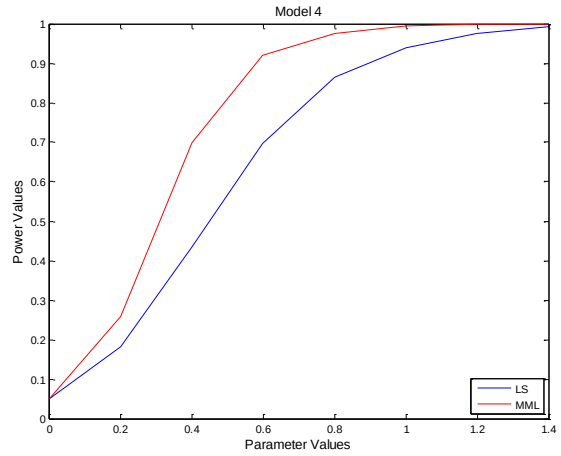


Figure 5.2 (Continued)

5.4 Application

In this study, the efficiency properties of LS and MML estimators for proposed models are also compared by using the expression and perturbation data given in Table 3.1 and Table 3.2, respectively.

In the application part of this study, firstly a multiple linear regression model with nonstochastic covariates is constructed for each of the genes in the network using LS estimation method to estimate the regulatory influences of genes on one another. Since the number of covariates (number of genes in the network) is equal to the number of experiments (number of observations), parameter estimates cannot be obtained. Therefore, it is assumed that the network is not fully connected, that is, some of the model parameters are zero. Also, the constructed model must be dynamically stable, i.e., gene expression levels must settle to steady state over time. Gardner et al. (2003) states that the multiple linear regression model becomes dynamically stable when the maximum number of covariates in the model is equal to 5. Hence, it is assumed that 3 of the parameters are equal to zero in each regression model. For each gene, it is built 9 chosen 5 number of regression models and the one with the smallest SSE is selected as the best model. When the Q-Q plots of residuals are examined to determine their distribution, the distribution of residuals is obtained as Weibull for each model. For conciseness, just one of the obtained Q-Q plot of residuals for Weibull distribution is given by Figure 5.3.

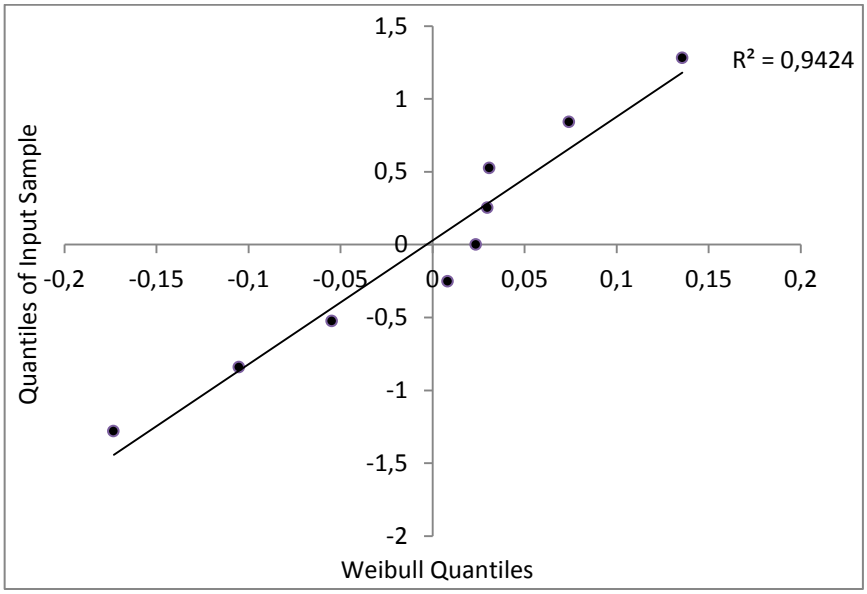


Figure 5.3: Q-Q Plot of residuals for Weibull distribution with $p = 8$.

Then, the method of MML is used to estimate the model parameters assuming that the error term in the model comes from a Weibull distribution. In Table 5.12, the first and second rows shows the influences of genes given in the columns on each gene given in the rows estimated by LS and MML methods, respectively. Given in the third and fourth rows are the variances of the LS and MML estimators. Also, the table gives the p-values of F and F^* statistics denoted by p and p^*, respectively to test significance of the constructed models. In some models, it is seen that F and F^* test statistics are not consistent in testing the equality of model parameters. For example, in the model constructed for the gene *lexA*, F^* statistic rejects the null hypothesis while the F statistic fails to reject the null hypothesis. Since the MML estimators have smaller variances, it can be concluded that the results of F^* statistic are more reliable.

Table 5.12: Constructed multiple linear regression model with nonstochastic covariates for every gene in the SOS subnetwork.

Genes	*Estimation Method*	*recA*	*lexA*	*ssb*	*recF*	*dinI*	*umuDC*	*rpoD*	*rpoH*	*rpoS*	*p*	p^*
recA	*LS*	-0.666	-0.147	-0.010	0	0.111	0	-0.011	0	0	0.88	0.08
	MML	-0.677	-0.219	-0.014	0	0.109	0	-0.020	0	0		
	Var(LS)	0.371	1.373	0.003	-	0.070	-	0.029	-	-		
	Var(MML)	0.008	0.050	0.000	-	0.002	-	0.001	-	-		
lexA	*LS*	0.546	-3.565	-0.060	0	0.138	-0.281	0	0	0	0.15	$0.02^{\dagger}$
	MML	0.549	-3.537	-0.059	0	0.138	-0.275	0	0	0		
	Var(LS)	0.116	0.471	0.001	-	0.025	0.023	-	-	-		
	Var(MML)	0.010	0.104	0.000	-	0.002	0.004	-	-	-		
ssb	*LS*	0.090	-0.285	-1.277	0	0.056	0	0.031	0	0	$0.00^{\dagger}$	$0.00^{\dagger}$
	MML	0.094	-0.286	-1.277	0	0.057	0	0.033	0	0		
	Var(LS)	0.083	0.307	0.001	-	0.016	-	0.007	-	-		
	Var(MML)	0.007	0.044	0.000	-	0.002	-	0.001	-	-		
recF	*LS*	0.127	0.897	0	-1.773	0	0	0.203	0	1.120	0.99	0.93
	MML	0.121	1.008	0	-1.999	0	0	0.229	0	1.270		
	Var(LS)	3.613	11.838	-	21.246	-	-	0.289	-	6.309		
	Var(MML)	0.888	3.701	-	5.272	-	-	0.091	-	2.202		
dinI	*LS*	0.222	0	0	0	-2.202	0.123	-0.091	0.008	0	0.10	$0.01^{\dagger}$
	MML	0.248	0	0	0	-2.197	0.132	-0.082	0.008	0		
	Var(LS)	0.955	-	-	-	0.207	0.200	0.086	0.001	-		
	Var(MML)	0.089	-	-	-	0.020	0.020	0.009	0.000	-		

Table 5.12 (Continued)

Genes	*Estimation Method*	*recA*	*lexA*	*ssb*	*recF*	*dinI*	*umuDC*	*rpoD*	*rpoH*	*rpoS*	*p*	p^*
umuDC	*LS*	0.247	-0.534	-0.017	0	0.210	-1.179	0	0	0	0.13	0.01†
	MML	0.258	-0.498	-0.015	0	0.209	-1.171	0	0	0		
	Var(LS)	0.241	0.974	0.002	-	0.051	0.048	-	-	-		
	Var(MML)	0.021	0.217	0.000	-	0.004	0.008	-	-	-		
rpoD	*LS*	-0.303	0	-0.025	0	0.009	0	-1.551	0.019	0	0.04†	0.00†
	MML	-0.310	0	-0.025	0	0.008	0	-1.552	0.019	0		
	Var(LS)	0.525	-	0.004	-	0.105	-	0.046	0.001	-		
	Var(MML)	0.053	-	0.000	-	0.011	-	0.005	0.000	-		
rpoH	*LS*	0.116	0	0.005	0	-0.011	-0.024	0	-0.482	0	0.00†	0.00†
	MML	0.112	0	0.004	0	-0.011	-0.025	0	-0.483	0		
	Var(LS)	0.046	-	0.000	-	0.010	0.010	-	0.000	-		
	Var(MML)	0.004	-	0.000	-	0.001	0.001	-	0.000	-		
rpoS	*LS*	0.575	0	0	-2.112	0.789	0	0.007	0	-4.644	0.81	0.16
	MML	0.554	0	0	-2.146	0.786	0	0.002	0	-4.662		
	Var(LS)	3.611	-	-	22.959	0.800	-	0.307	-	7.336		
	Var(MML)	0.902	-	-	5.343	0.188	-	0.093	-	2.168		

†Indicates statistically significant fit for the row.

As mentioned in Chapter 3, if the null hypothesis is rejected, then it is necessary to test significance of individual parameters. For example, in the model constructed for the gene *ssb*, both F and F^* statistics reject the null hypothesis. To determine which genes have a significant effect on this gene, individual t-tests are conducted. The p-values of T_j and T_j^* statistics obtained by obtained using LS and MML estimator, respectively are given in the last two rows of Table 5.13. From this table, it is concluded that the effects of the genes *dinI* and *rpoD* on the gene *ssb* are significant when the LS estimators are used in the testing procedure. However, when the MML estimators are used in the testing procedure, it is seen that all genes have a significant effect on the gene *ssb.*

Table 5.13: Individual t-tests of model constructed for the gene *ssb.*

Genes	*Estimation Method*	*recA*	*lexA*	*ssb*	*recF*	*dinI*	*umuDC*	*rpoD*	*rpoH*	*rpoS*
ssb	*LS*	0.090	-0.285	-1.277	0	0.056	0	0.031	0	0
	MML	0.094	-0.286	-1.277	0	0.057	0	0.033	0	0
	Var(LS)	0.083	0.307	0.001	-	0.016	-	0.007	-	-
	Var(MML)	0.007	0.044	0.000	-	0.002	-	0.001	-	-
	p	0.36	0.42	0.00†	-	0.04†	-	0.02†	-	-
	p^*	0.00†	0.00†	0.00†	-	0.00†	-	0.00†	-	-

†Indicates statistically significant coefficient.

Since the explanatory variables in the multiple linear regression model represent the gene expression changes and they are subject to the measurement errors, the explanatory variables are treated as stochastic in the second part of the application and the stochastic multiple linear regression analysis is applied to estimate the regulatory influences of genes on one another. Therefore, firstly the Q-Q plots of gene expression changes are examined to determine their distributions. Obtained Q-Q plots show that the gene expression changes have also a Weibull distribution. For conciseness, again one of the obtained Q-Q plot of gene expression changes for Weibull distribution is given by Figure 5.4.

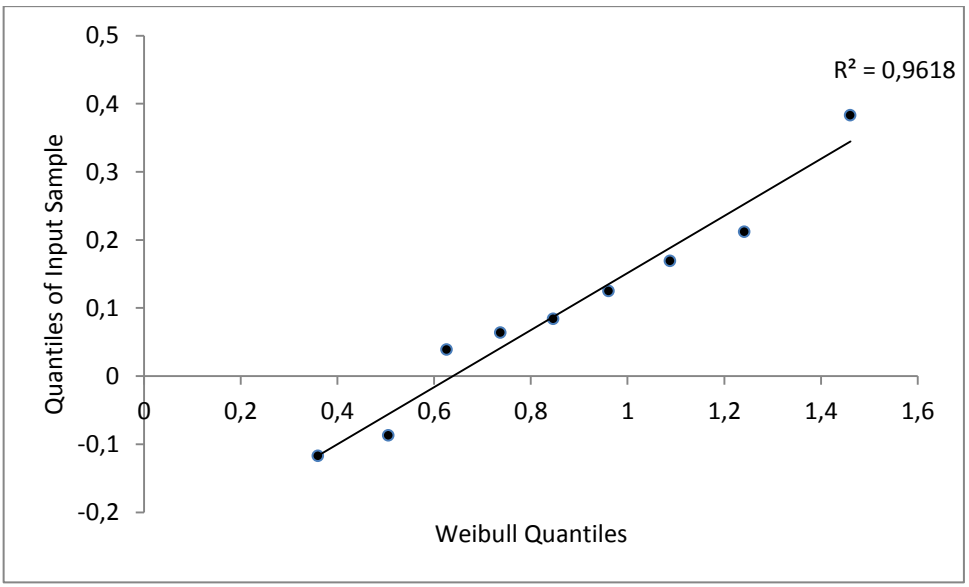


Figure 5.4: Q-Q Plot of gene expression for Weibull distribution with $p = 2.2$.

Then, the method of LS and MML are used to estimate the model parameters assuming that the error term and explanatory variables in the model comes from a Weibull distribution. In this part, stochastic multiple linear regression model for the gene *lexA* is constructed. This gene is selected since the F and F^* statistics give inconsistent results for the significance of model in the multiple linear regression analysis with nonstochastic covariates. In Table 5.14, the first and second rows show the influences of genes in the network on the gene *lexA* estimated by LS and MML methods, respectively. Given in the third and fourth rows are the bootstrap variances of the LS and MML estimators. Table indicates that the MML estimators have smaller variances and hence they are more precise. Furthermore, when the results given Table 5.12 and 5.14 are compared, it is seen that LS and MML estimates obtained for stochastic multiple linear regression model are different from those obtained for multiple linear regression model with nonstochastic covariates. Also, p-values of F and F^* statistics obtained to test the significance of the constructed models are given in the last two columns of the tables. Null hypothesis is rejected by the test statistic F^* in both of the stochastic and nonstochastic regression analyses. However, when the results of F statistic given in Table 5.12 and 5.14 are compared, it is seen that F statistic rejects the null hypothesis in stochastic regression analysis while it fails to reject the null hypothesis in nonstochastic regression analysis.

Table 5.14: Constructed multiple linear regression model with stochastic covariates for gene *lexA* in the SOS subnetwork.

Genes	*Estimation Method*	*recA*	*lexA*	*ssb*	*recF*	*dinI*	*umuDC*	*rpoD*	*rpoH*	*rpoS*	p	p^*
lexA	*LS*	0.298	-0.941	-0.379	0	0.195	-0.452	0	0	0	0.03†	0.00†
	MML	0.345	-1.066	-0.353	0	0.209	-0.531	0	0	0		
	Var(LS)	0.210	0.175	0.173	-	0.141	0.154	-	-	-		
	Var(MML)	0.197	0.165	0.156	-	0.134	0.901	-	-	-		

†Indicates statistically significant fit for the row.

CHAPTER 6

SUMMARY AND CONCLUSIONS

In this thesis, a biological background of cells, genes, DNA and RNA molecules is given in the framework of gene expression and gene regulation. Then, microarray technology used to measure gene expression levels is explained and issues about transformation and normalization of microarray data are explored to describe the preparation of data for the statistical analysis.

The most widely used methods such as Boolean networks, Gaussion models, Bayesian networks and Ordinary differential equations to reconstruct GRNs are presented. Especially, NIR algorithm which is a first order differential equation model is examined in detail since it is the motivation of this study.

The distribution of gene expression data following an external perturbation experiment is explored and determined as a distribution from Weibull family by examining their Q-Q plots and by matching (approximately) the sample skewness and kurtosis with the corresponding theoretical values of the distribution. Therefore, a theoretical background for Weibull distribution is given briefly.

To determine the regulatory relationships between genes in the network, a multiple linear regression analysis applied to gene expression and perturbation data and a theoretical explanation of the multiple linear regression model is summarized.

It is shown that the error term in the multiple linear regression model has also a Weibull distribution. Under the assumption of Weibull distributed error term, the method of LS and MML are implemented to estimate the model parameters. An extensive simulation study is carried out to examine the bias and efficiency properties of obtained LS and MML estimators. In addition, robustness properties of LS and MML estimators are explored based on the misspecified and contamination models as a plausible alternative for Weibull distribution.

For the regression model with nonstochastic covariates, F and F^* statistics are obtained to test the equality of model parameters based on LS and MML estimators, respectively. Furthermore, power values of these tests are compared under true distribution and some plausible distributions.

Since the explanatory variables represents the expression levels of genes in the network and it is shown that they come from a Weibull distribution, these variables are considered as stochastic in this study. Hence, stochastic multiple linear regression model is also used for inferring GRNs and model parameters are estimated by using method of LS and MML. Since the variances of LS estimators of model parameters are very sensitive to the location and scale parameters of the explanatory variables in the regression analysis, it is proposed to use re-parameterized model to rectify this situation.

Bias, efficiency and robustness comparisons of LS and MML estimators are also made for the stochastic multiple linear regression model assuming both explanatory variables and errors come from a Weibull distribution. In addition, test statistics for the significance of the constructed models are obtained based on LS and MML estimators and power values of them are compared under true distribution and some plausible distributions.

A real-life gene expression data are analyzed to illustrate the proposed multiple linear regression models by implementing LS and MML estimation methods.

On the basis of this research, the following conclusions can be stated:

- MML estimators derived for the proposed models are computationally and theoretically straightforward since they are the explicit solutions of the likelihood functions.

- For the multiple linear regression model with nonstochastic covariates and Weibull distributed error term, the MML estimators $\hat{\gamma}_0$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\sigma}$ are unbiased (or have negligible bias) and remarkably more efficient than the corresponding LS estimators. Also, relative efficiencies of LS estimators decrease as the sample size increases, that is, the variances of MML estimators become smaller for the larger sample sizes.

- For the stochastic multiple linear regression model with Weibull distributed covariates and Weibull distributed error term, it is indicated that MML estimators $\hat{\mu}_j, \hat{\sigma}_j$, $\hat{\rho}_{j.I_j}$ $\hat{\gamma}_0$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\sigma}$ are unbiased estimators and have better efficiency properties than the LS estimators.

- When the plausible alternatives are considered for Weibull distribution, it is shown that MML estimators obtained for the proposed models are also robust to deviations from the assumed distribution.

- Test statistic F^* obtained by using MML estimators is more powerful than the test statistic based on the LS estimators.

To sum up, multiple linear regression analysis used to construct GRNs depends on the normality assumption of errors. However, this assumption is not satisfied in most real life situations and non-normality complicates the data analysis and results in inefficient estimators. Therefore, this study examines the true distributions of errors and provides an efficient, robust and powerful estimation technique by implementing the method of MML under the true distributions of errors. Also, this study considers the explanatory variables as stochastic since they represents the gene expression changes and gene expression changes are subject to the measurement errors and proposes stochastic multiple linear regression analysis for the reconstruction of GRNs by considering the regulatory relationships between the genes.

As a future study, it is planned to handle the relationships between the model covariates by applying ridge regression analysis to the gene expression data and estimate the model parameters by using the MML estimation method.

Furthermore, it is considered to extend this study by using adaptive modified maximum likelihood (AMML) estimation method for inferring GRNs. This method is used when the nature of the underlying distribution cannot be investigated by a researcher (Dönmez, 2010).

REFERENCES

Akkaya A.D. and Tiku M.L. (2008). Robust estimation in multiple linear regression model with non-Gaussian noise. Automatica, 44, 407-417.

Alizadeh A. A., Eisen M. B., Davis R. E., Ma C., Lossos I. S., Rosenwald A., Boldrick J. C., Sabet H., Tran T., Yu X., Powell J. I., Yang L., Marti G. E., Moore T., Hudson J. Jr., Lu L., Lewis D. B., Tibshirani R., Sherlock G., Chan W. C., Greiner T.C., Weisenburger D. D., Armitage J. O., Warnke R., and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403, 503-511.

Amaratunga, D. and Cabrera, C. (2004). Exploration and Analysis of DNA Microarray and Protein Array Data. Wiley–Interscience: New Jersey.

Arfin S. M., Long A. D., Ito E. T., Tolleri L., Riehle M. M., Paegle E. S., and Hatfield G. W. (2000). Global gene expression profiling in Escherichia coli K12. J. Biol. Chem., 275, 29672-29684.

Bammert, G. F., and Fostel, J. M. (2000). Genome-wide expression patterns in Saccharomyces cerevisiae:comparison of drug treatments and genetic alterations affecting biosynthesis of ergosterol. Antimicrobial Agents Chemother, 44, 1255-1265.

Bansal M., Gatta G. D., and di Bernardo, D. (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics, 22, 7, 815–822.

Bansal M., Belcastro V., Ambesi-Impiombato A. and Bernardo D. (2007). How to infer gene networks from expression profiles. Mol. Syst., Biol., 3, 78.

Bar-Joseph Z., Gerber G. K., Lee T. I., Rinaldi N. J., Yoo J. Y., Robert F., Gordon D. B., Fraenkel E., Jaakkola T. S., Young R. A., and Gifford D. K. (2003).

Computational discovery of gene modules and regulatory networks. Nat. Biotechnol. , 21, 1337-1342.

Bhattacharyya, G. K. (1985). The asymptotics of maximum likelihood and related estimators based on type II censored data. J. Amer. Statist. Assoc., 80, 398-404.

Bilban M., Head S., Desoye G., and Quaranta, V. (2000). DNA microarrays: a novel approach to investigate genomics in trophoblast invasion: A review. Placenta. 21 (Suppl A), 99-105.

Boscolo R., Sabatti C., Liao J. C., Roychowdhury V.P. (2005). A generalized framework for network component analysis. IEEE-ACM Trans. Comput. Biol. Bioinform., 2, 289-301.

Britton R. A., Eichenberger P., Gonzalez-Pastor J. E., Fawcett P., Monson R., Losick R., and Grossman A. D. (2002). Genome-wide analysis of the stationary-phase sigma factor (sigma-H) regulon of Bacillus subtilis. J. Bacteriol, 184. 4881-4890.

Cai M., Bazerque J. A., and Giannakis G. B. (2013). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. PLoS Comput. Biol., 9, e1003068.

Chan K., Baker S., Kim C. C., Detweiler C. S., Dougan G. and Falkow S. (2003). Genomic comparison of Salmonella enterica serovars and Salmonella bongori by use of a S. Enterica serovar Typhimurium DNA microarray. Journal of Bacteriology, 185, 553-563.

Chen T., He H. L., Church G. M., (1999). Modeling gene expression with differential equations. Pacific Symposium on Biocomputing, 4, 29-40.

Chen K. C., Wang T. Y., Tseng H. H., Huang C. Y. F., Kao C. Y. (2005). A stochastic differential equation model for quantifying transcriptional regulatory network in saccharomyces cerevisiae. Bioinformatics, 21, 12, 2883-2890.

Cho R., Campbell J., Winzeler E., Steinmetz L., Conway A., Wodicka, L., Wolfsberg T., Gabrielian A., Landsman, D., and Lockart D. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell, 2, 65-73.

Cho D., Cho K., and Zhang B. (2006). Identification of biochemical networks by S-tree based genetic programming. Bioinformatics, 22, 1631–1640.

Cho K. H., Choo S. M., Jung S. H., Kim J. R., Choi H. S., and Kim, J. (2007). Reverse engineering of gene regulatory networks. IET System Biology, 1,149-163.

Chu S., Derisi J., Eisen M., Mullholland J., Botstein D., Brown P., and Herskowitz I. (1998). The transcriptional program of sporulation in budding yeast. Science, 282, 699-705.

Clarke P. A., Poele R., Wooster R., and Workman P. (2001). Gene expression microarray analysis in cancer biology, pharmacology and drug development: progress and potential. Biochemical Pharmacology, 62, 1311-1336.

Claverie J. M. (1999). Computational methods for the identification of differential and coordinated gene expression. Human Molecular Genetics, 8, 1821-1832.

Cleveland W.S (1979). Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association. 74, 829-836.

Cold Spring Harbor Laboratory. http://rulai.cshl.edu/TRED/GRN/HIF.htm. (last accessed on 24.09.2014).

Crick F. H. (1970). Central dogma of molecular biology. Nature, 227, 561-563.

de Saizieu A., Gardes C., Flint N., Wagner C., Kamber M., Mitchell T. J., Keck W., Amrein K. E. and Lange R. **(**2000**).** Microarray-based identification of a novel Streptococcus pneumoniae regulon controlled by an autoinduced peptide. J. Bacteriol, 182, 4696-4703.

Dobra A., Jones B., Hans C., Nevis J. and West M. (2004). Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis, 90, 196-212.

Dönmez, A. (2010). Adaptive estimation and hypothesis testing methods. Ph.D Thesis, Middle East Technical University: Ankara.

Draghici S. (2003). Data Analysis Tools for DNA Microarrays. Chapman & Hall/CRC, London.

Emmert-Streib F., Glazko G. V., Altay G., and De Matos Simoes R. (2012). Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Front. Genet., 3, 8, doi: 10.3389/fgene.2012.00008

Faith J. J., Driscoll M. E., Fusaro V. A., Cosgrove E. J., Hayete B., Juhn F. S., Schneider S. J., and Gardner T. S.(2008). Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental meta-data. Nucleic Acids Research, 36 (Database issue), D866-870.

Fang Y., Brass A., Hoyle D. C., Hayes A., Bashein A., Oliver S. G., Waddington D., Rattray M. (2003). A model-based analysis of microarray experimental error and normalisation. Nucleic Acids Research, 31, E96.

Friedman N., Murphy K., and Russell S. (1998). Learning the structure of dynamic probabilistic networks. Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI ’98), 139–147.

Friedman N., Linial M., Nachman I. and Pe’er D. (2000). Using Bayesian networks to analyze expression data. Journal of Computational Biology, 7, 601–620.

Gardner T. S., Bernardo D., Lorenz D., and Collins J.J. (2003). Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. Science, 301, 102-105.

Gingeras T. R., Ghandour G., Wang E., Berno A., Small P. M., Drobniewski F., Alland D., Desmond E., Holodniy M., and Drenkow J. (1998). Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic Mycobacterium DNA arrays. Genome Res., 8, 435-448.

Gross C., Kelleher M, Iyer V. R., Brown P. O., and Winge D.R. (2000). Identification of the copper regulon in Saccharomyces cerevisiae by DNA microarrays. J. Biol. Chem., 275, 32310-32316.

Hartemink A. J., Gifford D. K., Jaakkola T. S. and Young R. A. (2001). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. Pacific Symposium on Biocomputing, 6, 422–433.

Harter H.L. (1964). Exact confidence bounds, based on one order statistic, for the parameters of an exponential population. Tecnometrics, 6, 301-317.

Hecker M. (2007). Gene Regulatory Network Reconstruction Best Practice Guide. Bio-Control Jena GmbH, Technical Report.

Islam M. Q., Tiku M. L., and Yıldırım F. (2001). Non-normal regression: I. Skew distributions. Commun. Stat.-Theory Meth., 30, 6, 993-1020.

Islam, M. Q. and Tiku, M. L. (2004). Multiple linear regression model under non-normality. Commun. Stat.-Theory Meth., 33, 2443-2467.

Islam M. Q. and Tiku M. L. (2010). Multiple linear regression model with stochastic design variables. Journal of Applied Statistics, 37, 6, 923-943.

Kaderali L. and Radde N. (2008). Inferring gene regulatory networks from expression data. Computational Intelligence in Bioinformatics, 94, 33-74.

Kao K. C., Yang Y. L., Boscolo R., Sabatti C., Roychowdhury V., and Liao J.C. (2004). Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis. Proc. Natl Acad. Sci. USA, 101, 641-646.

Karlebach G. and Shamir R. (2008). Modelling and analysis of gene regulatory networks. Nat.Rev.Mol.Cell Biol., 9, 770-780.

Kauffman S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of Theoretical Biology, 22, 3, 437-467.

Kikuchi S, Tominaga D, Arita M, Takahashi K, and Tomita M (2003). Dynamic modelling of genetic networks using genetic algorithm and S-system. Bioinformatics, 19, 643-650.

Kuhn K. M, DeRisi J. L, Brown P. O, and Sarnow P. (2001). Global and specific translational regulation in the genomic response of Saccharomyces cerevisiae to a rapid transfer from a fermentable to a non-fermentable carbon source. Mol. Cell. Biol., 21, 916-927.

Laub M. T., McAdams H. H., Feldblyum T., Fraser C. M., and Shapiro L. (2000). Global analysis of gene network controlling a bacteria cell cycle. Science, 290, 2144-2148.

Laubenbacher R., and Stigler B. (2004). A computational algebra approach to the reverse engineering of gene regulatory networks. J. Theo. Biol, 229, 523-537.

Lee K. R., Kapadia C. H., and Dwight B. B. (1980). On estimating the scale parameter of Rayleigh distribution from censored samples. Statist. Hefte, 21, 14-20.

Lee M.-L. T. (2004). Analysis of Microarray Gene Expression Data. Kluwer Academic Publishers: Boston.

Lee W. P. and Tzou W. S. (2009). Computational methods for discovering gene networks from expression data. Briefing in Bioinformatics, 10, 4, 408-23.

Li X., Rao S., Jiang W., Li C., Xiao Y., Guo Z., Zhang Q., Wang L., Du L., Li J., Li L., Zhang T., and Wang Q. K. (2006). Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling. BMC Bioinformatics, 7, 26-20.

Liang S., Fuhrman S. and Somogyi R. (1998). REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. Proc. Pacific Symp. on Biocomputing, 3, 18-29.

Lou X. J., Schena M., Horrigan F. T., Lawn R. M., and Davis R. W. (2001). Expression monitoring using cDNA microarrays. A general protocol. Methods in Molecular Biology, 175, 323- 340.

Macoska J. A. (2002). The progressing clinical utility of DNA microarrays, CA-Canc. J. Clin., 52, 50-9.

Margolin A., Nemenman I., Basso K., Wiggins C., Stolovitzky G., Favera R., and Califano A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics, 7 (Suppl 1), S7.

Murphy K.P. and Mian S. (1999). Modelling gene expression data using dynamic bayesian networks. Technical Report, Computer Science Division, University of California, Berkeley, CA.

Nachman I., Regev A., and Friedman N. (2004). Inferring quantitative models of regulatory networks from expression data. Bioinformatics, 20, 248–256.

Nadon R. and Shoemaker J. (2002). Statistical issues with microarrays: processing and analysis. Trends in Genetics. 18, 5, 265-71.

Oral E. (2006). Binary regression with stochastic covariates. Communications in Statistics-Theory and Methods, 35, 1429-1447.

Ott S. , Imoto S., and Miyano S. (2004) Finding optimal models for small gene networks. Pac. Symp. Biocomput., 557–567.
Panse C. and Kshirsagar M. (2013). Survey on Modelling Methods applicable to gene regulatory, International Journal on Bioinformatics & Biosciences, 3, 3, 13-23.

Parmigiani G., Garrett E. S., Irizarry R. A. and Zeger S. L. (2003). The Analysis of Gene Expression Data, Springer-Verlag.

Perrin B. E., Ralaivola L., Mazurie A., Bottani S. 2, Mallet J., and d'Alché–Buc F. (2003). Gene networks inference using dynamic Bayesian networks. Bioinformatics, 19, 138-148.

Puthenpura S. and Sinha N. K. (1986). Modified maximum likelihood method for the robust estimation of system parameters from very noisy data. Automatica, 22, 231-235.

Quackenbush J. (2002). Microarray data normalization and transformation. Nature Genetics Supplement, 32, 496-501.

Sabatti C. and James G. M. (2006). Bayesian sparse hidden components analysis for transcription regulation networks. Bioinformatics, 22, 737-744.

Sazak H. S., Tiku M. L. and Islam M. Q. (2006). Regression analysis with a stochastic design variable. International Statistical Review, 74, 77-88.

Schena M., Shalon D., Davis R. W. and Brown P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270, 467-470.

Schuchhardt J., Beule D., Malik A., Wolski E., Eickhoff H., Lehrach H., and Herzel H. (2000). Normalization strategies for cDNA microarrays. Nucleic Acids Research, 28, E47.

Shmulevich I., Dougherty E. R., Kim S. and Zhang W. (2002). Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics, 18, 261–274.

Silvescu A. and Honavar V. (1997). Temporal Boolean Network Models of Genetic Networks and Their Inference from Gene Expression Time Series. Complex Systems, 11, 1-18.

Smith W. B., Zeis C. D., and Syler G. W. (1973). Three parameter lognormal estimation from censored data. J. Indian Statistical Association, 11, 15-31.

Smyth G. K., Yang Y.-H., and Speed T. P. (2003). Statistical issues in cDNA microarray data analysis. Methods in Molecular Biology, 224, 111-136.

Tan W. Y. (1985). On Tiku's robust procedure-a Bayesian insight. J. Statist. Plann. and Inf., 11, 329-340.

Tanaka T. S., Jaradat S. A., Lim M. K., Kargul G. J., Wang X., Grahovac M. J., Pantano S., Sano Y., Piao Y., Nagaraja R., Doi H., Wood III W. H., Becker K. G., and Ko M. S. (2000). Genome-wide expression profiling of midgestation placenta and embryo using a 15, 000 mouse developmental cDNA microarray. Proceedings of the National Academy of Sciences (USA), 97, 9127-9132.

Tao H., Bausch C., Richmond C., Blattner F. R., and Con way T. (1999). Functional genomics:expression analysis of Escherichia coli growing on mnimal and rich media. J. Bacteriol, 181, 6425-6440.

Tiku M. L. (1967). Estimating the mean and Standard deviation from censored normal samples. Biometrika, 54, 155-165.

Tiku M. L., Tan W. Y., and Balakrishnan N. (1986). Robust Inference. Marcel Dekker: New York.

Tiku M. L. and Suresh R. P. (1992). A new method of estimation for location and scale parameters. J. Stat. Plan. Inf., 30, 281-292.

Tiku M. L. and Akkaya A. D. (2004). Robust Estimation and Hypothesis Testing. New Age International Publishers:New Delhi.

Tiku M. L. and Akkaya A. D. (2010). Random design in regression models (invited paper). JISAS. 64, 2, 157-170.

Tominaga D. and Okamoto M. (1998). Design of canonical model describing complex nonlinear dynamics. Proc. IFAC Int. Conf., CAB7, 85-90.

Tseng G. C., Oh M. K., Rohlin L., Liao J. C., and Wong W. H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Research, 29, 2549-2557.

van't Veer L. J., Dai H., van de Vijver M. J., He Y. D., Hart A. A., Moa M., Peterse H. L., van der Kooy K., Maron M. J., Witteveen A. T., Schreiber G. J., Kerkhoven R. M., Roberts C., Linsley P. S., Bernards R., and Friend S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415, 530-536.

Vaughan D. C and Tiku M. L. (2000). Estimation and hypothesis testing for a non-normal bivariate distribution with applications. Journal of Mathematical and Computer Modelling, 32, 3-67.

Vierstraete A. The central dogma of molecular biology. http://users.ugent.be/~avierstr/principles/centraldogma.html. (last accessed on 24.09.2014).

Wang Z., Xu W., San Lucas F. A. and Liu Y. (2013). Incorporating prior knowledge into Gene Network Study. Bioinformatics 29, 20, 2633-2640.

Watson J. D. and Crick F. H. (1958). On protein synthesis. The Symposia of the Society for Experimental Biology, 12, 138-163.

Wei Y., Lee J. M., Richmond C., Blattner F. R., Rafalski J. A., and LaRossa R. A. (2001). High-density microarray-mediated gene expression profiling of Escherichia coli. J. Bacteriol, 183, 545-556.

Whittaker J. (1990). Graphical Models in Applied Multivariate Statistics. JohnWiley & Sons: New York, USA.

Wilson M., DeRisi J., Kristensen H. H., Imboden P., Rane S., Brown P. O., Schoolnik G. K. (1999). Exploring drug-induced alterations in gene expression in Mycobacterium tuberculosis by microarray hybridization. Proc. Natl. Acad. Sci. (USA), 96, 22, 12833-12838.

Xing B. and Laan, M. J. (2005). A statistical method for constructing transcriptional regulatory networks using gene expression and sequence data. Journal of Computational Biology, 12, 229-246.

Yang Y. H., Dudoit S., Luu P., Lin D. M., Peng V., Ngai J. and Speed T. P. (2002). Normalisation for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research, 30, E15.

Young R. A. (2000). Biomedical discovery with DNA arrays. Cell, 102, 9-15.

Yue H., Eastman P. S., Wang B. B., Minor J., Doctolero M. H., Nuttall R. L., Stack R., Becker J. W., Montgomery J. R., Vainer M., and Johnston R. (2001). An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. Nucleic Acids Research, 29, E41.

APPENDIX A

SIMULATION RESULTS FOR LARGE SAMPLE SIZES

Table A.1: Monte Carlo averages, variances, MSEs and REs for multiple linear regression with nonstochastic covariates; $q = 3$, $\sigma = 1$, $\gamma_0 = 0$ and $\gamma_j = 1$ $(j = 1,2,..,q)$.

	$n = 30$									
	$p = 2$					$p = 4$				
	γ_0	γ_1	γ_2	γ_3	σ	γ_0	γ_1	γ_2	γ_3	σ
Mean (LS)	0.018	1.000	0.992	0.993	0.988	0.010	1.003	1.000	0.999	0.987
Mean (MML)	0.036	1.001	0.994	0.994	0.953	0.024	1.003	1.000	0.998	0.971
n × var (LS)	2.708	3.073	2.887	2.951	0.623	1.110	0.867	0.872	0.906	0.494
n × var (MML)	2.091	2.450	2.273	2.336	0.481	1.030	0.798	0.809	0.841	0.411
RE(LS)	77	80	79	79	77	93	92	93	93	83
	$p = 6$					$p = 8$				
Mean (LS)	0.009	0.999	1.000	1.002	0.990	0.009	0.011	1.000	0.998	1.000
Mean (MML)	0.013	0.998	1.000	1.002	0.980	0.013	0.011	1.000	0.999	0.999
n × var (LS)	0.932	0.443	0.447	0.454	0.580	0.932	0.875	0.268	0.267	0.266
n × var (MML)	0.866	0.396	0.400	0.411	0.526	0.866	0.786	0.239	0.234	0.238
RE(LS)	93	89	89	90	91	93	90	89	87	89

Table A.1 (Continued)

$n = 50$										
	$p = 2$					$p = 4$				
	γ_0	γ_1	γ_2	γ_3	σ	γ_0	γ_1	γ_2	γ_3	σ
Mean (LS)	0.005	1.002	0.999	0.997	0.993	0.007	0.999	0.999	0.997	0.995
Mean (MML)	0.014	1.003	0.997	1.001	0.966	0.018	0.999	0.999	0.996	0.983
n × var (LS)	2.531	2.734	2.907	2.731	0.600	1.085	0.807	0.829	0.807	0.485
n × var (MML)	1.785	1.987	2.136	1.980	0.457	1.002	0.739	0.768	0.746	0.417
RE(LS)	71	73	73	72	76	92	92	93	92	86
	$p = 6$					$p = 8$				
	γ_0	γ_1	γ_2	γ_3	σ	γ_0	γ_1	γ_2	γ_3	σ
Mean (LS)	0.005	1.003	1.000	1.001	0.994	0.004	0.999	0.999	1.000	0.996
Mean (MML)	0.009	1.003	1.001	1.000	0.986	0.006	1.000	0.999	0.999	0.990
n × var (LS)	0.879	0.400	0.421	0.422	0.553	0.819	0.245	0.254	0.265	0.598
n × var (MML)	0.819	0.362	0.377	0.368	0.503	0.761	0.216	0.219	0.235	0.545
RE(LS)	93	90	89	87	91	93	88	86	89	91

Table A.2: Monte Carlo averages, variances, MSEs and REs for stochastic multiple linear regression; $q = 3$.

		$n = 30$				
		Mean		$n \times$ *Variance*		
Parameter	*True Value*	*LS*	*MML*	*LS*	*MML*	*RE(LS)*
		$p = 4,\ p_1 = 2,\ p_2 = 4,\ p_3 = 6$				
μ_1	0	0.007	0.008	0.4	0.349	87
μ_2	0	0.007	0.004	0.433	0.366	85
μ_3	0	0.012	0.013	0.542	0.497	92
σ_1	1	0.992	0.989	0.559	0.480	86
σ_2	1	0.942	0.944	0.272	0.221	81
σ_3	1	0.943	0.942	0.160	0.139	87
ρ_{21}	0.5	0.480	0.476	0.850	0.786	92
$\rho_{31.2}$	0.5	0.529	0.517	0.268	0.232	87
$\rho_{32.1}$	0.5	0.461	0.473	0.325	0.288	89
γ_0	0	0.028	0.031	5.651	5.227	92
γ_1	1	0.994	0.996	3.431	3.137	91
γ_2	1	0.937	0.949	2.409	2.233	93
γ_3	1	0.849	0.899	5.524	5.145	93
σ	1	0.994	0.991	0.506	0.427	84
		$p = 8,\ p_1 = 2,\ p_2 = 4,\ p_3 = 6$				
μ_1	0	0.011	0.009	0.398	0.368	92
μ_2	0	0.007	0.008	0.430	0.400	93
μ_3	0	0.012	0.014	0.599	0.568	95
σ_1	1	0.990	0.993	0.574	0.515	90
σ_2	1	0.944	0.946	0.282	0.251	89
σ_3	1	0.844	0.893	0.174	0.160	92
ρ_{21}	0.5	0.474	0.473	0.873	0.820	94
$\rho_{31.2}$	0.5	0.526	0.536	0.259	0.222	86
$\rho_{32.1}$	0.5	0.444	0.446	0.326	0.290	89
γ_0	0	0.040	0.038	3.791	3.709	98
γ_1	1	0.985	0.993	1.381	1.318	95
γ_2	1	0.940	0.944	0.914	0.895	98
γ_3	1	0.882	0.894	1.761	1.656	94
σ	1	0.986	0.992	0.643	0.595	92

Table A.2 (Continued)

Parameter	True Value	$n = 50$				
		Mean		$n \times$ *Variance*		
		LS	*MML*	*LS*	*MML*	*RE(LS)*
		$p = 4,\ p_1 = 2,\ p_2 = 4,\ p_3 = 6$				
μ_1	0	0.003	0.004	0.405	0.367	91
μ_2	0	0.004	0.005	0.397	0.348	88
μ_3	0	0.001	0.010	0.575	0.537	93
σ_1	1	0.995	0.995	0.585	0.533	91
σ_2	1	0.947	0.946	0.261	0.243	93
σ_3	1	0.864	0.871	0.167	0.149	89
ρ_{21}	0.5	0.473	0.472	0.850	0.793	93
$\rho_{31.2}$	0.5	0.524	0.511	0.272	0.257	94
$\rho_{32.1}$	0.5	0.448	0.443	0.31	0.300	97
γ_0	0	0.004	0.031	5.415	5.203	96
γ_1	1	0.988	1.000	3.192	2.907	91
γ_2	1	0.941	0.947	2.156	2.004	93
γ_3	1	0.877	0.884	4.776	4.505	94
σ	1	0.993	0.996	0.479	0.461	96
		$p = 8,\ p_1 = 2,\ p_2 = 4,\ p_3 = 6$				
μ_1	0	0.006	0.004	0.412	0.379	92
μ_2	0	0.001	0.003	0.390	0.363	93
μ_3	0	0.005	0.006	0.578	0.503	87
σ_1	1	0.996	0.993	0.552	0.502	91
σ_2	1	0.950	0.947	0.253	0.229	90
σ_3	1	0.845	0.843	0.160	0.139	87
ρ_{21}	0.5	0.469	0.473	0.814	0.769	94
$\rho_{31.2}$	0.5	0.521	0.519	0.254	0.244	96
$\rho_{32.1}$	0.5	0.444	0.444	0.318	0.295	93
γ_0	0	0.020	0.013	3.704	3.318	90
γ_1	1	0.995	0.992	1.254	1.151	92
γ_2	1	0.952	0.960	0.882	0.784	89
γ_3	1	0.881	0.892	1.633	1.452	89
σ	1	0.997	0.996	0.605	0.595	98

Table A.3: Robustness comparisons for multiple linear regression with nonstochastic covariates, $n = 50$, $q = 3$, $\sigma = 1$, $\gamma_0 = 0$ and $\gamma_j = 1$ $(j = 1, 2, .., q)$.

	Model1					*Model2*				
	γ_0	γ_1	γ_2	γ_3	σ	γ_0	γ_1	γ_2	γ_3	σ
Mean (LS)	-0.096	1.001	0.998	1.003	0.991	0.059	1.001	1.000	0.999	0.994
Mean (MML)	-0.101	1.000	0.998	1.002	0.993	0.064	1.000	0.999	0.999	0.996
n × var (LS)	0.763	0.272	0.261	0.250	0.564	0.841	0.254	0.248	0.256	0.640
n × var (MML)	0.674	0.242	0.230	0.220	0.494	0.739	0.221	0.215	0.218	0.546
RE(LS)	88	89	88	88	88	88	87	87	85	85
	Model3					*Model4*				
	γ_0	γ_1	γ_2	γ_3	σ	γ_0	γ_1	γ_2	γ_3	σ
Mean (LS)	-0.073	1.001	0.999	1.000	0.991	0.002	0.998	1.000	0.997	0.987
Mean (MML)	-0.084	1.002	1.000	1.000	0.999	0.000	0.998	1.000	0.997	0.986
n × var (LS)	0.640	0.244	0.249	0.259	0.421	0.882	0.260	0.255	0.249	0.650
n × var (MML)	0.629	0.215	0.214	0.225	0.411	0.810	0.227	0.231	0.224	0.592
RE(LS)	98	88	86	87	98	92	87	90	90	91

Table A.3 (Continued)

Model5

	γ_0	γ_1	γ_2	γ_3	σ
Mean (LS)	-0.091	0.997	1.000	0.999	0.992
Mean (MML)	-0.084	0.997	0.999	1.000	0.995
$n \times var$ (LS)	1.183	0.373	0.367	0.361	0.551
$n \times var$ (MML)	0.901	0.269	0.270	0.271	0.472
RE(LS)	76	72	74	75	85

Table A.4: Robustness comparisons for stochastic multiple linear regression; $n = 50,\ q = 3$.

		Model 1					*Model 2*				
		Mean		$n \times Variance$			*Mean*		$n \times Variance$		
Parameter	*TrueValue*	*LS*	*MML*	*LS*	*MML*	*RE(LS)*	*LS*	*MML*	*LS*	*MML*	*RE(LS)*
μ_1	0	0.002	0.003	0.418	0.380	91	0.005	0.003	0.379	0.352	93
μ_2	0	0.003	0.003	0.499	0.459	92	0.002	0.004	0.512	0.471	92
μ_3	0	0.004	0.006	0.546	0.513	94	0.005	0.005	0.533	0.485	91
σ_1	1	0.997	0.997	0.569	0.512	90	0.993	0.995	0.567	0.494	87
σ_2	1	0.992	0.991	0.575	0.512	89	0.950	0.970	0.280	0.246	88
σ_3	1	0.953	0.944	0.559	0.487	87	0.958	0.968	0.156	0.138	88
ρ_{21}	0.5	0.533	0.532	0.249	0.229	92	0.513	0.512	0.244	0.220	90
$\rho_{31.2}$	0.5	0.510	0.500	0.268	0.241	90	0.523	0.519	0.270	0.246	91
$\rho_{32.1}$	0.5	0.515	0.512	0.419	0.390	93	0.529	0.523	0.394	0.362	92
γ_0	0	-0.091	-0.089	4.833	4.398	91	-0.078	-0.078	4.288	3.988	93
γ_1	1	0.994	0.995	1.552	1.459	94	0.996	0.997	1.323	1.138	86
γ_2	1	0.920	0.939	1.360	1.292	95	0.981	0.989	1.197	1.041	87
γ_3	1	0.949	0.951	1.998	1.838	92	0.958	0.968	1.645	1.464	89
σ	1	0.991	0.993	1.036	0.912	88	0.996	0.998	0.410	0.390	95

Table A.4 (Continued)

		Model 3					Model 4				
		Mean		$n \times$ Variance			Mean		$n \times$ Variance		
Parameter	True Value	LS	MML	LS	MML	RE(LS)	LS	MML	LS	MML	RE(LS)
μ_1	0	0.006	0.004	0.379	0.345	91	0.005	0.003	0.385	0.354	92
μ_2	0	0.007	0.002	0.473	0.431	91	0.003	0.006	0.501	0.451	90
μ_3	0	0.007	0.001	0.513	0.472	92	0.004	0.006	0.561	0.511	91
σ_1	1	0.991	0.995	0.551	0.485	88	0.995	0.995	0.550	0.489	89
σ_2	1	0.976	0.980	0.263	0.229	87	0.960	0.979	0.578	0.514	89
σ_3	1	0.943	0.952	0.158	0.136	86	0.953	0.961	0.556	0.490	88
ρ_{21}	0.5	0.523	0.521	0.243	0.211	87	0.519	0.512	0.265	0.241	91
$\rho_{31.2}$	0.5	0.529	0.528	0.246	0.222	90	0.519	0.516	0.265	0.247	93
$\rho_{32.1}$	0.5	0.523	0.522	0.431	0.371	86	0.522	0.519	0.421	0.387	92
γ_0	0	0.015	0.004	4.095	3.726	91	-0.021	-0.020	4.810	4.329	90
γ_1	1	0.995	0.997	1.341	1.180	88	0.997	0.999	1.699	1.461	86
γ_2	1	0.978	0.982	1.145	1.019	89	0.971	0.985	1.661	1.412	85
γ_3	1	0.943	0.954	1.621	1.427	88	0.964	0.973	1.837	1.598	87
σ	1	0.991	0.996	0.623	0.573	92	1.103	1.100	0.990	0.842	85

APPENDIX B

MATLAB CODE FOR ESTIMATION AND HYPOTHESIS TESTING IN MULTIPLE LINEAR REGRESSION ANALYSIS WITH NONSTOCHASTIC COVARIATES

```
clear all
n=input('n=');
q=input('q=');
sigma=input('sigma=');
p=input('p=');
ga0=input('ga0=');
ga=[];
for i=1:q
    ga(i,1)=input('ga=');
end
gax=[ga0;ga];
LSE_PAR=[];
MML_PAR=[];
LSE_PARx=[];
F_lse=[];
F_mml=[];
aa=0;
bb=0;
for ii=1:100000/n
x=[];
w=[];
y=[];
xx=[];
for j=1:q
    x(:,j)=rand(n,1);
end
xx=[ones(n,1) x];
mean_x=mean(x,1);
mean_xx=mean(xx,1);
w=wblrnd(sigma,p,n,1);
y=ga0+x*ga+w;

%LSE for Multiple Linear Regression Model

ga_lse=(inv(xx'*xx))*(xx'*y);
sum1=0;
for i=1:n
    sum2=0;
```

```
    for j=1:(q+1)
        sum2=sum2+ga_lse(j)*xx(i,j);
    end
    sum1=sum1+(y(i)-sum2)^2;

end
sigma_lse=sqrt(sum1/((n-q-1)*(gamma(1+2/p)-(gamma(1+1/p))^2)));
ga0_lse=ga_lse(1,1)-(gamma(1+1/p))*sigma_lse;
ga_lse(1,1)=ga0_lse;

%MMLE for Multiple Linear Regression Model

alpha1=[];
beta1=[];
alpha2=[];
beta2=[];
eta=[];
delta=[];

for i=1:n
    t(i,1)=(-log(1-i/(n+1)))^(1/p);
    alpha1(i,1)=2*((t(i))^(-1));
    alpha2(i,1)=(2-p)*((t(i))^(p-1));
    beta1(i,1)=(t(i))^(-2);
    beta2(i,1)=(p-1)*((t(i))^(p-2));
    eta(i,1)=(p-1)*beta1(i,1)+p*beta2(i,1);
    delta(i,1)=(p-1)*alpha1(i,1)-p*alpha2(i,1);
end

m=sum(eta);
de=sum(delta);

ga0i=ga_lse(1,1);
gai=[];
for j=1:q
    gai(j,1)=ga_lse(j+1);
end

%iteration

for i=1:3
    e=[];
    data=[];
    for l=1:n
    e(l,1)=y(l)-ga0i-x(l,:)*gai;
    end
    data=[e y x];
    [data1,I]=sort(data(:,1));
    data_ordered=data(I,:);
    e_or=data_ordered(:,1);
    yc=data_ordered(:,2);
    xc=[];
    X=[];
```

```
    K=[];
    L=[];
    for j=1:q
        xc(:,j)=data_ordered(:,j+2);
    end
    mean_yc=(eta'*yc)/m;
    for j=1:q
        mean_xc(1,j)=(eta'*xc(:,j))/m;
    end
    for l=1:n
        for j=1:q
            X(l,j)=xc(l,j)-mean_xc(j);
        end
    end
    for l=1:n
        Y(l,1)=yc(l)-mean_yc;
    end
    K=(inv(X'*(diag(eta))*X))*(X'*(diag(eta))*Y);
    D=(inv(X'*(diag(eta))*X))*(X'*(diag(delta))*ones(n,1));
    B1=[];
    B1=Y-X*K;
    C1=[];
    C1=(Y-X*K).^2;
    B=0;
    for l=1:n
        B=B+(delta(l))*B1(l);
    end
    C=0;
    for l=1:n
        C=C+(eta(l))*C1(l);
    end
    sigma_mml=(-B+sqrt(B^2+4*n*C))/(2*sqrt(n*(n-q-1)));
    ga_mml=K-D*sigma_mml;
    ga0_mml=mean_yc-mean_xc*ga_mml-(de/m)*sigma_mml;
    ga0i=ga0_mml;
    gai=ga_mml;
end
ga0_mml=ga0i;
ga_mml=gai;
LSE_PAR(ii,:)=[ga_lse' sigma_lse];
MML_PAR(ii,:)=[ga0_mml ga_mml' sigma_mml];

ga_lse1=[];

for l=2:q+1
    ga_lse1(l-1,1)=ga_lse(l,1);
end

F_lse(ii)=(ga_lse1'*x'*y)/(q*(sigma_lse^2));
F_mml(ii)=(ga_mml'*X'*diag(eta)*Y)/(q*(sigma_mml^2));

if F_lse(ii)>=finv(0.95,q,n-q-1);
    aa=aa+1;
end
```

```
if F_mml(ii)>=finv(0.95,q,n-q-1)
    bb=bb+1;
end

end
mean_LSE=(mean(LSE_PAR,1))';
mean_MML=(mean(MML_PAR,1))';
var_LSE=n*(var(LSE_PAR,1))';
var_MML=n*(var(MML_PAR,1))';
RE=100*(var_MML./var_LSE);
tablo=[mean_LSE mean_MML var_LSE var_MML RE]';
power_lse=aa/(100000/n);
power_mml=bb/(100000/n);
power=[power_lse power_mml];
```

APPENDIX C

MATLAB CODE FOR ESTIMATION AND HYPOTHESIS TESTING IN STOCHASTIC MULTIPLE LINEAR REGRESSION ANALYSIS

```
clear all
n=input('n=');
p=input('p=');
p1=input('p1=');
p2=input('p2=');
p3=input('p3=');
q=3;
mu1=0;
mu2=0;
mu3=0;
sigma1=1;
sigma2=1;
sigma3=1;
ro21=0.5;
ro31=0.5;
ro32=0.5;
sigma=1;
teta21=ro21*(sigma2/sigma1);
teta31=ro31*(sigma3/sigma1);
teta32=ro32*(sigma3/sigma2);
mu1_I=0;
mu2_I=0;
mu3_I=0;
sigma1_I=1;
sigma2_I=sqrt(0.75);
sigma3_I=sqrt(0.5625);
aa=0;
bb=0;
for j=1:100000/n
w1=wblrnd(sigma1_I,p1,n,1);
x1=w1;
w2=wblrnd(sigma2_I,p2,n,1);
x2=sqrt(0.75)*w2+0.5*x1;
w3=wblrnd(sigma3_I,p3,n,1);
x3=sqrt(0.5625)*w3+0.5*x1+0.5*x2;
u1=(x1-mu1)/sigma1;
u2=(x2-mu2)/sigma2;
u3=(x3-mu3)/sigma3;
X1=x1-mean(x1);
X2=x2-mean(x2);
```

```
X3=x3-mean(x3);
e=wblrnd(sigma,8,n,1);
y=u1+u2+u3+e;
Y=y-mean(y);
Xj1=X1;
Xj2=[X1 X2];

%LSE for Stochastic Multiple Linear Regression

teta21_lse=inv(Xj1'*Xj1)*(Xj1'*x2);
teta_lse=inv(Xj2'*Xj2)*(Xj2'*x3);
teta31_lse=teta_lse(1);
teta32_lse=teta_lse(2);
sigma1_I_lse=sqrt(X1'*X1)/sqrt((n-1)*(gamma(1+2/p1)-
(gamma(1+1/p1))^2));
sigma2_I_lse=sqrt((X2-teta21_lse*X1)'*(X2-teta21_lse*X1))/sqrt((n-
2)*(gamma(1+2/p2)-(gamma(1+1/p2))^2));
sigma3_I_lse=sqrt((X3-teta31_lse*X1-teta32_lse*X2)'*(X3-
teta31_lse*X1-teta32_lse*X2))/sqrt((n-3)*(gamma(1+2/p3)-
(gamma(1+1/p3))^2));
mu1_I_lse=mean(x1)-sigma1_I_lse*gamma(1+1/p1);
mu2_I_lse=mean(x2)-teta21_lse*mean(x1)-sigma2_I_lse*gamma(1+1/p2);
mu3_I_lse=mean(x3)-teta31_lse*mean(x1)-teta32_lse*mean(x2)-
sigma3_I_lse*gamma(1+1/p3);
mu1_lse=mu1_I_lse;
mu2_lse=mu2_I_lse+teta21_lse*mu1_I_lse;
mu3_lse=mu3_I_lse+teta31_lse*mu1_I_lse+teta32_lse*mu2_I_lse;
sigma1_lse=sigma1_I_lse;
sigma2_lse=sqrt(sigma2_I_lse^2+(teta21_lse^2)*(sigma1_I_lse^2));
sigma3_lse=sqrt(sigma3_I_lse^2+(teta31_lse^2)*(sigma1_I_lse^2)+(te
ta32_lse^2)*(sigma2_I_lse^2));
ro21_lse=teta21_lse*(sigma1_lse/sigma2_lse);
ro31_lse=teta31_lse*(sigma1_lse/sigma3_lse);
ro32_lse=teta32_lse*(sigma2_lse/sigma3_lse);
u1_lse=(x1-mu1_lse)/sigma1_lse;
u2_lse=(x2-mu2_lse)/sigma2_lse;
u3_lse=(x3-mu3_lse)/sigma3_lse;
U1_lse=u1_lse-mean(u1_lse);
U2_lse=u2_lse-mean(u2_lse);
U3_lse=u3_lse-mean(u3_lse);
U=[U1_lse U2_lse U3_lse];
ga_lse=inv(U'*U)*(U'*Y);
A=(Y-ga_lse(1)*U(:,1)-ga_lse(2)*U(:,2)-ga_lse(3)*U(:,3));
sigma_lse=sqrt(A'*A)/sqrt((n-q-1)*(gamma(1+2/p)-
(gamma(1+1/p))^2));
ga0_lse=mean(y)-ga_lse(1)*mean(u1_lse)-ga_lse(2)*mean(u2_lse)-
ga_lse(3)*mean(u3_lse)-gamma(1+1/p)*sigma_lse;

%MML for Stochastic Multiple Linear Regession

mu1_Ii=mu1_I_lse;
mu2_Ii=mu2_I_lse;
mu3_Ii=mu3_I_lse;
teta21i=teta21_lse;
```

```
teta31i=teta31_lse;
teta32i=teta32_lse;
for ii=1:3
e1=x1-mu1_Ii;
e2=x2-mu2_Ii-teta21i*x1;
e3=x3-mu3_Ii-teta31i*x1-teta32i*x2;
data1=[e1 x1];
[data2,I]=sort(data1(:,1));
data_ordered1=data1(I,:);
e1_or=data_ordered1(:,1); %ordered e
xc1=data_ordered1(:,2);
data2=[e2 x1 x2];
[data3,J]=sort(data2(:,1));
data_ordered2=data2(J,:);
e1_or=data_ordered2(:,1); %ordered e
xc21=data_ordered2(:,2);
xc22=data_ordered2(:,3);
data3=[e3 x1 x2 x3];
[data4,M]=sort(data3(:,1));
data_ordered3=data3(M,:);
e1_or=data_ordered3(:,1); %ordered e
xc31=data_ordered3(:,2);
xc32=data_ordered3(:,3);
xc33=data_ordered3(:,4);
for i=1:n
        t1(i,1)=(-log(1-i/(n+1)))^(1/p1);
        t2(i,1)=(-log(1-i/(n+1)))^(1/p2);
        t3(i,1)=(-log(1-i/(n+1)))^(1/p3);
        alpha11(i,1)=2*((t1(i,1))^(-1));
        alpha21(i,1)=2*((t2(i,1))^(-1));
        alpha31(i,1)=2*((t3(i,1))^(-1));
        alpha12(i,1)=(2-p1)*((t1(i,1))^(p1-1));
        alpha22(i,1)=(2-p2)*((t2(i,1))^(p2-1));
        alpha32(i,1)=(2-p3)*((t3(i,1))^(p3-1));
        beta11(i,1)=(t1(i,1))^(-2);
        beta21(i,1)=(t2(i,1))^(-2);
        beta31(i,1)=(t3(i,1))^(-2);
        beta12(i,1)=(p1-1)*((t1(i,1))^(p1-2));
        beta22(i,1)=(p2-1)*((t2(i,1))^(p2-2));
        beta32(i,1)=(p3-1)*((t3(i,1))^(p3-2));
        eta1(i,1)=(p1-1)*beta11(i,1)+(p1)*beta12(i,1);
        eta2(i,1)=(p2-1)*beta21(i,1)+(p2)*beta22(i,1);
        eta3(i,1)=(p3-1)*beta31(i,1)+(p3)*beta32(i,1);
        delta1(i,1)=(p1-1)*alpha11(i,1)-(p1)*alpha12(i,1);
        delta2(i,1)=(p2-1)*alpha21(i,1)-(p2)*alpha22(i,1);
        delta3(i,1)=(p3-1)*alpha31(i,1)-(p3)*alpha32(i,1);
end
de1=sum(delta1);
de2=sum(delta2);
de3=sum(delta3);
m1=sum(eta1);
m2=sum(eta2);
m3=sum(eta3);
%j=1
X1=xc1-eta1'*xc1/m1;
```

```
B1=delta1'*X1;
C1=eta1'*(X1.^2);
sigma1_I_mml=(-B1+sqrt(B1^2+4*n*C1))/(2*n);
mu1_I_mml=eta1'*xc1/m1-(de1/m1)*sigma1_I_mml;
mu1_mml=mu1_I_mml;
sigma1_mml=sigma1_I_mml;
%j=2
X21=xc21-eta2'*xc21/m2;
X22=xc22-eta2'*xc22/m2;
K1=inv(X21'*diag(eta2)*X21)*(X21'*diag(eta2)*X22);
D1=inv(X21'*diag(eta2)*X21)*(X21'*diag(delta2)*ones(n,1));
B21=delta2'*(X22-K1*X21);
C21=eta2'*((X22-K1*X21).^2);
sigma2_I_mml=(-B21+sqrt(B21^2+4*n*C21))/(2*n);
teta21_mml=K1-D1*sigma2_I_mml;
mu2_I_mml=eta2'*xc22/m2-teta21_mml*(eta2'*xc21/m2)-
(de2/m2)*sigma2_I_mml;
mu2_mml=mu2_I_mml+teta21_mml*mu1_mml;
sigma2_mml=sqrt(sigma2_I_mml^2+(teta21_mml^2)*sigma1_I_mml^2);
ro21_mml=teta21_mml*(sigma1_mml/sigma2_mml);
%j=3
X31=xc31-eta3'*xc31/m3;
X32=xc32-eta3'*xc32/m3;
X33=xc33-eta3'*xc33/m3;
Xj=[X31 X32];
K2=inv(Xj'*diag(eta3)*Xj)*(Xj'*diag(eta3)*X33);
D2=inv(Xj'*diag(eta3)*Xj)*(Xj'*diag(delta3)*ones(n,1));
B31=delta3'*(X33-K2(1)*X31-K2(2)*X32);
C31=eta3'*((X33-K2(1)*X31-K2(2)*X32).^2);
sigma3_I_mml=(-B31+sqrt(B31^2+4*n*C31))/(2*n);
teta31_mml=K2(1)-D2(1)*sigma3_I_mml;
teta32_mml=K2(2)-D2(2)*sigma3_I_mml;
mu3_I_mml=eta3'*xc33/m3-teta31_mml*(eta3'*xc31/m3)-
teta32_mml*(eta3'*xc32/m3)-(de3/m3)*sigma3_I_mml;
mu3_mml=mu3_I_mml+teta31_mml*mu1_mml-teta32_mml*mu2_mml;
sigma3_mml=sqrt(sigma3_I_mml^2+(teta31_mml^2)*sigma1_I_mml^2+(teta
32_mml^2)*sigma2_I_mml^2);
ro31_mml=teta31_mml*(sigma1_mml/sigma3_mml);
ro32_mml=teta32_mml*(sigma2_mml/sigma3_mml);
mu1_Ii=mu1_I_mml;
mu2_Ii=mu2_I_mml;
mu3_Ii=mu3_I_mml;
teta21i=teta21_mml;
teta31i=teta31_mml;
teta32i=teta32_mml;
end
ga0i=ga0_lse;
gai=ga_lse;
for i=1:n
    t(i,1)=(-log(1-i/(n+1)))^(1/p);
    alpha1(i,1)=2*((t(i))^(-1));
    alpha2(i,1)=(2-p)*((t(i))^(p-1));
    beta1(i,1)=(t(i))^(-2);
    beta2(i,1)=(p-1)*((t(i))^(p-2));
    eta(i,1)=(p-1)*beta1(i,1)+p*beta2(i,1);
```

```
    delta(i,1)=(p-1)*alpha1(i,1)-p*alpha2(i,1);
end
m=sum(eta);
de=sum(delta);
u1i=u1_lse;
u2i=u2_lse;
u3i=u3_lse;
for ii=1:3
    e=[];
    e=y-ga0i-gai(1)*u1i-gai(2)*u2i-gai(3)*u3i;
    data5=[];
    data5=[e y x1 x2 x3];
    [data6,L]=sort(data5(:,1));
    data_ordered5=data5(L,:);
    e_or2=data_ordered5(:,1); %ordered e
    yc=data_ordered5(:,2); %concomitant (y)
    %concomitant (x)
    xc=[];
    xc1=data_ordered5(:,3);
    xc2=data_ordered5(:,4);
    xc3=data_ordered5(:,5);
    u1_mml=(xc1-mu1_mml)/sigma1_mml;
    u2_mml=(xc2-mu2_mml)/sigma2_mml;
    u3_mml=(xc3-mu3_mml)/sigma3_mml;
    U1_mml=u1_mml-eta'*u1_mml/m;
    U2_mml=u2_mml-eta'*u2_mml/m;
    U3_mml=u3_mml-eta'*u3_mml/m;
    Y=yc-eta'*yc/m;
    K=[];
    D=[];
    U_mml=[U1_mml U2_mml U3_mml];
    K=(inv(U_mml'*(diag(eta))*U_mml))*(U_mml'*(diag(eta))*Y);

D=(inv(U_mml'*(diag(eta))*U_mml))*(U_mml'*(diag(delta))*ones(n,1))
;
    B=delta'*(Y-K(1)*U1_mml-K(2)*U2_mml-K(3)*U3_mml);
    C=eta'*((Y-K(1)*U1_mml-K(2)*U2_mml-K(3)*U3_mml).^2);
    sigma_mml=(-B+sqrt(B^2+4*n*C))/(2*sqrt(n*(n-q-1)));
    ga_mml=K-D*sigma_mml;
    ga0_mml=eta'*yc/m-ga_mml(1)*eta'*u1_mml/m-
ga_mml(2)*eta'*u2_mml/m-ga_mml(3)*eta'*u3_mml/m;
    ga0i=ga0_mml;
    gai=ga_mml;
    u1i=u1_mml;
u2i=u2_mml;
u3i=u3_mml;
end
LSE_PAR(j,:)=[mu1_lse mu2_lse mu3_lse sigma1_lse sigma2_lse
sigma3_lse ro21_lse ro31_lse ro32_lse ga0_lse ga_lse' sigma_lse];
MML_PAR(j,:)=[mu1_mml mu2_mml mu3_mml sigma1_mml sigma2_mml
sigma3_mml ro21_mml ro31_mml ro32_mml ga0_mml ga_mml' sigma_mml];
ulse=[u1_lse u2_lse u3_lse];
umml=[u1_mml u2_mml u3_mml];
F_lse(j)=(ga_lse'*ulse'*y)/(q*(sigma_lse^2));
F_mml(j)=(ga_mml'*umml'*diag(eta)*Y)/(q*(sigma_mml^2));
```

```
if F_lse(j)>=finv(0.95,q,n-q-1);
    aa=aa+1;
end
if F_mml(j)>=finv(0.95,q,n-q-1)
    bb=bb+1;
end
end
mean_LSE=(mean(LSE_PAR,1))'; % simulated means of LSEs
mean_MML=(mean(MML_PAR,1))'; % simulated means of MMLs
var_LSE=n*(var(LSE_PAR,1))'; % simulated means of LSEs
var_MML=n*(var(MML_PAR,1))'; % simulated means of MMLs
RE=100*(var_MML./var_LSE);
tablo=[mean_LSE mean_MML var_LSE var_MML RE];
tablo1=tablo';
power_lse=aa/(100000/n)
power_mml=bb/(100000/n)
power=[power_lse power_mml]
```

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Balcı, Sibel
Nationality: Turkish (TC)
Date and Place of Birth: 7 September 1979, Ankara
email: sbalci@metu.edu.tr

EDUCATION

Degree	Institution	Year of Graduation
MS	METU Statistics	2007
BS	Hacettepe University Statistics	2001
High School	Ayrancı Anadolu Lisesi, Ankara	1996

ACADEMIC EXPERIENCE

Year	Place	Enrollment
2004-2014	METU Department of Statistics	Research Assistant

FOREIGN LANGUAGES

Advanced English

PUBLICATIONS

Akkaya A. D. and Balci S. (2014). Chapter Translation in: Bioinformatics for Biologists (Regulatory Network Inference). Translation Editor: Kaya Z., Nobel: Ankara.

Balci S., Akkaya A. D. and Ulgen B. E. (2013). Modified maximum likelihood estimators using ranked set sampling. J. of Comp. and Appl. Math., 238, 171-179.