Fall 12-2021

A Copula Model Approach to Identify the Differential Gene Expression

Prasansha Liyanaarachchi
Old Dominion University, prasanshasjb@gmail.com

A COPULA MODEL APPROACH TO IDENTIFY THE DIFFERENTIAL GENE EXPRESSION

by

Prasansha Liyanaarachchi
B.S. September 2011, University of Peradeniya, Sri Lanka
M.S. May 2015, Sam Houston State University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

COMPUTATIONAL AND APPLIED MATHEMATICS

OLD DOMINION UNIVERSITY
December 2021

Approved by:

N. Rao Chaganty (Director)

Lucia Tabacu (Member)

Sinjini Sikdar (Member)

Hadiza Galadima (Member)

ABSTRACT

A COPULA MODEL APPROACH TO IDENTIFY THE DIFFERENTIAL GENE EXPRESSION

Prasansha Liyanaarachchi
Old Dominion University, 2021
Director: Dr. N. Rao Chaganty

Deoxyribonucleic acid, more commonly known as DNA, is a complex double helix-shaped molecule present in all living organisms and hosts thousands of genes. However, only a few genes exhibit differential expression and play a vital role in a particular disease such as breast cancer. Microarray technology is one of the modern technologies developed to study these gene expressions. There are two major microarray technologies available for expression analysis: Spotted cDNA array and oligonucleotide array. The focus of our research is the statistical analysis of data that arises from the spotted cDNA microarray. Numerous models have been proposed in the literature to identify differentially expressed genes from the red and green intensities measured by the cDNA microarrays. Motivated by the Bayesian models described in Newton et al. (2001) and Mav and Chaganty (2004), we propose two models for the joint distribution of the red and green intensities using a Gaussian copula, which accounts for the dependence. In both models, we assume the marginals are distributed as gamma. The differentially expressed genes were identified by calculating the Bayes estimates of the differential expression under the first proposed copula model. The second copula model incorporates a latent Bernoulli variable, which indicates differential expression. The EM algorithm is applied to calculate the posterior probabilities of differential expression for the second model. The posterior probabilities rank the genes. We conducted two simulation studies to check the parameter estimation for the Gaussian copula-based models. We show that our models improve the models given in Newton et al. (2001) and Mav and Chaganty (2004). We have also studied the use of Weibull distribution instead of gamma distribution for the marginals. Our analysis shows that the copula models with Weibull marginals provide a better fit and improve the identification of genes. Finally, we illustrate the application of our models on samples of *Escherichia coli* microarrays data.

I dedicate this dissertation to my parents, Udaya and Mahesha Liyanaarachchi, my sisters, Prarthana and Aradhana, and my husband Chamara Ranatunga.

ACKNOWLEDGEMENTS

I want to express my deep and sincere gratitude to my advisor Dr. N. Rao Chaganty, for so many reasons. First, this dissertation would not be possible without him. His guidance and patience throughout my dissertation are exceptional. Second, as my dissertation advisor, professor, and program director, the help and support I had from him are enormous, and I am truly thankful. It has been an enriching journey to work under him, and I feel blessed to have him as my advisor.

I should especially thank Dr. Lucia Tabacu, Dr. Sinjini Sikdar, and Dr. Hadiza Galadima for kindly agreeing to serve on my dissertation committee. They are always helpful and flexible, and without their kind cooperation, this work would not be possible or successful.

I am thankful to the faculty, the staff, and dear friends at the Department of Mathematics and Statistics for providing a pleasant and friendly environment. In addition, I am undoubtedly grateful to the Chesapeake Bay Program at ODU for the financial support. Finally, working at the Benthic Ecology lab under the supervision of Micheal Lane was the best thing I earned during my student life. He is more like a family to me, and thank you for being tremendous support throughout my academic and personal life.

I take this opportunity to express my deepest gratitude to my parents, sisters, and in-laws for their love, support, and understanding throughout the years. Of course, I wish my father-in-law could see me today, and I genuinely miss him at this moment. But, more importantly, thank you, Chamara, my husband, who is highly supportive throughout this journey. The sacrifices you have made for me are priceless.

Having a handful of caring friends made my stay in the United States enjoyable and memorable. Thank you for being such a friendly bunch. There are these seven people far away from me who taught me to love myself. Without you, my life would be tedious and stressful. Thank you from the bottom of my heart.

Last but not the least, I would like to express gratitude to all the teachers, professors, and well-wishers I met in my entire life. Thank you for providing me with all the knowledge and moral support I need to face my future.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

BIOLOGICAL BACKGROUND

1.1 INTRODUCTION

Microarray technology is one of the latest approaches used for research fields such as medical science and basic biology. There has been a rapid increment in the number of microarray studies over the last decade. For instance, the number of publications exceeds more than 150,000 in PubMed search for "microarray" in November 2021. In this section, we describe briefly about microarray technology that we use in this dissertation.

1.2 BACKGROUND OF MICROARRAY TECHNOLOGY

There are two major microarray technologies available for expression analysis: Spotted cDNA array and oligonucleotide array. Dr. Patrick Brown and colleagues developed the spotted cDNA microarray at Stanford University in 1995 (Schena et al. (1995)), and the oligonucleotide array was first commercially released (using the trade name GeneChip) in 1996 by Affymetrix Corporation (Santa Clara, CA). For cDNA microarrays, both the targets and probes are the cDNA molecules, while for the oligonucleotide arrays, the targets are cDNA molecules and the probes are well-chosen small segments of cDNA, known as oligos. Thus, even though the primary concern of this dissertation will be the spotted cDNA microarray, the methods illustrated here may be adapted to analyze data produced by the Affymetrix chip.

1.2.1 DNA, PROTEINS AND CENTRAL DOGMA

Here we briefly review basic genetic notions and microarray technology and experiments. An excellent treatise is in the books by Baldi and Hatfield (2002), Schena (2003), and Speed (2003). A complex molecule containing all the information required for an organism to develop, maintain, and reproduce is called Deoxyribonucleic acid, more commonly known as DNA. It is also considered the primary unit of heredity in an organism.

DNA molecule is a double-stranded polymer with a double helix structure consisting of four basic molecular units called nucleotides. They are adenine (A), guanine (G), cytosine

(C), and thymine (T), usually referred to as "bases" (see Figure 1). The nucleotides always pair together in the same way, A with T, C with G. This establishment between bases is called complementary bases.

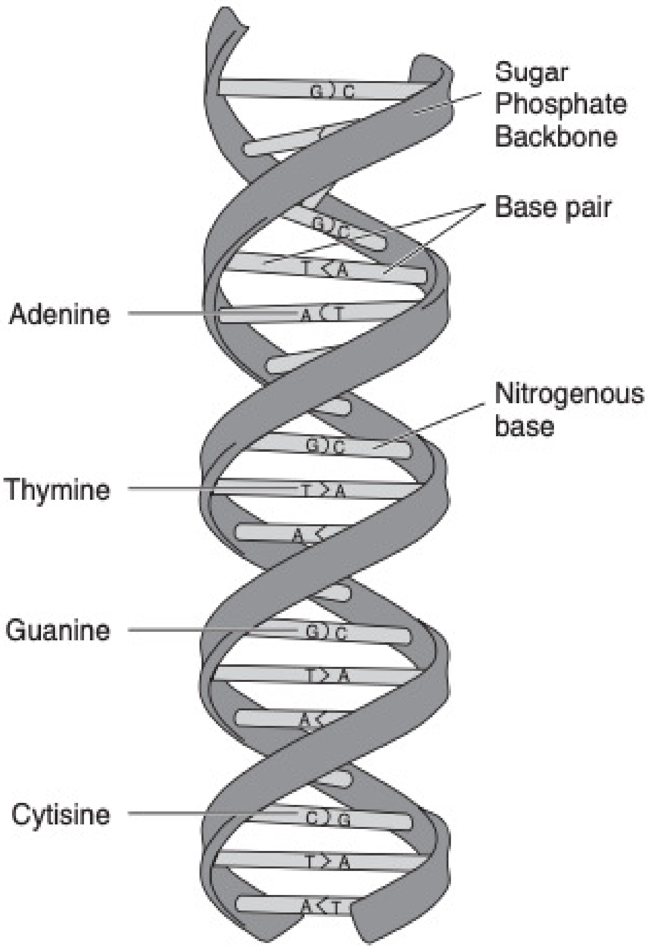


Figure 1. Double standard helix structure of DNA.

A DNA molecule is divided up into functional units called genes. Proteins are the functional products of most known genes. The Central Dogma of Molecular Biology establishes the correspondence between the DNA and the amino acid sequence of a protein.

Many genes provide instructions for building protein, and this process takes place in two stages known as transcription and translation (see Figure 2). During the transcription stage, the information stored in the gene's DNA is transferred to a similar molecule called ribonucleic acid (RNA), and this process is called gene expression. The expression level of a gene indicates the approximate number of copies of RNA, the gene produces in a cell. The type of RNA that consists of the instructions to make a protein is called messenger RNA (mRNA). The translation stage is the process of producing proteins from the instructions stored on an mRNA. This mRNA can be converted into complementary DNA via reverse transcription, which usually serves as samples in microarray experiments.

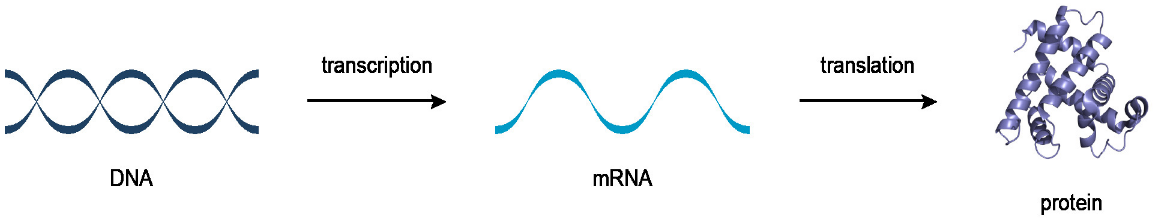


Figure 2. The stages of protein synthesis.

1.2.2 NUCLEIC ACIDS HYBRIDIZATION

A nucleic acid hybridization is a fundamental tool in molecular genetics. Combining two complementary single-stranded nucleic acid molecules and letting them establish a single, double-stranded molecule through base pairing is defined as hybridization. This tool can determine the degree of sequence identity between nucleic acids and can capture the specific sequences.

Hybridization has been used to identify genes in cellular DNA for more than four decades now (Alwine et al. (1977)). Microarrays are based on the same principle but differ in quantity.

While traditional hybridization techniques, such as "Southern blot" can detect one gene at a time, microarrays are intended to do the same with thousands of genes in a single experiment.

1.3 DNA MICROARRAYS

DNA microarray, also known as a gene chip or DNA chip, is a standard laboratory tool for detecting thousands of gene expressions or mutations in a single experiment. In a DNA microarray, thousands of strands of polynucleotide (probes) are located on microscope slides or silicon chips, or nylon membranes. One tiny spot on this slides represents a known DNA sequence or a gene.

These days DNA sequencing technology is used for some tests for which microarrays were used in the past. However, microarray is less expensive than DNA sequencing technology, so they are still used for very large studies and clinical tests.

1.3.1 MICROARRAY TECHNOLOGY

There are two types of microarray experiments: cDNA and oligonucleotide microarrays. First, RNA is extracted from the subject cells to start a microarray experiment. Next, some of its molecules are substituted by others containing a fluorescent dye. The resulting labeled transcripts are called targets. For cDNA microarrays, both the targets and probes are the cDNA molecules, while for the oligonucleotide arrays, the targets are cDNA molecules and the probes are well-chosen small segments of cDNA, known as oligos.

A two-channel array is a term commonly used to refer to a cDNA microarray. In this technique (see Figure 3), samples are prepared from both the experimental sample and a reference sample and labeled using two fluorescent dyes (Cyanine 3 or Cy3 (green) and Cyanine 5 or Cy5 (red)) on a chip. Usually, the experimental sample is labeled with Cy5 (Liu et al. (2010)). There are thousands of spots on a chip, and each spot represents a gene. The brightness of each fluorescent site can be measured using a laser microscope scanner. The colored spots denote genes expressed in one of the samples or may be both, while grey areas reveal the genes expressed in neither type of sample.

In oligonucleotide chips techniques, the Affymetrix system hybridized only one sample per chip (see Figure 4), which means the sample is labeled with one fluorescent dye. This requires more slides per experiment and does not enjoy the advantage of using competitive hybridization; however, it simplifies experimental design and is based on more sensitive technology.

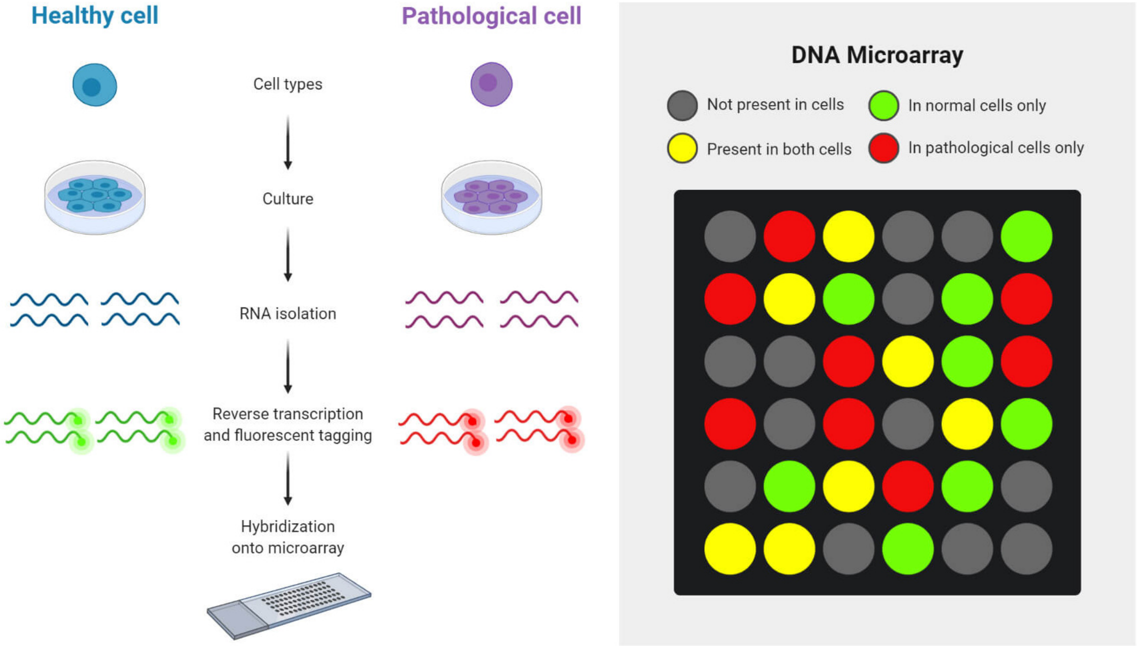


Figure 3. Two color cDNA chip.

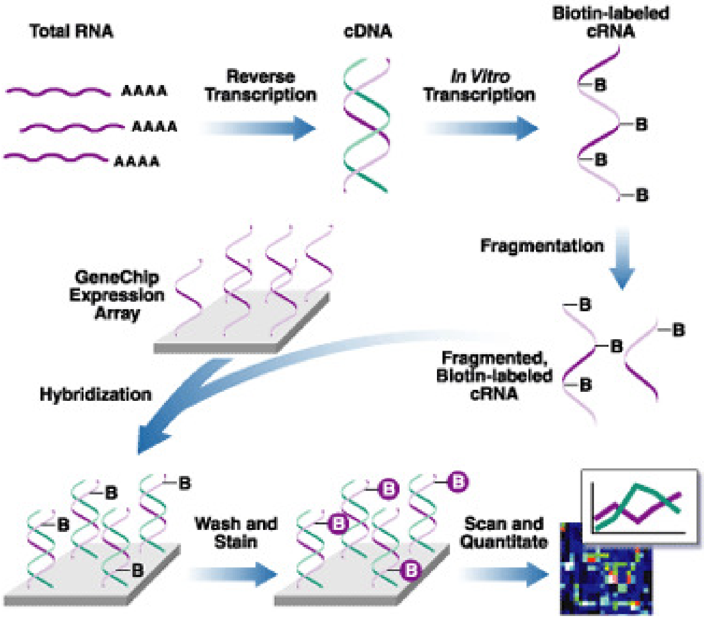


Figure 4. One color affymetrix chips.

1.3.2 PREPROCESSING OF MICROARRAY DATA

A microarray experiment produces a set of images transformed into numerical values representing absolute or relative intensities. Before the analysis, it is necessary to perform some additional operations on the data. Reducing data dimensionality and variability are the two main goals of data preprocessing (Sebastiani et al. (2003)). In this thesis, we mainly focus on the cDNA microarray. Hence the most common preprocessing steps based on cDNA microarray data will be discussed in this section.

- Normalization aims to correct for systematic differences between genes or array. For example, in a two-channel cDNA microarray experiment, several noise sources create recurring sources of biases which causes experimental errors. These experimental errors can be removed using normalization techniques.

- Nonlinear Transformations: Usually, the corrected intensity values are highly skewed. It is common to pass intensity values through a nonlinear function. Log-transforming the raw data is strongly recommended as it usually produces normally distributed data.

- Filtering: Even before or after normalization, it is common to have some genes with negative or small-expression levels. This step can reduce the data dimensionality and variability by removing those gene measurements that are not sufficiently accurate or not adequately differentiated. The elimination of these genes is done if those measurements fail to satisfy some simple criteria. Commonly used criteria include a minimum threshold for the standard deviation of the expression values and a threshold on the maximum percentage of missing values.

1.4 DIFFERENTIALLY EXPRESSED GENES (DEG)

Microarray technology is one of the latest approaches used for research fields such as cancer research, medical science, and basic biology. Basically, microarray data consists of thousands of genes, but a small number of informative differentially expressed genes may be critical elements for a disease such as cancer. Hence it is essential to select those differentially expressed genes out of numerous genes. Several methods for the identification of differentially expressed gene exists in the literature.

1.4.1 SINGLE-SLIDE METHODS

In single-slide experiments, there are two fluorescent intensity measurements (R, G), for each gene or spot, representing the expression level of the gene in the red (Cy5) and green (Cy3) labeled mRNA samples, respectively. Thus, many methods were proposed for the detection of DEG in single-slide cDNA microarray experiments.

The fold-change method is one approach used in the early analysis of microarray data (Schena et al. (1995, 1996); DeRisi et al. (1996)). This simple approach relied on some specified threshold on fold change to capture the DEG. However, under a few conditions, such as data is not correctly normalized, this method is subject to being biased (Sreekumar (2008)), and because of not considering the statistical variation, the procedure is unreliable.

Later the approaches based on probabilistic modeling of R and G were used to find DEGs. In this approach, a rule was derived based on distributional assumptions of (R, G) to identify the differential expression of a gene. Chen et al. (1997) have proposed a data-dependent rule for choosing a threshold for the ratio R/G based on distributional assumptions, including normality and constant coefficient of variation. This method's major drawback is that it has ignored the information contained in the product RG.

To avoid this problem, Newton et al. (2001) suggested a hierarchical model (Gamma–Gamma–Bernoulli) to capture DEGs based on the posterior odds of change (the odds are functions of $R+G$ and RG). This method assumes that R and G are independent and approximately normally distributed.

Mav and Chaganty (2004) have shown that the R and G are positively correlated. To incorporate the dependence, they have built a bivariate distribution with gamma marginals and a positive correlation between R and G. They also incorporated a latent Bernoulli variable. Finally, the EM algorithm was applied to calculate the posterior probabilities. The higher posterior probabilities identify the DEGs.

1.4.2 MULTIPLE-SLIDE METHODS

Statistical methods for identifying DEGs in multiple-slide experiments have dragged little attention relative to the single-slide experiments. However, cluster analysis methods are a common approach that can apply to multiple-slide experiments. In cluster analysis methods, genes are grouped with correlated expression profiles across experimental conditions (Ross et al. (2000); Alizadeh et al. (2000)). Then, the DEGs are identified based on visual inspection of the resulting cluster. Hierarchical clustering, K-means, and SOM's (Self-Organizing Maps) are the most commonly used cluster algorithms. Hierarchical clustering was the first algorithm used in microarray research to cluster genes (Eisen et al. (1998)). We cite the work of Tavazoie et al. (1999) on the K–means algorithm and the work of Tamayo et al. (1999), the first use of SOM's for gene clustering from microarrays. Such methods are called 'unsupervised' since the expression profiles can be clustered together without using covariates or responses for the samples hybridized to the slide.

Supervised methods can further be classified as parametric, nonparametric, and semi-parametric statistical methods. The t-test for two samples is a more direct and appropriate parametric approach that exists in the literature. The two-sample t-statistic is the most common statistic for testing for the mean difference of two samples, and these t-tests may be either equal variance or unequal variance. However, there will always be some genes in the microarray, with small sum of squares across replicates, which leads the absolute t-values to be large regardless of whether their averages are large or not. To avoid this difficulty, Tusher et al. (2001) have proposed a modified t-statistic.

In nonparametric methods, the distribution of random errors is estimated without any parametric assumption. Tusher et al. (2001) have used the method of statistical analysis of microarrays (SAM) to determine the genes with statistically significant changes in expression. A score is assigned to each gene based on a change in gene expression relative to the standard deviation of repeated measures to perform SAM. The genes with scores greater than an adjustable threshold are considered potentially significant. The percentage of such genes determined by chance is the false discovery rate (FDR). This nonparametric approach can be applicable for small sample sizes. The nonparametric Empirical Bayes (EB) was introduced by Efron et al. (2001) to identify DEGs. They have avoided the parametric assumption about gene expression by using a simple nonparametric mixture before modeling the population of affected and unaffected genes. This method allows the analyst to handle multiple testing issues that arise when dealing with many simultaneous tests, establishing a close connection between the estimated posterior probabilities and a local version of the FDR. Lee et al. (2003) applied a nonparametric statistical approach, mixture model method (MMM), as a solution to the unstable selection process of DEGs due to the small sample size and a large number of variables.

Semiparametric models can be much more flexible than parametric models while enjoying the interpretability not shared by nonparametric models. Cox proportional hazards model (Gui and Li (2005); Ma et al. (2009)), Additive risk model (Ma and Huang (2007)) and AFT model (Engler and Li (2009)) are the three most extensively used semiparametric prognosis models to analyze gene expression.

1.5 REAL DATA EXAMPLES

The source for the data is the experiment designed to study gene expression levels in *Escherichia coli (E. coli)*, initially described in Richmond et al. (1999). The *E. coli* genome consists of approximately 4.6 million base pairs (Mbp) but is suspected of encoding only about forty-two hundred genes. To study differential gene expressions in *E. coli*, Richmond et al. (1999) used two traditional treatments which affect gene expression levels. The first treatment is induction with isopropyl-β-D-thiogalactopyranoside (IPTG), which provides a simple test of the methods since only a few gene transcripts are expected to change, and secondly, the Heat Shock treatment, which allows global regulatory effects to be observed. A single colony of *E. coli* K-12 was divided into five samples for the experiments.

IPTG treatment was performed independently on two samples (IPTG-A and IPTG-B), while one sample (control) was untreated. Heat Shock induction was carried out by treating the culture to a 50^0C shaking water bath for seven minutes on the remaining two samples (Heat Shock-A and Heat Shock-B). Following hybridization of the samples on *E. coli* microarrays, signal intensities for each spot were determined using ScanAlyze software. The average fluorescence intensity for each site was measured, and background was determined as the median pixel intensity in a square surrounding each spot. The red and green signal intensities were recalculated and normalized after background subtraction. The *E. coli* data was made publicly available by Newton et al. (2001). We are interested in proposing a Gaussian copula-based joint distribution of red and green intensities.

1.6 OVERVIEW OF THE DISSERTATION

This dissertation proposes and develops two Bayesian Gaussian copula models to identify differentially expressed genes in cDNA microarray. In Chapter 2, we present a brief review of copulas and Gaussian copula as a particular case of interest. Further, we discuss the applications of copulas in different fields in the last section of Chapter 2.

In Chapter 3, we extend the work done in Newton et al. (2001) and Mav and Chaganty (2004) by replacing the joint probability distribution of intensities with a Gaussian copula-based joint distribution. The differentially expressed genes can be identified by calculating the Bayes estimate of the differential expression under this model. Moreover, the relationship between the copula parameter and the linear correlation is derived. In this chapter, we conduct two simulation methods to evaluate the parameter estimation procedure. First, we apply the proposed Gaussian copula model to study the differential gene expressions in *E. coli* (Richmond et al. (1999)). Finally, we show that this model is an improvement over the models given in Newton et al. (2001) and Mav and Chaganty (2004), by comparing the log-likelihood values.

Motivated by the models described in the papers by Newton et al. (2001) and Mav and Chaganty (2004), we propose another Gaussian copula model which incorporates a latent Bernoulli variable, which can be applied to capture differentially expressed genes, in Chapter 4. We use the EM algorithm to calculate the posterior probabilities. The higher posterior probabilities identify the differentially expressed genes. We present two simulation studies to check our parameter estimation methods. The proposed Gaussian copula model with a latent Bernoulli variable is applied on *E. coli* (Richmond et al. (1999)) and a comparison of the log-likelihood values to the model introduced in Mav and Chaganty (2004). We end the chapter by selecting the Gaussian copula model incorporated with a latent Bernoulli variable over the model discussed in Chapter 3 as the best model after studying the AICs for both models.

The model proposed in Chapter 4 uses gamma marginals for the red and green intensities. In Chapter 5, we consider the same model in Chapter 4 but with Weibull marginals. The extreme flexibility of the Weibull distribution allows it to model symmetric, left-skewed, and right-skewed data. We also cover the same topics we covered in the previous chapter.

In Chapter 6, we present a summary of results obtained in this dissertation. Finally, the Appendix section contains important R programs that we developed for this dissertation.

CHAPTER 2

BRIEF INTRODUCTION TO COPULAS

2.1 INTRODUCTION

In this chapter we review the basics concepts of copulas, and discuss the most important copula, namely the Gaussian copula that is related to the multivariate normal distribution. Later, we will use the bivariate Gaussian copula to model the dependence and to construct a joint distribution for the red and green intensities that arise in cDNA microarrays.

2.2 COPULAS

Copula functions are useful for constructing bivariate or in general multivariate distributions with given marginal distributions. The term “copula” was first used by Sklar (1959) meaning that it “ties” the marginal uniform distributions to create a joint distribution function. Since its introduction, the literature and applications of the copulas has grown rapidly. Some classic books on the topic include Joe (1997), Nelsen (1996), and Nelsen (2006)). Joe (1997) investigated the dependence concepts for bivariate and multivariate random variables. He discussed the fundamental properties of the bivariate and multivariate copulas. A more comprehensive coverage of copula models and their applications is given in Joe (2015).

Definition 1. *A d-dimensional copula is a function $C : [0,1]^d \to [0,1]$ with the following properties:*

1. $C(1,\ldots,1,u_i,1,\ldots,1) = u_i, \forall\ i = 1,2,\ldots,d$ and $u_i \in [0,1]$.

2. $C(u_1,u_2,\ldots,u_d) = 0$ if at least one $u_i = 0$ for $1 \le i \le d$.

3. For any $u_{i_1}, u_{i_2} \in [0,1]$ with $u_{i_1} \le u_{i_2}$, for $i = 1,2,\ldots,d$,

$$\sum_{j_1=1}^{2}\sum_{j_2=1}^{2}\cdots\sum_{j_d=1}^{2}(-1)^{j_1+j_2+\cdots+j_d}C(u_{1j_1},u_{2j_2}\ldots,u_{dj_d}) \ge 0.$$

2.2.1 EXAMPLES OF COPULAS

There are numerous copulas available in the literature. Some well known copulas are listed below.

Example 1. *The first and simplest is the Independence Copula given by*

$$C(u_1, u_2, \dots, u_d) \;\; = \;\; \prod_{j=1}^{d} u_j. \tag{1}$$

Example 2. *The Multivariate Gaussian Copula with latent correlation matrix $\boldsymbol{R}$ is a function given by*

$$C(u_1, u_2, \dots, u_d; \boldsymbol{R}) \;\; = \;\; \Phi_d(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \dots, \Phi^{-1}(u_d); \mathbf{0}, \boldsymbol{R}), \tag{2}$$

where Φ is the cumulative distribution function of standard normal and $\Phi_d(.; \boldsymbol{\mu}, \Sigma)$ is the cumulative distribution function of a d-variate normal with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Note that d-dimensional Gaussian copula reduces to the Independence Copula when $\Sigma = I$, the identity matrix.

Theorem 1. *(Sklar's Theorem). Let $X_1, X_2, \dots, X_d$ be random variables with marginal distribution functions $F_1, F_2, \dots, F_d$ respectively. Suppose F is joint cumulative distribution function.*

1. *Then there exists a function C such that for all $x_1, x_2, \dots, x_d \in (-\infty, \infty)$*

   $$F(x_1, x_2, \dots, x_d) = C(F_1(x_1), F_2(x_2), \dots, F_d(x_d)), \tag{3}$$

   Conversely, if $u_i = F_i(x_i)$ then $x_i = F_i^{-1}(u_i)$, and the copula function can be extracted from (3) as

   $$C(u_1, u_2, \dots, u_d) = F(F_1^{-1}(u_1), F_2^{-1}(u_2), \dots, F_d^{-1}(u_d)). \tag{4}$$

2. *If $X_1, X_2, \dots, X_d$ are continuous random variables defined on real line, then C is unique. Otherwise, C is uniquely determined on the d-dimensional rectangle $Range(F_1) \times Range(F_2) \times Range(F_d)$.*

Equations (3), (4) are the basis for the construction of multivariate distributions using copulas.

2.2.2 MULTIVARIATE PROBABILITY DENSITY FUNCTIONS

Suppose F_i is the marginal cumulative distribution function of $X_i, i = 1, 2, \ldots, d$. For a copula model, the joint cumulative distribution function for the vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$ is given by

$$F(\boldsymbol{x}) = C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)), \tag{5}$$

where C a d-dimensional copula. If $\boldsymbol{X}$ is continuous then its probability density function is

$$f(\boldsymbol{x}) = c(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) \prod_{i=1}^{d} f_i(x_i), \tag{6}$$

where $f_i(x)$ is the marginal probability density function of X_i and

$$c(u_1, u_2, \ldots, u_d) = \frac{\partial^d C(u_1, u_2, \ldots, u_d)}{\partial u_1 \, \partial u_2 \, \ldots \, \partial u_d},$$

is the density of the copula C. On the other hand, if $\boldsymbol{X}$ is a discrete random vector then the d-dimensional joint probability mass function is given by

$$f(x_1, x_2, \ldots, x_d) = \sum_{j_1=1}^{2} \sum_{j_2=1}^{2} \cdots \sum_{j_d=1}^{2} (-1)^{j_1+j_2+\cdots+j_d} C(u_{1j_1}, u_{2j_2}, \ldots, u_{dj_d}), \tag{7}$$

where $u_{i1}(x_i) = F_i(x_i{}^-)$ and $u_{i2}(x_i) = F_i(x_i)$. $F_i(x_i{}^-)$ is the left hand limit of F_i at x_i.

2.3 BIVARIATE COPULA DISTRIBUTIONS

Here we present some examples of copulas in the bivariate case ($d = 2$). Some of these have natural extensions to the multivariate case. The first and simplest is the independent copula given by $C(u_1,\, u_2) = u_1\, u_2$, $0 \leq u_i \leq 1$ for $i = 1, 2$. Clearly this corresponds to the case where the two uniformly distributed randomly variables are independent. Next, a very popular copula is the bivariate Gaussian copula. It is given by

$$C(u_1,\; u_2;\; \gamma) \;\; = \;\; \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2);\; \gamma), \quad u_i \in [0, 1] \;\text{ for }\; i = 1, 2, \tag{8}$$

where Φ is the cumulative distribution function of standard normal and Φ_2 is the cumulative distribution function of a standard bivariate normal distribution with correlation γ. Taking the partial derivatives of (8) we get the probability density function of the Gaussian copula as

$$c(u_1, u_2) \;\; = \;\; \frac{1}{\sqrt{1-\gamma^2}} \; \exp\left[-\frac{1}{2} \left(\frac{\gamma^2 (z_1^2 + z_2^2) - 2\,\gamma\, z_1\, z_2}{1-\gamma^2} \right) \right], \tag{9}$$

where $z_i = \Phi^{-1}(u_i)$, for $i = 1,\, 2$.

Let R_1 and R_2 be two non-negative random variables with cumulative distribution functions $F_1(r_1)$ and $F_2(r_2)$ respectively. A copula based joint cumulative distribution function for R_1 and R_2 is given by

$$F(r_1,\, r_2) = C(F_1(r_1),\, F_2(r_2);\; \gamma), \;\; \text{for} \;\; r_1 > 0,\, r_2 > 0, \tag{10}$$

where C is the Gaussian copula given in (8). The density function is given by

$$\begin{aligned} f(r_1,\, r_2) &= c(F_1(r_1),\, F_2(r_2);\; \gamma)\, f_1(r_1)\, f_2(r_2) \\ &= c(u_1,\, u_2;\, \gamma)\, f_1(r_1)\, f_2(r_2), \end{aligned} \tag{11}$$

where $u_i = F_i(r_i)$, $c(u_1,\, u_2;\, \gamma)$ as in (9) and $f_i(r_i)$ is the probability density function of R_i.

2.4 APPLICATIONS OF COPULA IN DIFFERENT FIELDS

Copula models are an important and vigorously growing modeling tools applicable in many fields where the main interest is the dependence between random variables of any type. For example, copulas were widely used in the field of finance. The approach of Clayton canonical vine copula to analyze systemic risk in financial markets by Low (2018) is a recent example of the financial application of copulas.

Engineering is another major field that has successfully employed copula functions. Some applications of copulas can be found in the paper by Yang et al. (2017) on the reliability of tower and tower-line systems under spatiotemporally changing wind or earthquake loads. Zhang et al. (2015) used copulas to study on long-term performance assessment and design of offshore structures.

Copulas have had a growing impact in the field of meteorology and climate research lately. Numerous successful applications can be found over the last decade in the climate research field. For example, Mesbahzadeh et al. (2019) has discussed copulas for joint modeling of precipitation and temperature, which are two main climatic factors impacting agricultural production, meteorological and hydrological phenomena. Cong and Brady (2012) presents a copula modeling framework to model the interdependence of rainfall and temperature. Few of the copula-based approaches can be found in the field of Geodesy. As an example, Modiri et al. (2018, 2020) have combined copula with singular spectrum analysis for polar motion prediction and to improve the accuracy of the forecasted length of day.

Bayesian nonparametric conditional copula estimation has been used to analyze the influence of socioeconomic status on the relationship between twins' cognitive abilities. See Valle et al. (2017) for examples of copula applications in social sciences.

Copulas are also being used in the field of medicine. For example, a high dimensional latent Gaussian copula model for mixed data in imaging genetics (Zhang et al. (2018)) is an excellent example of the copula in the field of magnetic resonance imaging (MRI). Brain research (Qian et al. (2017)) and oncology (Bao et al. (2009)) are some other areas of medicine where copulas are used.

Bioinformatics is another field where copula methods that have been widely applied. Owzar et al. (2007) have incorporated copulas to detect prognostic genes associated with survival outcomes in microarray studies. Yuan et al. (2008) proposed a semiparametric copula method for microarray-SNP genomewide association analysis using pedigree data. A unified copula VC approach that allows the analysis of traits with a variety of distributions was developed by Li et al. (2006). Escarela and Carriere (2003) proposed a fully parametric model for the analysis of competing risks data where the types of failure may not be independent. They have shown that with the proposed copula model, more accurate inferences can be obtained than using a simpler model. Most recently, Kasa et al. (2020) have published a paper about Gaussian mixture copulas for high-dimensional clustering and dependency-based subtyping.

Only a few works in the literature demonstrate copula methods applications in microarray data for gene selection. For instance, Chaba (2006) has developed a semi-parametric copula-based algorithm for gene selection that does not depend on the distributions of the covariates. They assumed marginal distributions are continuous and have validated the result in a melanoma dataset. Furthermore, a clustering algorithm based on copula functions on microarray data, called 'CoClust' was proposed by Di Lascio (2008). This dissertation addresses the need for a copula-based approach in microarray data to identify differentially expressed genes, which has not been addressed so far in the scientific literature.

CHAPTER 3

BAYESIAN COPULA MODEL

3.1 INTRODUCTION

In Chapter 2, we presented a brief review of copulas. This chapter develops a Gaussian copula-based model for the joint distribution of cDNA microarrays' red and green intensities. As an application, we apply the model to real data sets to identify differentially expressed genes.

3.2 MOTIVATION

Escherichia coli (E. coli) is a bacteria that generally live in the intestines of people and animals. The source of data for this dissertation is the experiment designed to study gene expression levels in *E. coli*, initially described in Richmond et al. (1999). The *E. coli* genome consists of approximately 4.6 million base pairs (Mbp) but is suspected of encoding only about forty-two hundred genes. To study differential gene expressions in *E. coli*, Richmond et al. (1999) used two traditional treatments which affect gene expression levels. The first treatment is induction with isopropyl-β-D-thiogalactopyranoside (IPTG), which provides a simple test of the methods since only a few gene transcripts are expected to change, and secondly, the Heat Shock treatment, which allows global regulatory effects to be observed. A single colony of *E. coli* K-12 was divided into five samples for the experiments.

IPTG treatment was performed independently on two samples (IPTG-A and IPTG-B), while one sample (control) was untreated. Heat Shock induction was carried out by treating the culture to a 50^0C shaking water bath for seven minutes on the remaining two samples (Heat Shock-A and Heat Shock-B). Following hybridization of the samples on *E. coli* microarrays, signal intensities for each spot were determined using ScanAlyze software. The average fluorescence intensity for each spot was measured, and background was chosen as the median pixel intensity in a square surrounding each spot. The red and green signal intensities were recalculated and normalized after background subtraction. The *E. coli* data was made publicly available by Newton et al. (2001).

Newton et al. (2001) have proposed a Bayesian hierarchical model with a latent variable to identify differentially expressed genes. Here the marginal distributions of red and green intensities were modeled as gamma distributions with common shape parameter but different scale parameters. Newton et al. (2001) assumed that the red and green intensities measured on the same gene are independent. They have applied the suggested hierarchical model on *E. coli* data.

Figure 7 contains the scatter plots of red (R_1) and green (R_2) intensities. Clearly the uncorrelated assumption is not true. Mav and Chaganty (2004) have remodeled the red and green intensities by a bivariate distribution with gamma marginals and a positive correlation between the variables. By applying the new model on the same *E. coli* data, they have shown that their model is an improvement over the model given in the Newton et al. (2001).

In this chapter, we extend and replace the bivariate distribution in the Bayesian model proposed by Mav and Chaganty (2004) with Gaussian Copula joint distribution with gamma marginals. The performance of our extended model in terms of log-likelihood analysis is assessed via applying on *E. coli* data.

3.3 BAYESIAN COPULA MODEL FOR EXPRESSION LEVEL

The typical objective when analyzing data arising from microarray experiments is to identify genes that are differentially expressed. In this section, we will propose a Bayesian copula model that can filter the differentially expressed genes.

Consider a microarray consisting of n genes. Let R_{1j} and R_{2j} denote the red and green intensities of gene j, respectively. In literature, the concepts based on the red and green intensity ratio have been widely used to identify differentially expressed genes. Some of those were discussed briefly in section 1.4.1. To filter the differentially expressed genes, we will use the ratio of expected expression levels which are given by $\eta_j = E(R_{1j})/E(R_{2j})$ for $j = 1, \ldots, n$.

As explained in section 3.2, this study is mainly based on the *E. coli* data. Empirical plots show the gamma distribution is an appropriate model for the marginal distributions of red and green intensities. For the model simplicity purposes we assume R_{ij} and R_{2j} are gamma distributions with common shape parameterα but different scale parameters $1/\theta_{1j}$ and $1/\theta_{2j}$ for $j = 1, 2, \ldots, n$. The probability density function of R_{ij} is given by

$$f_i(r_{ij}; \theta_{ij}, \alpha) \;=\; \frac{1}{\Gamma(\alpha)}\, \theta_{ij}^{\,\alpha}\, r_{ij}^{\alpha-1}\, \exp\left(-\theta_{ij}\, r_{ij}\right), \qquad i = 1, 2;\; j = 1, \ldots, n. \tag{12}$$

To model the dependence between two intensities, we assume the joint distribution of (R_{1j}, R_{2j}) is given by the bivariate Gaussian copula (11) , which can be written as

$$f(r_{1j},\, r_{2j}; \theta_{1j}, \theta_{2j}, \alpha, \gamma) \;=\; c(u_{1j},\, u_{2j};\, \gamma)\; f_1(r_{1j})\; f_2(r_{2j}), \tag{13}$$

where $u_{ij} = F_i(r_{ij})$ and $F_i(.)$ is the cumulative distribution function of a gamma distribution with parameters $(\alpha, 1/\theta_{ij})$. Note that

$$c(u_{1j},\, u_{2j};\, \gamma) \;=\; \frac{1}{\sqrt{1-\gamma^2}}\, \exp\left[-\frac{1}{2}\left(\frac{\gamma^2(z_{1j}^2+z_{2j}^2)-2\gamma z_{1j}z_{2j}}{1-\gamma^2}\right)\right], \tag{14}$$

where $z_{ij} = \Phi^{-1}(F_i(r_{ij})) = \Phi^{-1}(u_{ij})$ and γ is the parameter for the copula density. To simplify the notation, we write $c(u_{1j},\, u_{2j})$ in place of $c(u_{1j},\, u_{2j}; \gamma)$ from now on.

The model stated in (13) consists of $2n+2$ unknown parameters that needs to be estimated. Since there are too many unknown parameters, we adopt the empirical Bayes approach to make the model parsimonious. This requires specification of prior distributions for the gene specific parameters θ_{1j} and θ_{2j}'s. We assume independent gamma distributions with parameters α_0 and $1/\nu$ as the prior distributions for θ_{ij}'s. The prior density $\pi(\theta_{ij})$ is

$$\pi(\theta_{ij}; \nu, \alpha_0) \;=\; \frac{1}{\Gamma(\alpha_0)}\, \nu^{\,\alpha_0}\, \theta_{ij}^{\alpha_0-1}\, \exp\left(-\nu\theta_{ij}\right), \;\text{ for }\; i=1,2;\; j=1,\ldots,n. \tag{15}$$

Multiplying (13) and (15) we get the joint density of $(R_{1j},\, R_{2j})$ and $(\theta_{1j},\, \theta_{2j})$ as

$$f(r_{1j},\, r_{2j}, \theta_{1j},\, \theta_{2j}; \Upsilon) \;=\; \left(\frac{\nu^{\alpha_0}}{\Gamma(\alpha)\,\Gamma(\alpha_0)}\right)^2 c(u_{1j}, u_{2j}) \prod_{i=1}^{2}\left[r_{ij}^{\alpha-1}\,\theta_{ij}^{\alpha+\alpha_0-1}\, \exp\left(-\theta_{ij}\left(r_{ij}+\nu\right)\right)\right], \tag{16}$$

where $\boldsymbol{\Upsilon} = (\alpha, \alpha_0, \nu, \gamma)$ is the vector of model parameters. Recall, this model has gene specific parameters $(\theta_{1j}, \theta_{2j})$ for $j = 1, \ldots, n$. The marginal density of $\mathbf{R}_j = (R_{1j},\, R_{2j})$ is

$$\begin{aligned} f_m(r_{1j},\, r_{2j}; \boldsymbol{\Upsilon}) \;&=\; \int_0^\infty \int_0^\infty f(r_{1j},\, r_{2j}; \theta_{1j}, \theta_{2j}; \boldsymbol{\Upsilon})\; d\theta_{1j}\; d\theta_{2j} \\ &=\; \left(\frac{\nu^{\alpha_0}}{\Gamma(\alpha)\,\Gamma(\alpha_0)}\right)^2 \int_0^\infty \int_0^\infty c(F_1(r_{1j}), F_2(r_{2j}))\; \times \\ &\qquad \prod_{i=1}^{2}\left[r_{ij}^{\alpha-1}\,\theta_{ij}^{\alpha+\alpha_0-1}\, \exp\left(-\theta_{ij}\left(r_{ij}+\nu\right)\right)\right]\, d\theta_{1j}\, d\theta_{2j}. \end{aligned} \tag{17}$$

Here $F_i(r_{ij})$ is the cumulative distribution function of gamma with parameters α and $1/\theta_{ij}$ for $i = 1, 2$ and $j = 1, 2, \ldots, n$.

The double integral in equation (17) does not simplify because of the presence of the Gaussian copula function $c(u_{1j}, u_{2j})$ in the integrand. Numerical computation of (17) is also challenging. To compute the double integral we could use the R libraries such as cubature by Narasimhan et al. (2021) or pracma by Borchers (2021). We were not successful with these packages and encountered numerous errors with the functions embedded in these packages to evaluate the double integral iteratively. To overcome the computational problems we have developed our own R code to evaluate the double integral and obtain the marginal density of $(R_{1j},\, R_{2j})$. This R code is given in Appendix A.

3.4 PARAMETER ESTIMATION PROCEDURE

The marginal bivariate density of red and green intensities given in (17) has four unknown parameters given by the vector $\boldsymbol{\Upsilon} = (\alpha, \alpha_0, \nu, \gamma)$. The maximum likelihood is the efficient method for estimating these parameters. This method entails maximizing the likelihood or alternatively the log-likelihood, which is the logarithm of the likelihood function. For n genes the log-likelihood is given by

$$\begin{aligned} l\,(\boldsymbol{\Upsilon}) &= \sum_{j=1}^{n} \log f_m(r_{1j}, r_{2j}; \boldsymbol{\Upsilon}) \\ &= \sum_{j=1}^{n} \log \left[\left(\frac{\nu^{\alpha_0}}{\Gamma(\alpha)\,\Gamma(\alpha_0)} \right)^2 \int_0^\infty \int_0^\infty c(F_1(r_{1j}), F_2(r_{2j})) \times \right. \\ &\qquad \left. \prod_{i=1}^{2} \left[r_{ij}^{\alpha-1}\, \theta_{ij}^{\alpha+\alpha_0-1}\, \exp\left(-\theta_{ij}\,(r_{ij}+\nu)\right) \right] d\theta_{1j}\, d\theta_{2j} \right] \\ &= 2n \left[\alpha_0\, \log(\nu) \,-\, \log\left(\Gamma(\alpha)\,\Gamma(\alpha_0)\right)\right] \,+\, \sum_{j=1}^{n} \log \left[\int_0^\infty \int_0^\infty c(F_1(r_{1j}), F_2(r_{2j})) \times \right. \\ &\qquad \left. \prod_{i=1}^{2} \left[r_{ij}^{\alpha-1}\, \theta_{ij}^{\alpha+\alpha_0-1}\, \exp\left(-\theta_{ij}\,(r_{ij}+\nu)\right) \right] d\theta_{1j}\, d\theta_{2j} \right]. \qquad (18) \end{aligned}$$

Maximizing (18) will yield the maximum likelihood estimate of the unknown parameter vector $\boldsymbol{\Upsilon}$.

3.4.1 ESTIMATION

A numerical optimization routine is required to obtain the maximum likelihood estimator of $\boldsymbol{\Upsilon} = (\alpha, \alpha_0, \nu, \gamma)$, since the log-likelihood (18) is highly nonlinear. The quasi-Newton (or variable metric) algorithm given in Nash (1979) is an ideal choice for this situation. The algorithm can be described as follows:

Step 1 Start with an initial estimate $\widehat{\boldsymbol{\Upsilon}}_{int}$ of $\boldsymbol{\Upsilon}$.

Step 2 At the i th step compute $\widehat{\boldsymbol{\Upsilon}}_{i+1} = \widehat{\boldsymbol{\Upsilon}}_i - c\,B(\widehat{\boldsymbol{\Upsilon}}_i)g(\widehat{\boldsymbol{\Upsilon}}_i)$ where $g(\boldsymbol{\Upsilon}) = \partial l(\boldsymbol{\Upsilon})/\partial\boldsymbol{\Upsilon}$ and $B(\boldsymbol{\Upsilon})$ is an approximation to the inverse of Hessian matrix, $[\partial^2 l(\boldsymbol{\Upsilon})/\partial\Upsilon_j\partial\Upsilon_k]^{-1}$, and c is a constant.

Step 3 Repeat Step 2 until $\widehat{\boldsymbol{\Upsilon}}_{i+1} \cong \widehat{\boldsymbol{\Upsilon}}_i$ and take $\widehat{\boldsymbol{\Upsilon}} = \widehat{\boldsymbol{\Upsilon}}_{i+1}$ as the MLE of $\boldsymbol{\Upsilon}$.

The function optim in the R package stats provides algorithms for general purpose optimization. We used the quasi-Newton method "BFGS", which was published simultaneously by Broyden (1970); Fletcher (1970); Goldfarb (1970); Shanno (1970). The estimation of gradient function is carried out using finite-difference approximation. The Hessian matrix is the square matrix of second order partial derivatives given by

$$\frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\boldsymbol{\Upsilon}\partial\boldsymbol{\Upsilon}'} = \begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\alpha^2} & \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\alpha\partial\alpha_0} & \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\alpha\partial\nu} & \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\alpha\partial\gamma} \\ \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\alpha_0\partial\alpha} & \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial{\alpha_0}^2} & \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\alpha_0\partial\nu} & \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\alpha_0\partial\gamma} \\ \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\nu\partial\alpha} & \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\nu\partial\alpha_0} & \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\nu^2} & \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\nu\partial\gamma} \\ \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\gamma\partial\alpha} & \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\gamma\partial\alpha_0} & \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\gamma\partial\nu} & \frac{\partial^2 l(\boldsymbol{\Upsilon})}{\partial\gamma^2} \end{pmatrix}.$$

This matrix can be calculated numerically at the point of maximum of the log-likelihood function using the method "Richardson" of function Hessian in the R package numDeriv by Gilbert and Varadhan (2019). The square-root of the diagonal elements of inverse Hessian gives us the standard errors of the maximum likelihood estimates.

3.5 DIFFERENTIALLY EXPRESSED GENES

Our ultimate goal of modeling using (17) is to identify the differentially expressed genes in cDNA microarray. Recall that we are interested in estimating $\eta_j = E(R_{1j})/E(R_{2j}) = \theta_{2j}/\theta_{1j}$ for $j = 1, \ldots, n$. Consider the transformation $\mu_j = \theta_{1j}$ and $\eta_j = \theta_{2j}/\theta_{1j}$. The inverse transformation is $\theta_{1j} = \mu_j$ and $\theta_{2j} = \eta_j\,\mu_j$ and the Jacobian is given by

$$J = \begin{vmatrix} \mu_j & \eta_j \\ 0 & 1 \end{vmatrix} = \mu_j.$$

By the transformation theorem the joint density of $\mathbf{R}_j$, η_j and μ_j is given by

$$g(r_{1j},\, r_{2j}, \eta_j,\, \mu_j; \boldsymbol{\Upsilon}) \;=\; f(r_{1j},\, r_{2j}, \mu_j,\, \eta_j\mu_j; \boldsymbol{\Upsilon})\,\mu_j\,, \qquad \eta_j\,,\, \mu_j > 0.$$

The conditional posterior distribution of η_j and μ_j given $\mathbf{R}_j$ is

$$g(\eta_j,\, \mu_j | r_{1j},\, r_{2j}; \boldsymbol{\Upsilon}) = \frac{f(r_{1j},\, r_{2j}, \mu_j,\, \eta_j\mu_j; \boldsymbol{\Upsilon})\,\mu_j}{f_m(r_{1j},\, r_{2j}; \boldsymbol{\Upsilon})}\,, \qquad \eta_j\,,\, \mu_j > 0.$$

The Bayes estimate of the differential expression of the jth gene is

$$E(\eta_j|r_{1j},\, r_{2j};\boldsymbol{\Upsilon}) \;=\; \int_0^\infty\int_0^\infty \eta_j\, \frac{f(r_{1j},\, r_{2j}, \mu_j,\, \eta_j\mu_j;\boldsymbol{\Upsilon})\,\mu_j}{f_m(r_{1j},\, r_{2j};\boldsymbol{\Upsilon})}\, d\eta_j\, d\mu_j, \tag{19}$$

which we can calculate numerically. Let $\widetilde{\eta_j} = E(\eta_j|r_{1j},\, r_{2j};\widehat{\boldsymbol{\Upsilon}})$, where $\widehat{\boldsymbol{\Upsilon}}$ is the maximum likelihood estimate of $\boldsymbol{\Upsilon}$. We say the j th gene is up-regulated if $\widetilde{\eta_j}$ is greater than some specified value and down-regulated if it is less than that value.

3.6 RELATION BETWEEN COPULA PARAMETER AND CORRELATION COEFFICIENT

It is a well known fact that correlation coefficient of two random variables is the magnitude and the direction of the linear relationship between those two random variables. However it fails to capture nonlinear dependence. But the copula function is able to capture nonlinear dependence, specifically, dependence in the tail region for non-normal variables. In this section we will derive the relationship between linear correlation coefficient ρ between R_1 and R_2 and the Gaussian copula parameter γ.

Case 1. Suppose that R_i is distributed as gamma$(\alpha_i, 1/\theta_i)$ for $i = 1, 2$ and the joint distribution is given by the bivariate Gaussian copula with parameter γ. Note that the marginal mean and variance of R_i are α_i/θ_i and $\alpha_i/{\theta_i}^2$ respectively. The joint probability density function of $(R_1,\, R_2)$ is given by

$$\begin{aligned} f(r_1,\, r_2;\theta_1,\theta_2,\alpha_1,\alpha_2,\gamma) \;&=\; c(u_1,u_2)\; f_1(r_1)\; f_2(r_2) \\ &=\; \frac{1}{\sqrt{1-\gamma^2}}\, \exp\left[-\frac{1}{2}\left(\frac{\gamma^2(z_1^2+z_2^2) - 2\,\gamma\, z_1\, z_2}{1-\gamma^2}\right)\right] \\ &\qquad \times \prod_{i=1}^{2} \frac{1}{\Gamma(\alpha_1)}\, r_i^{\alpha_i-1}\,\theta_i^{\alpha_i}\, \exp\left(-\theta_i r_i\right), \end{aligned}$$

where $z_i = \Phi^{-1}(u_i)$, for $i = 1, 2$ and $u_i = F_i(r_i)$, and F_i is the cumulative distribution function. Therefore the expected value of R_1R_2 is

$$E[R_1R_2] = \int_0^\infty\int_0^\infty r_1r_2\; f(r_1,\, r_2;\theta_1,\theta_2,\alpha_1,\alpha_2,\gamma)\; dr_1\, dr_2.$$

If ρ is the correlation coefficient between R_1 and R_2 then we have

$$\rho = \frac{\theta_1\theta_2}{\sqrt{\alpha_1\alpha_2}}\int_0^\infty\int_0^\infty r_1r_2\; f(r_1,\, r_2;\theta_1,\theta_2,\alpha_1,\alpha_2,\gamma)\; dr_1\, dr_2 \;-\; \sqrt{\alpha_1\alpha_2}. \tag{20}$$

This can be numerically computed for different values of (α_i, θ_i), $i = 1, 2$, and γ.

Case 2. Suppose that R_i is distributed as gamma$(\alpha, 1/\theta_i)$ and θ_i is also distributed as gamma$(\alpha_0, 1/\nu)$ for $i = 1, 2$. Then the joint probability distribution of (R_i, θ_i) is

$$f_i(r_i, \theta_i; \alpha, \alpha_0, \nu) = \frac{\nu^{\alpha_0}}{\Gamma(\alpha)\,\Gamma(\alpha_0)}\, r_i^{\alpha-1}\, \theta_i^{\alpha+\alpha_0-1}\, \exp[-\theta_i(r_i\nu)].$$

We can show that the marginal probability density function of r_i is Beta distribution of the second type $(Beta_2)$ with parameters (ν, α, α_0).

$$\begin{aligned} f_i(r_i; \alpha, \alpha_0, \nu) &= \int_0^\infty f_i(r_i, \theta_i; \alpha, \alpha_0, \nu)\, d\theta_i \\ &= \frac{\Gamma(\alpha+\alpha_0)}{\Gamma(\alpha)\,\Gamma(\alpha_0)}\, \nu^{\alpha_0}\, \frac{{r_i}^{\alpha-1}}{(r_i+\nu)^{\alpha+\alpha_0}} \\ &\sim Beta_2(\nu, \alpha, \alpha_0). \end{aligned} \tag{21}$$

The marginal mean and the variance of R_i are

$$\begin{aligned} E(R_i) &= \int_0^\infty \int_0^\infty r_i f_i(r_i, \theta_i; \alpha, \alpha_0, \nu)\, dr_i\, d\theta_i \\ &= \frac{\alpha\,\nu^{\alpha_0}}{\Gamma(\alpha_0)} \int_0^\infty \theta_i^{\alpha_0-2}\, \exp(-\theta_i\nu)\, d\theta_i \\ &= \frac{\alpha\nu}{\alpha_0-1}, \\ Var(R_i) &= \int_0^\infty \int_0^\infty r_i^2 f_i(r_i, \theta_i; \alpha, \alpha_0, \nu)\, dr_i\, d\theta_i\, - \left(\frac{\alpha\nu}{\alpha_0-1}\right)^2 \\ &= \frac{\alpha\,(\alpha+1)\,\nu^{\alpha_0}}{\Gamma(\alpha_0)} \int_0^\infty \theta_i^{\alpha_0-3}\, \exp(-\theta_i\nu)\, d\theta_i\, - \left(\frac{\alpha\nu}{\alpha_0-1}\right)^2 \\ &= \frac{\alpha(\alpha+1)}{(\alpha_0-1)(\alpha_0-2)}\, \nu^2\, - \left(\frac{\alpha\nu}{\alpha_0-1}\right)^2 \\ &= \frac{\alpha(\alpha+\alpha_0-2)}{(\alpha_0-1)^2(\alpha_0-2)}\, \nu^2. \end{aligned} \tag{22}$$

Note that $E(R_1) = E(R_2)$ and $Var(R_1) = Var(R_2)$ are functions of (α, α_0, ν). Assuming the joint distribution of $(R_1,\, R_2)$ is determined by the Gaussian copula with parameter γ, equation (17) gives the marginal density of $(R_1,\ R_2)$. The expected value of the product of R_1R_2 is given by

$$\begin{aligned} E[R_1R_2] &= \int_0^\infty \int_0^\infty r_1 r_2 f_m(r_1,\ r_2; \alpha, \alpha_0, \nu, \gamma)\ dr_1 dr_2 \\ &= \left(\frac{\nu^{\alpha_0}}{\Gamma(\alpha)\,\Gamma(\alpha_0)}\right)^2 \int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty r_1\ r_2\ c(F_1(r_1), F_2(r_2)) \times \\ &\qquad \prod_{i=1}^{2} \left[r_{ij}^{\alpha-1}\, \theta_{ij}^{\alpha+\alpha_0-1}\ \exp\left(-\theta_{ij}\left(r_{ij}+\nu\right)\right)\right] d\theta_1\, d\theta_2\, dr_1\, dr_2. \end{aligned} \tag{23}$$

Equation (23) has a double integral with respect to θ_1 and θ_2, and another additional double integral with respect to r_1 and r_2. The function adaptIntegrate in the R package cubature is useful to numerically evaluate this multidimensional integral. We have developed an R function that uses adaptIntegrate to calculate (23), and it is given in Appendix A. The relationship between the copula parameter γ and ρ in this case is given by

$$\rho \ = \ \frac{\alpha_0 - 2}{\alpha + \alpha_0 - 2} \left[\frac{(\alpha_0 - 1)^2}{\alpha \nu^2}\ E[R_1R_2]\ -\ \alpha \right]. \tag{24}$$

3.7 SIMULATION STUDY

In this section we check our parameter estimation methods for the Bayesian Gaussian copula model on simulated data. The data is simulated for two sets of values of $\boldsymbol{\Upsilon} = (\alpha, \alpha_0, \nu, \gamma)$ with three sample sizes $n = 100, 500, 3000$. The data simulation steps are as follows.
Fix a value for $\boldsymbol{\Upsilon} = (\alpha, \alpha_0, \nu, \gamma)$.

Step 1 Generate n pairs of bivariate normal random variables $(x_{1j},\ x_{2j})$ from standard bivariate normal distribution (BVN) with correlation parameter γ.

Step 2 Calculate $(u_{1j},\ u_{2j}) = (\Phi(x_{1j}), \Phi(x_{2j}))$ for $j = 1, \ldots, n$ where Φ is the cumulative distribution function of standard normal.

Step 3 Generate θ_{ij} from a gamma distribution with parameters $(\alpha_0, 1/\nu)$ for $i = 1, 2$ and $j = 1, \ldots, n$.

Step 4 Calculate $(r_{1j},\ r_{2j}) = \left(F_{1j}^{-1}(u_{1j}), F_{2j}^{-1}(u_{2j})\right)$ where F_{ij} is the cumulative distribution function of a gamma distribution with parameters $(\alpha, 1/\theta_{ij})$.

For our first simulation, we have fixed the parameter values as $\alpha = 0.5, \alpha_0 = 10, \nu = 25$, and $\gamma = 0.9$. With these parameter values we simulated samples of sizes n = 100, 500, and 3000. The results of parameter estimation are given in Table 1 and the scatter plots of simulated data are shown in Figure 5.

In Table 1, ρ is the correlation coefficient calculated from simulated data and $\widehat{\rho}$ is the correlation coefficient calculated after substituting the estimated values of $(\alpha, \alpha_0, \nu, \gamma)$ in equation (24). The parameter estimates are closer to the true parameter values for large sample size, and the standard errors get smaller as the sample size increases. The values of the correlation coefficients ρ and $\widehat{\rho}$ are reasonably close for all sample sizes.

Table 1. Parameter estimates (standard errors) for the simulated data†.

n	$\widehat{\alpha}$	$\widehat{\alpha}_0$	$\widehat{\nu}$	$\widehat{\gamma}$	ρ	$\widehat{\rho}$
100	0.541	8.901	22.043	0.835	0.806	0.713
	(0.054)	(0.280)	(3.829)	(0.005)		
500	0.506	9.850	25.922	0.910	0.832	0.827
	(0.024)	(0.102)	(1.678)	(0.010)		
3000	0.540	10.119	25.999	0.890	0.771	0.805
	(0.011)	(0.016)	(0.421)	(<0.001)		

†True parameter values are $\alpha = 0.5, \alpha_0 = 10, \nu = 25$, and $\gamma = 0.9$.

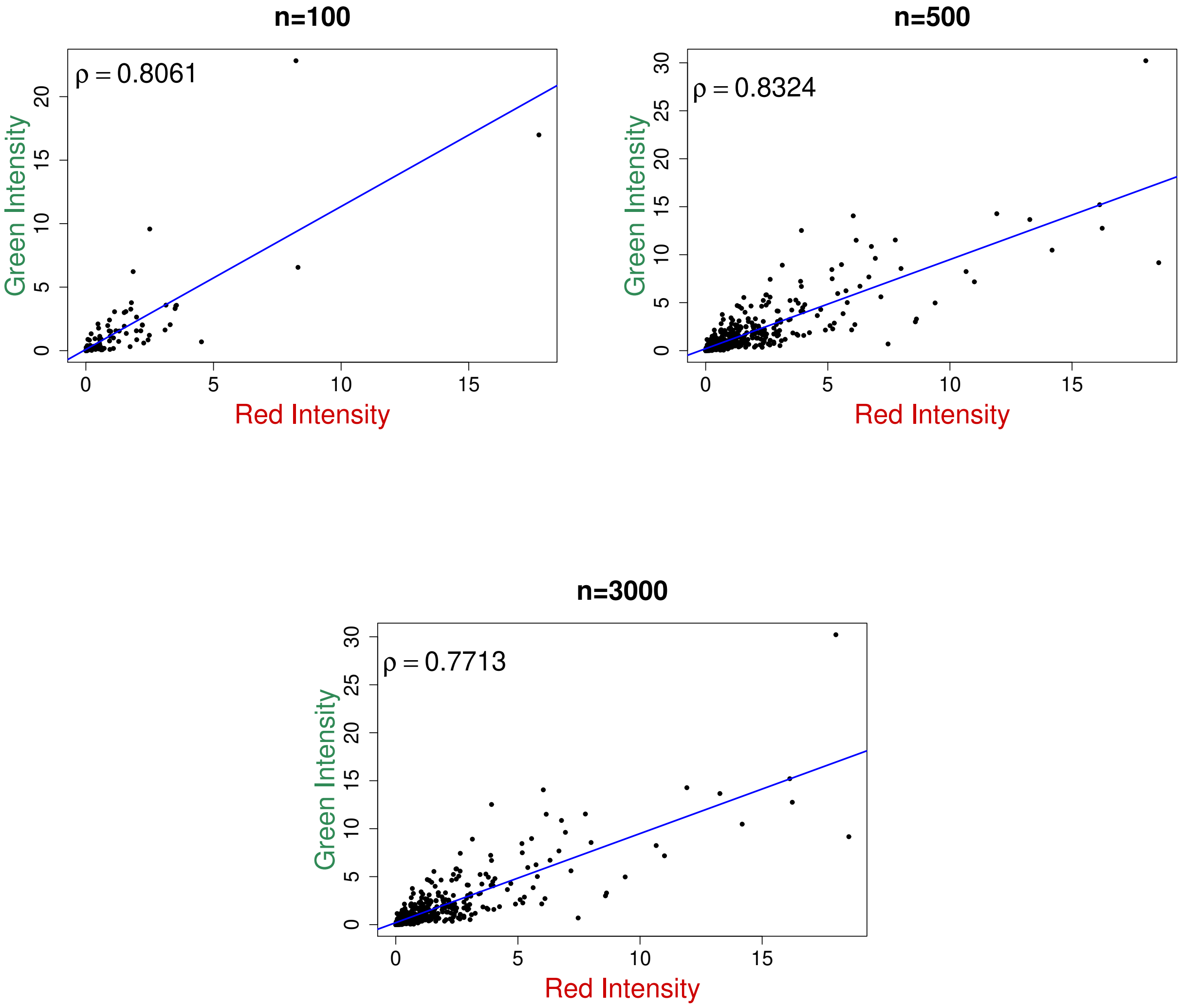


Figure 5. Scatter plots of simulated data with $\alpha = 0.5, \alpha_0 = 10, \nu = 25$, and $\gamma = 0.9$.

For our second simulation, we fixed the parameter values as $\alpha = 2, \alpha_0 = 27, \nu = 900$, and $\gamma = 0.8$, and as before we took three sample sizes 100, 500 and 3000. Figure 6 shows the scatter plots of simulated data, and Table 2 consists of parameter estimation results for this second simulation.

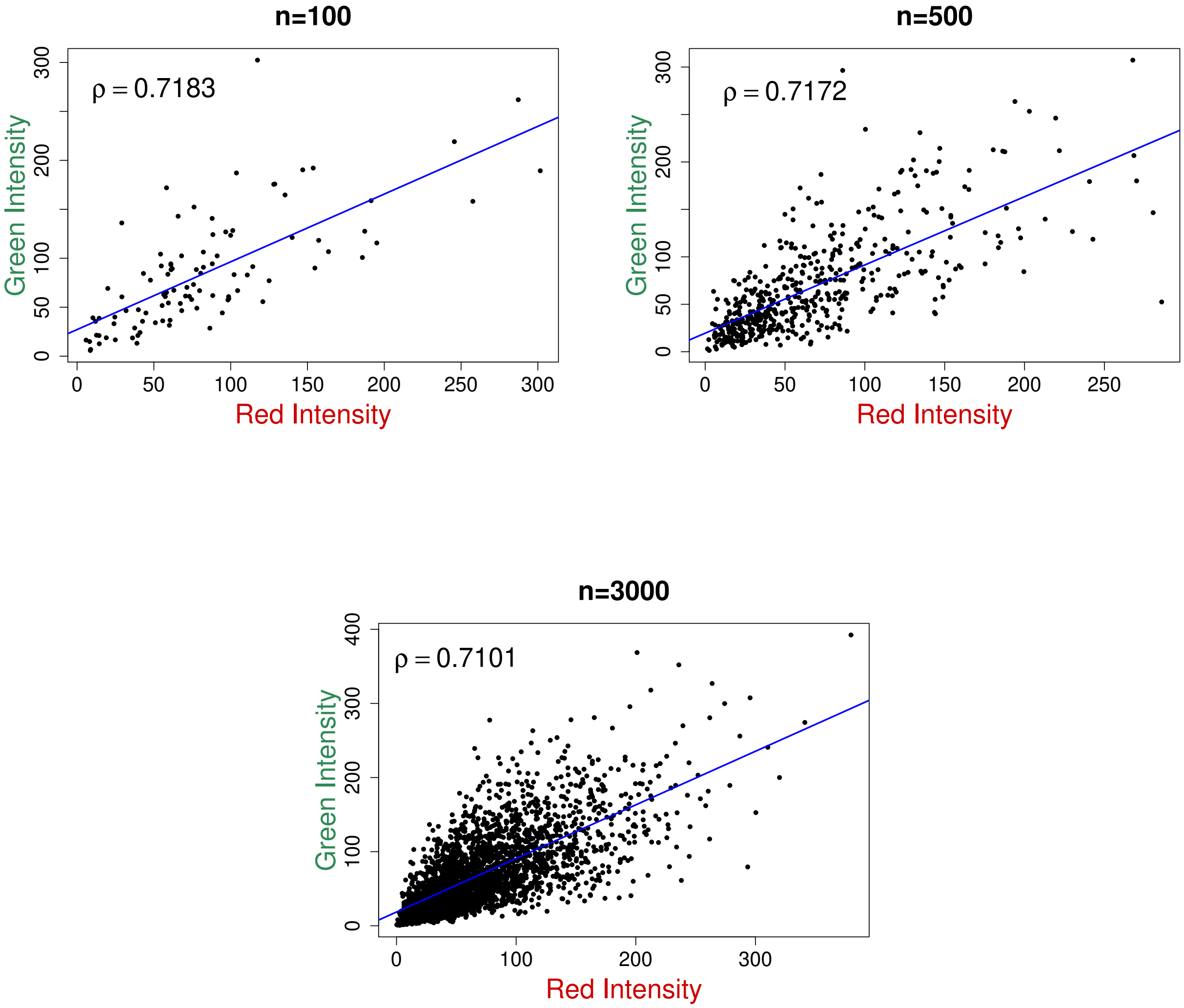


Figure 6. Scatter plots of simulated data with $\alpha = 2, \alpha_0 = 27, \nu = 900$, and $\gamma = 0.8$.

Table 2. Parameter estimates (standard errors) for the simulated data†.

n	$\widehat{\alpha}$	$\widehat{\alpha}_0$	$\widehat{\nu}$	$\widehat{\gamma}$	ρ	$\widehat{\rho}$
100	2.629	31.999	899.216	0.737	0.718	0.485
	(0.005)	(0.019)	(2.772)	(0.003)		
500	2.018	30.424	898.990	0.706	0.717	0.407
	(0.003)	(0.026)	(1.982)	(0.005)		
3000	2.187	26.256	898.991	0.863	0.710	0.775
	(0.003)	(0.002)	(0.753)	(0.001)		

†True parameter values are $\alpha = 2, \alpha_0 = 27, \nu = 900$, and $\gamma = 0.8$.

For $n = 100, 300$, the estimate $\widehat{\alpha}_0$ is an over estimate of α_0, and $\widehat{\rho}$ is terribly an under estimate of ρ, otherwise the results are consistent with the first simulation. All the parameter estimates are closer to their true values for lager sample size $n = 3000$. This is a good news because in practice n, which represents the number of genes, is in thousands.

3.8 ANALYSIS OF *E. COLI* DATA

In this section we apply the Bayesian Gaussian copula model that we had developed in Section 3.3 to some real data obtained from microarray experiments on *E. coli*. These data consists of observations from five microarrays. There are two IPTG treated samples labeled IPTG-A and IPTG-B, and two heatshock samples labeled as Heat Shock-A and Heat Shock-B and the fifth is a control (untreated). We have described these data earlier in Section 3.2. There are 4253, 4083, 4141, 4208, and 4071 genes in control, IPTG-A, IPTG-B, Heat Shock-A and Heat Shock-B samples, respectively. The first 15 observations taken from control sample *E. coli* are shown in Table 3. The “Bnumber" is a label associated with the gene.

The scatter plots for the red and green intensities for the five samples are shown in Figure 7, along with the sample correlation coefficients. Clearly, there is a high positive correlation between red and green intensities in all of the five samples. Figures 8 and 9 show the histograms of red and green intensities along with the nonparametric kernel density plots for the five microarray experiments. The positively skewed shape of the density curves suggest the assumption of gamma marginals is reasonable. Thus following Newton et al. (2001), as a parsimonious model, we assume the marginal distributions of the red and green intensities as gamma with common shape parameter but different scale parameters.

Table 3. Sample data for the *E. coli* example.

		Intensity	
Obs	Bnumber	Red (R_1)	Green (R_2)
1	b0001	1.4780	1.4107
2	b0002	13.0661	9.0702
3	b0003	22.4852	15.4512
4	b0004	12.8999	6.7668
5	b0005	4.5915	5.2459
6	b0006	29.8578	30.9245
7	b0007	14.1593	11.0870
8	b0008	157.8423	137.1544
9	b0009	9.5066	8.2216
10	b0010	19.6253	18.3938
11	b0011	7.8186	7.6815
12	b0012	10.7135	7.9901
13	b0013	1.8191	1.6862
14	b0014	36.8106	26.7476
15	b0015	28.1874	22.6589

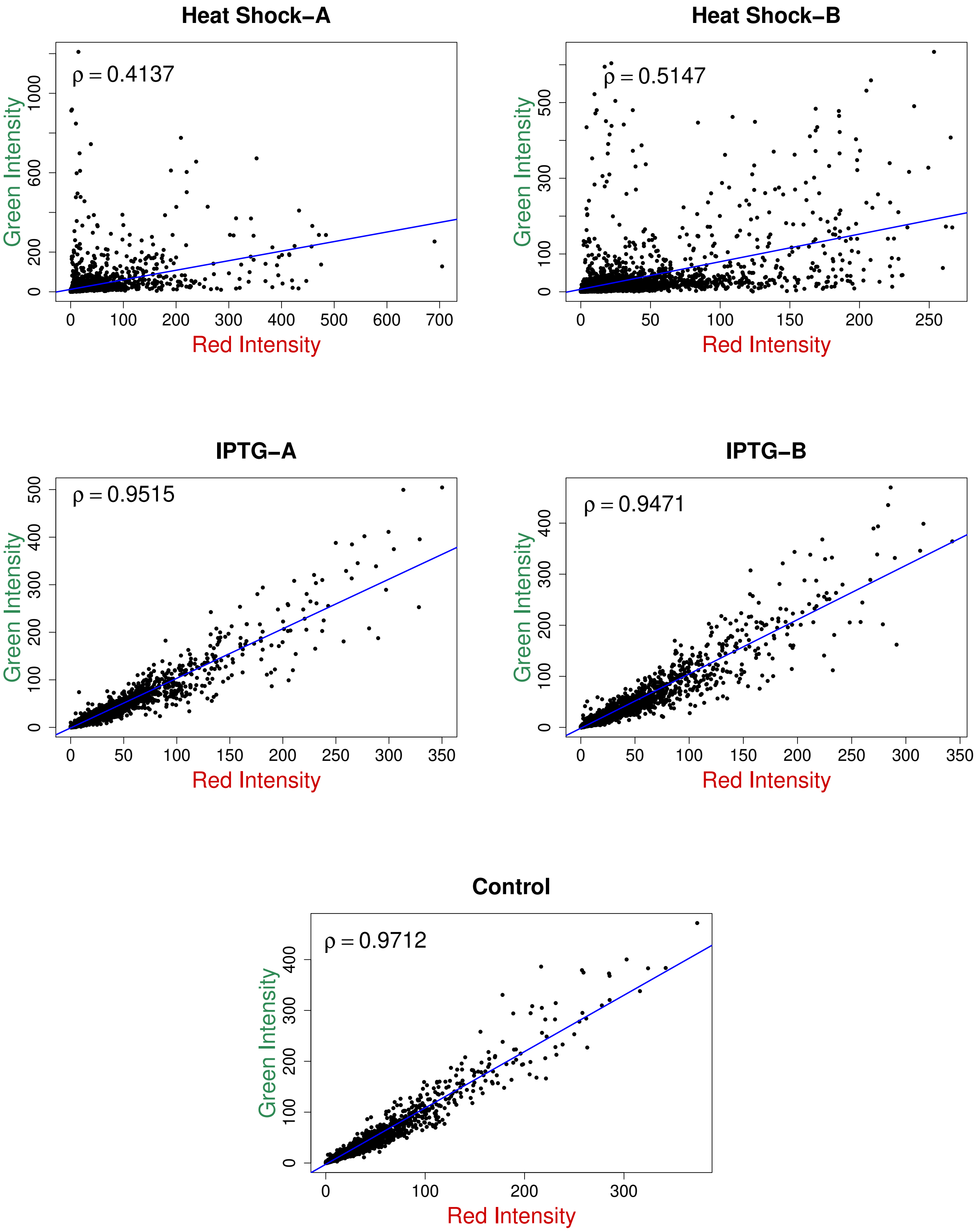


Figure 7. Scatter plots of red and green intensities.

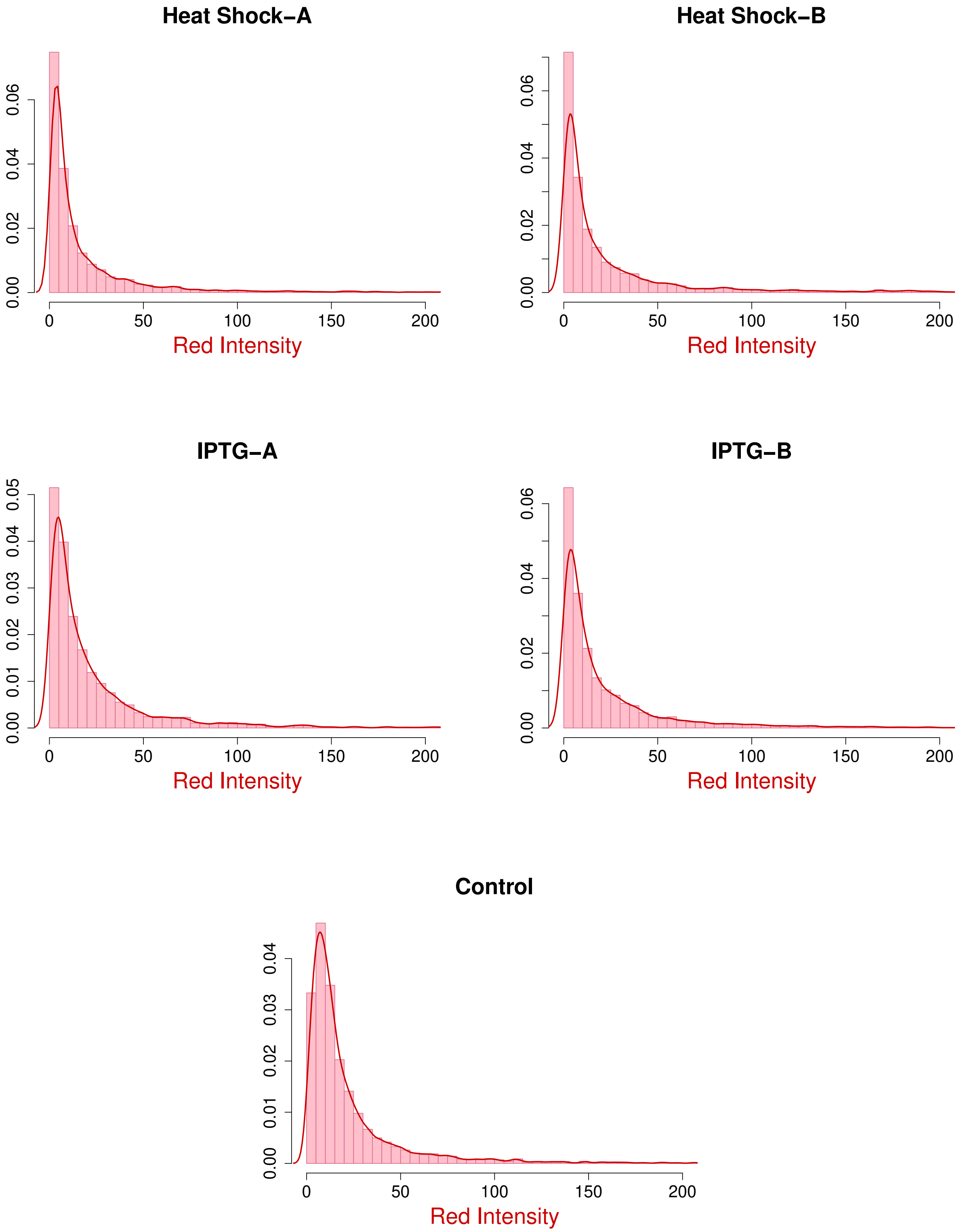


Figure 8. Histogram of red intensities with density plots.

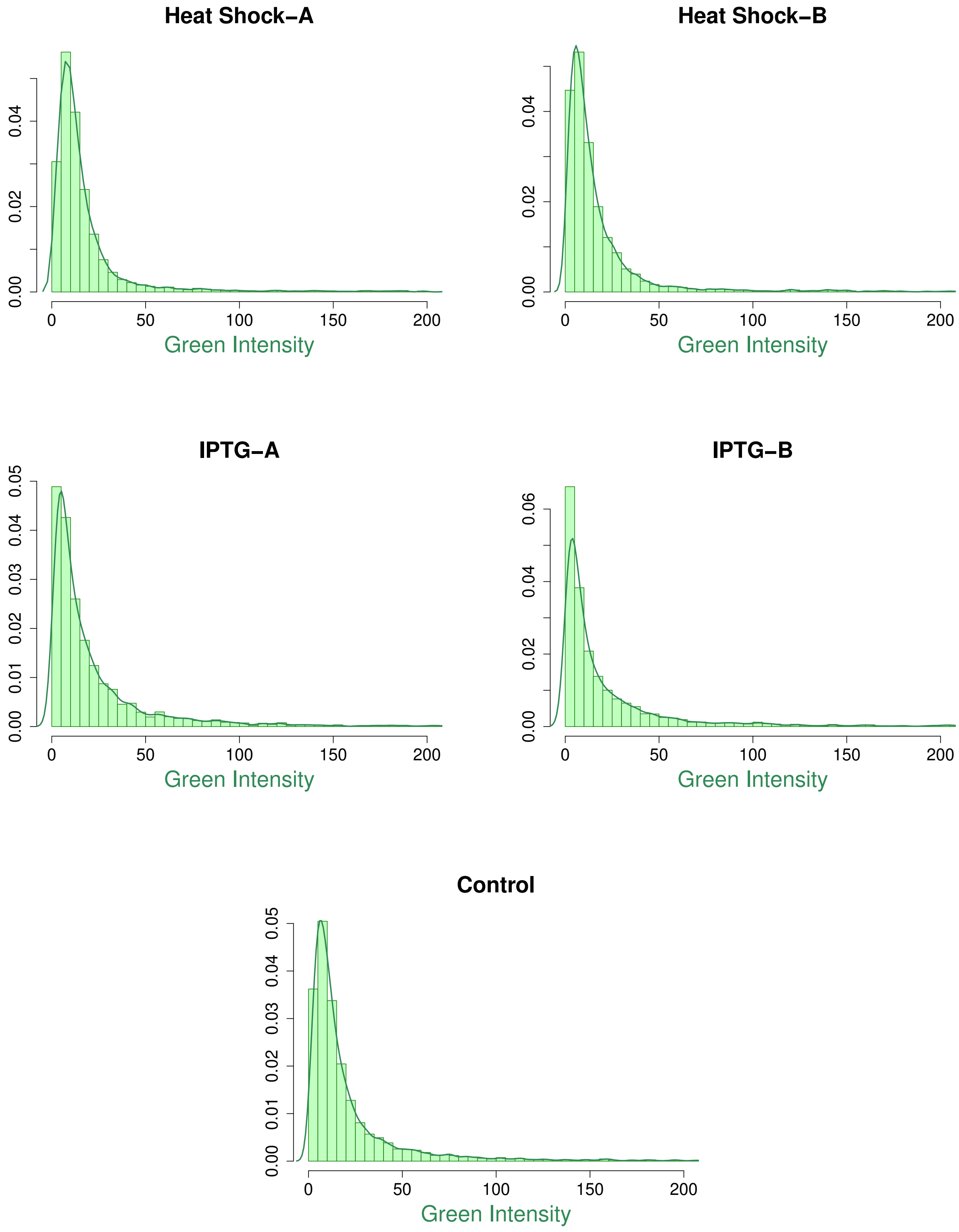


Figure 9. Histogram of green intensities with density plots.

Table 4 contains the parameter estimates and their standard errors for the Bayesian Gaussian copula models for the five microarray samples. The standard errors are small because the sample sizes are large; more than 4000 in all cases. This suggests the parameter estimates are fairly accurate.

Table 4. Parameter estimates (standard errors) for the *E. coli* data.

Microarray	α	α_0	ν	γ
Control	0.796	55.245	1529.876	0.9896
	(0.001)	(0.343)	(1.056)	(0.003)
IPTG-A	0.743	40.048	1149.604	0.9838
	(0.001)	(0.127)	(1.181)	(0.021)
IPTG-B	0.643	27.095	899.997	0.9839
	(0.003)	(0.059)	(0.678)	(0.019)
Heat Shock-A	1.777	4.644	24.999	0.8116
	(0.039)	(0.094)	(0.056)	(0.013)
Heat Shock-B	1.449	4.613	29.999	0.6507
	(0.024)	(0.083)	(0.102)	(0.015)

The empirical density plots along with the fitted density plots are shown in Figures 10 and 11 for the red and green intensities, respectively. The solid curves in Figures 10 and 11 are the fitted gamma marginals and the shaded curves are the empirical plots. Note the fitted marginals are gamma densities with the estimated parameter values in Table 4. These figures show the fitted marginals are very good for the IPTG and control samples but there is some improvement for the heat shock samples, especially the red intensities.

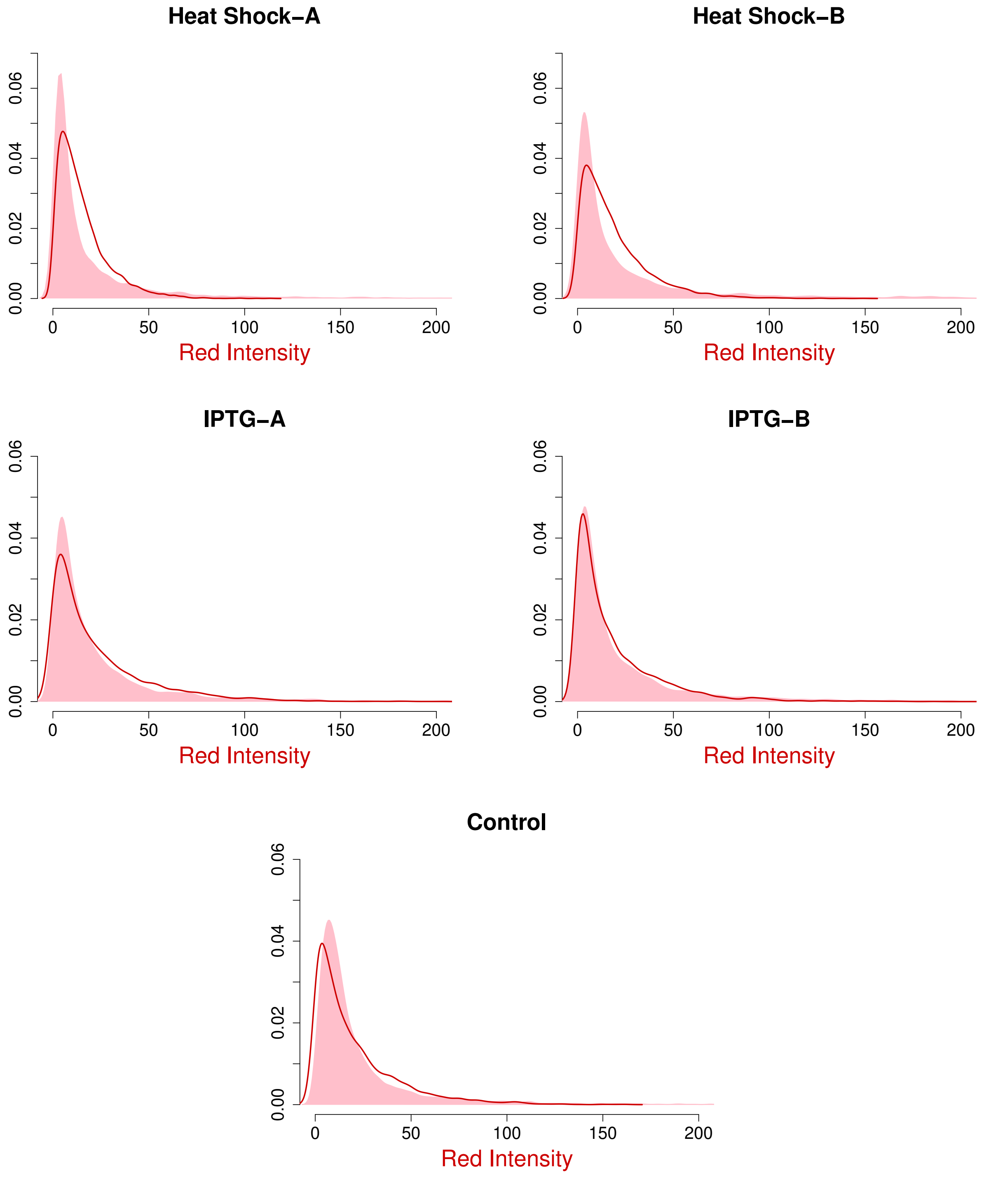


Figure 10. Density plots of red intensities.

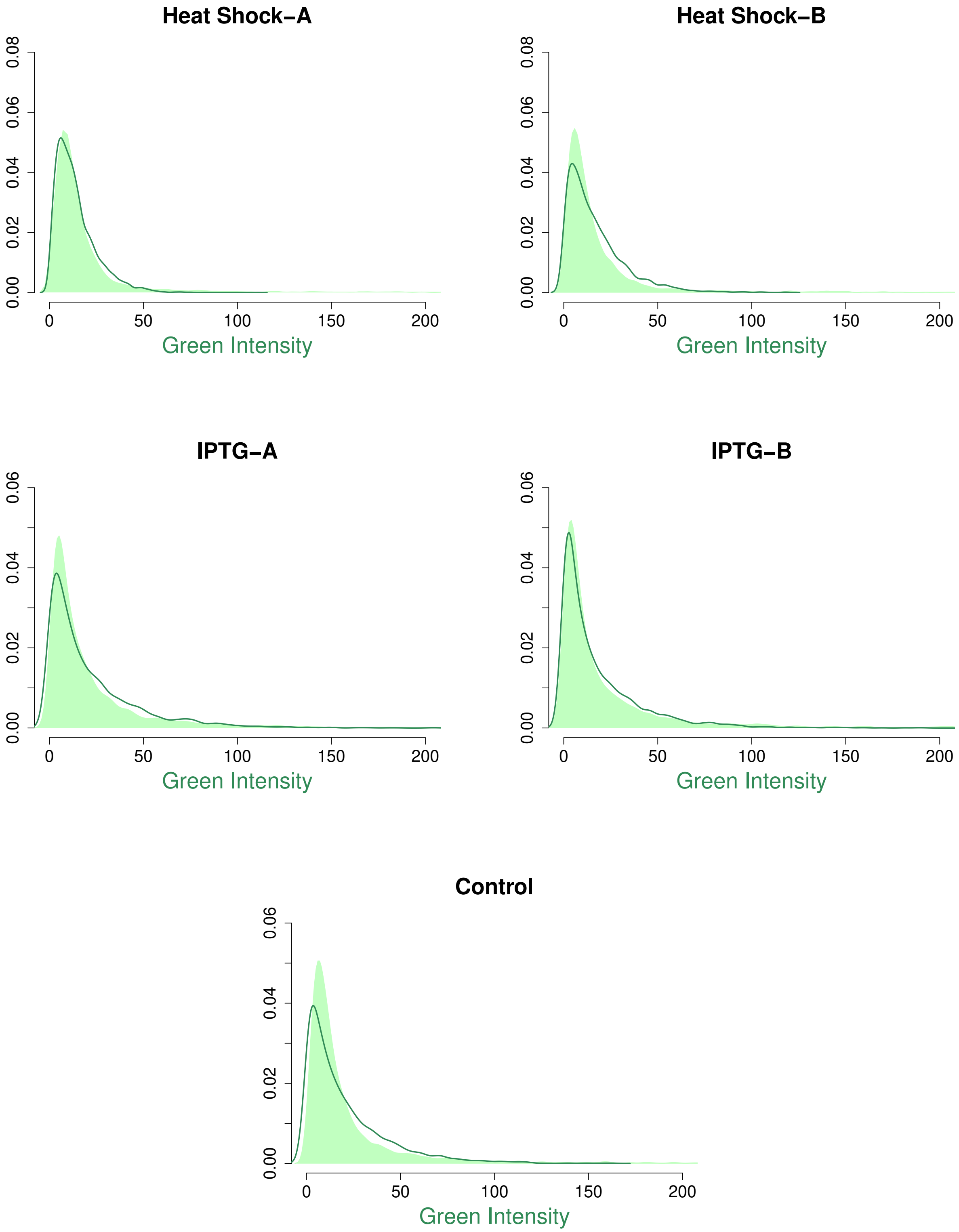


Figure 11. Density plots of green intensities.

Figure 12 shows the fitted bivariate density plots obtained using the parameter estimates in Table 4. In these plots the 45^0 line indicates equal red and green intensities and the points that fall on this line correspond to genes that are not differentially expressed. Using this criteria we can see we can see most of the genes in the control group are not differentially expressed. For the IPTG samples a few points lie away from the 45^0 line indicating the presence of differentially expressed genes in these samples. Finally, for the two Heat Shock samples a large number of points are away from the 45^0 line indicating there are a large number of differentially expressed genes in these samples.

Table 5. True and estimated correlation coefficients.

Microarray	ρ	$\widehat{\rho}$
Control	0.9712	0.9748
IPTG-A	0.9515	0.9163
IPTG-B	0.9471	0.9799
Heat Shock-A	0.4137	0.4723
Heat Shock-B	0.5147	0.3971

Table 5 displays the observed correlation (ρ) and correlation coefficient ($\widehat{\rho}$) calculated from the estimated copula parameter as in Table 4 using equation (24). Except for Heat Shock-B, for all the other four samples the values of ρ and $\widehat{\rho}$ are very close indicating that our copula model was fairly successful in quantifying the dependence between the two intensities.

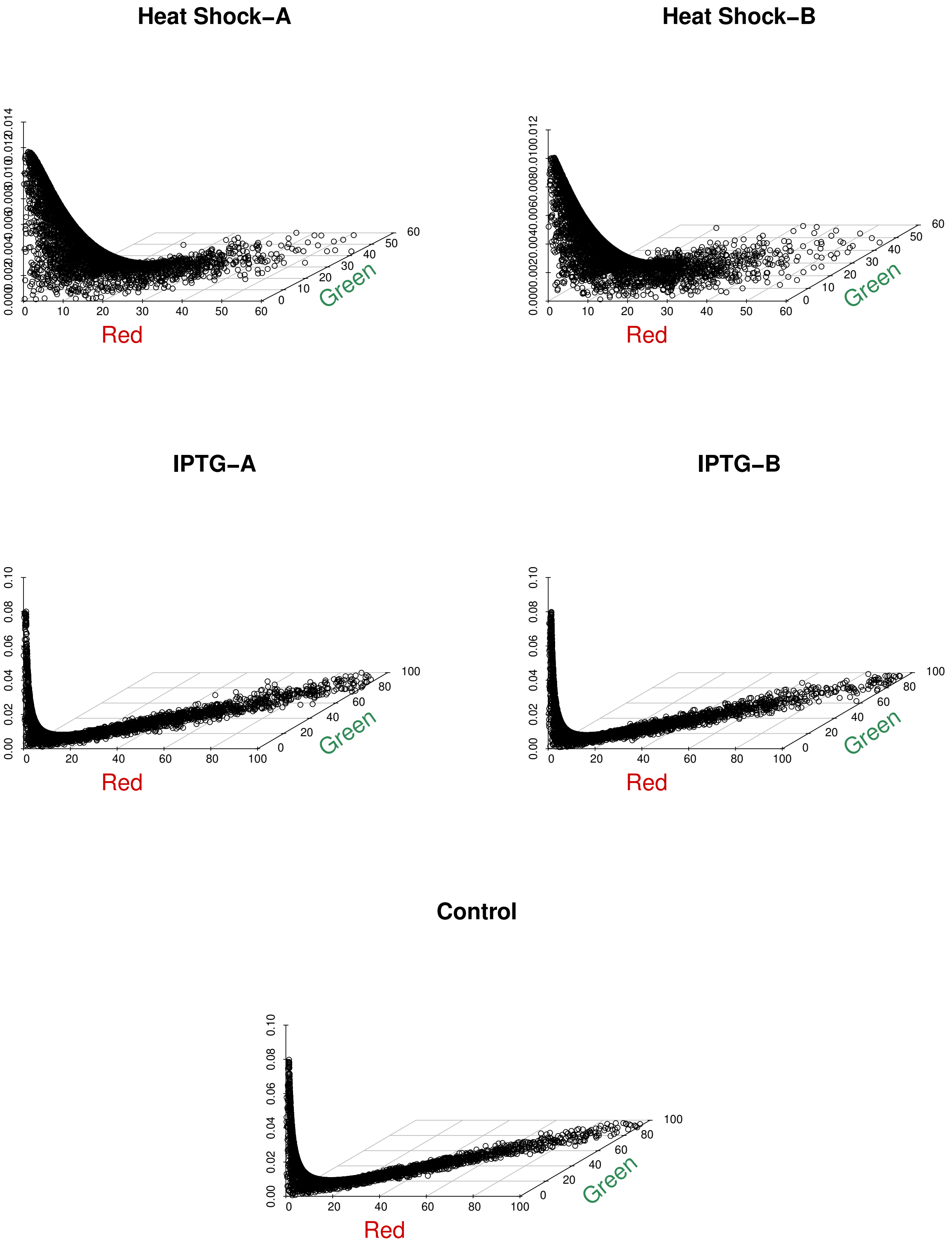


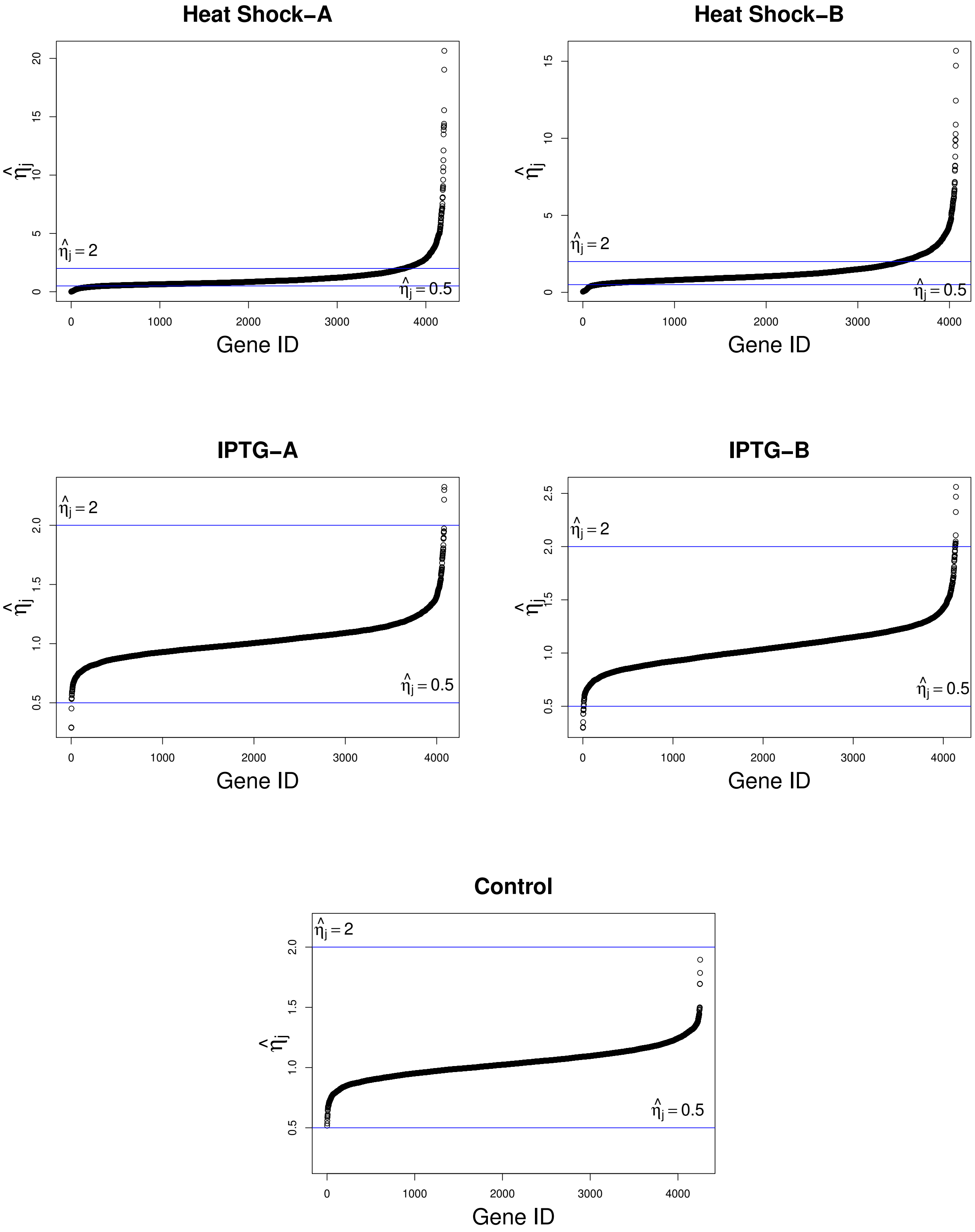
Figure 12. Estimated bivariate density plots of red and green intensities.

Table 6. Top 20 down-regulated genes.

#	Control		IPTG-A		IPTG-B		Heat Shock-A		Heat Shock-B	
	Gene id	$\widetilde{\eta_j}$	Gene id	$\widetilde{\eta_j}$	Gene id	$\widetilde{\eta_j}$	Gene id	$\widetilde{\eta_j}$	Gene id	$\widetilde{\eta_j}$
1	b0233	0.52	b4098	0.29	b4119	0.30	b3686	0.00	b3686	0.04
2	b1325	0.53	b4119	0.29	b4120	0.30	b3687	0.00	b3687	0.05
3	b0558	0.56	b4120	0.45	b4149	0.35	b0014	0.02	b4142	0.05
4	b2843	0.58	b0296	0.53	b0341	0.43	b1306	0.03	b0015	0.05
5	b2129	0.60	b4291	0.54	b4291	0.43	b1967	0.03	b0014	0.05
6	b1319	0.60	b1571	0.54	b1785	0.46	b1304	0.03	b3400	0.06
7	b3818	0.61	b0720	0.56	b1020	0.47	b1380	0.03	b1380	0.06
8	b1075	0.64	b1020	0.58	b0648	0.47	b2614	0.04	b2592	0.07
9	b1924	0.65	b3908	0.59	b0558	0.50	b0399	0.04	b1306	0.08
10	b2742	0.66	b1500	0.60	b2260	0.52	b1305	0.04	b0966	0.08
11	b1447	0.67	b0326	0.60	b0705	0.55	b1307	0.04	b4143	0.08
12	b3341	0.68	b3962	0.60	b0720	0.55	b4143	0.04	b1304	0.08
13	b3751	0.68	b4149	0.61	b3489	0.56	b4140	0.05	b1307	0.08
14	b2756	0.68	b0702	0.62	b0302	0.56	b3400	0.05	b1305	0.08
15	b1166	0.68	b1018	0.63	b3342	0.57	b4142	0.05	b2614	0.09
16	b2051	0.68	b4247	0.63	b1166	0.58	b3401	0.05	b0473	0.09
17	b2541	0.68	b1685	0.64	b0805	0.60	b1321	0.05	b0016	0.09
18	b3340	0.69	b2260	0.65	b1681	0.60	b0473	0.05	b3932	0.09
19	b3966	0.70	b0705	0.65	b2843	0.61	b1829	0.06	b1060	0.10
20	b2628	0.70	b0726	0.66	b3508	0.61	b4171	0.07	b1829	0.10

Table 7. Top 20 up-regulated genes.

	Control		IPTG-A		IPTG-B		Heat Shock-A		Heat Shock-B	
#	Gene id	$\widetilde{\eta_j}$	Gene id	$\widetilde{\eta_j}$	Gene id	$\widetilde{\eta_j}$	Gene id	$\widetilde{\eta_j}$	Gene id	$\widetilde{\eta_j}$
1	b4325	1.89	b2206	2.32	b1256	2.56	b3556	20.65	b3556	15.68
2	b0657	1.79	b1256	2.30	b2206	2.47	b2094	19.03	b1078	14.71
3	b2740	1.70	b0043	2.22	b0759	2.32	b1076	15.56	b0907	12.44
4	b0542	1.69	b1673	1.97	b4307	2.11	b1077	14.40	b1857	10.89
5	b0679	1.50	b2205	1.95	b1674	2.05	b1075	14.24	b1074	10.28
6	b4314	1.50	b2204	1.95	b2997	2.03	b1857	14.12	b1076	9.91
7	b2418	1.49	b2997	1.94	b2203	2.03	b0754	14.11	b0296	9.87
8	b3616	1.48	b0115	1.90	b2151	2.02	b1074	13.85	b2094	9.51
9	b4243	1.48	b0759	1.89	b0733	2.01	b2926	13.49	b1588	8.81
10	b0185	1.47	b2727	1.89	b0857	2.00	b1073	12.11	b1245	8.22
11	b1084	1.45	b2241	1.83	b2204	1.99	b2935	11.27	b4025	8.18
12	b3834	1.44	b0283	1.81	b2242	1.97	b3544	10.66	b4328	7.97
13	b2860	1.44	b2202	1.80	b2205	1.95	b1078	10.32	b1885	7.89
14	b0729	1.44	b2996	1.80	b2996	1.95	b0907	9.60	b2241	7.17
15	b1083	1.43	b0347	1.78	b2957	1.92	b2416	9.05	b1244	7.12
16	b1064	1.42	b0733	1.78	b2727	1.91	b2092	8.93	b1417	7.03
17	b2283	1.42	b2957	1.78	b2149	1.90	b2286	8.84	b0131	6.70
18	b3147	1.42	b2203	1.76	b2241	1.90	b3357	8.73	b1938	6.63
19	b1674	1.41	b2151	1.75	b0598	1.88	b0893	8.10	b1676	6.57
20	b0698	1.41	b2242	1.74	b0894	1.87	b1244	8.06	b1072	6.53



Heat Shock−A

Heat Shock−B

IPTG−A

IPTG−B

Control

Figure 13. Plots of $\widetilde{\eta_j}$.

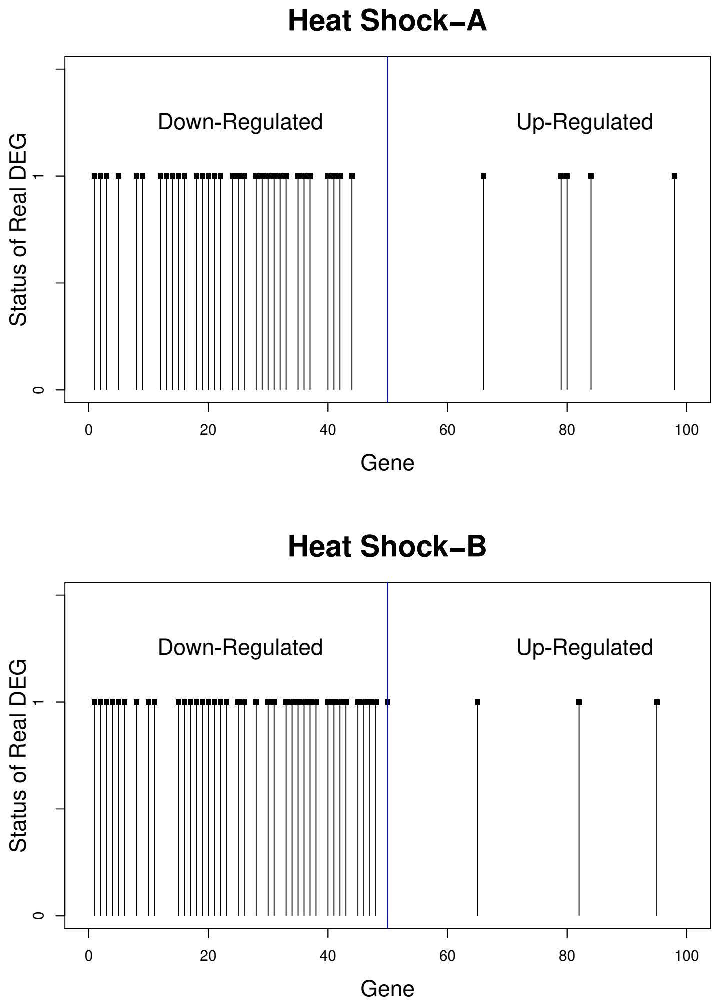


Figure 14. Plots of differentially expressed gene comparison

We calculated $\widetilde{\eta_j}$ using equation (19), the Bayes estimate $\widetilde{\eta_j}$ of η_j which is a measure of differential expression of the jth gene. Table 6 displays the top twenty down-regulated ($\widetilde{\eta_j}$ is small) genes and Table 7 lists the top twenty up-regulated ($\widetilde{\eta_j}$ is large) genes for all the five samples.

Plots of ordered $\widetilde{\eta_j}$ values for the five samples are displayed in Figure 13. These plots are S-shaped, and the left tails contain the down-regulated genes, whereas the right tails contain the up-regulated genes. In their paper, Richmond et al. (1999) have listed the genes that are significantly affected by Heat Shock and IPTG treatments. According to their findings, the control sample has none of the differentially expressed genes, IPTG samples have few, and Heat shock samples have a large number of differentially expressed genes. Therefore, by considering the number of differentially expressed genes and the plots of $\widetilde{\eta_j}$ of five microarrays, $\widetilde{\eta_j} = 2$ is a good candidate cut off value to filter up-regulated genes while $\widetilde{\eta_j} = 0.5$ is for down-regulated genes. The horizontal lines in Figure 13 indicate the possible cut off values to separate the normal genes from the two extremes.

The total number of differentially expressed genes for each microarray is listed in Table 8 along with the total number of differentially expressed genes filtered with the bivariate gamma model was proposed by Mav and Chaganty (2004).

Table 8. Total number of differentially expressed genes.

Microarray	# of Genes for which			
	$\widetilde{\eta_j} > 2$		$\widetilde{\eta_j} < 0.5$	
	Bivariate	Gaussian	Bivariate	Gaussian
	Gamma	Copula	Gamma	Copula
Control	0	0	0	0
IPTG-A	10	3	3	3
IPTG-B	21	9	7	8
Heat Shock-A	553	451	1007	439
Heat Shock-B	856	600	590	169

As expected, none of the genes are identified as differentially expressed in the control sample, and very few in IPTG-A and IPTG-B. Many genes have been filtered as up or down-regulated from both models for the Heat Shock-A and Heat Shock-B. The number of genes filtered from the Gaussian copula model is somewhat smaller than that from the Bivariate gamma model.

The best model cannot be determined by looking at this total number of genes. Therefore, the log-likelihoods for the two models under each microarray are compared and shown in Table 9. The log-likelihood values under our model are larger than that of the bivariate gamma model, which was proposed by Mav and Chaganty (2004) for each microarray. Hence we conclude the Bayesian Gaussian copula model has an improvement over the model given in Mav and Chaganty (2004). Further, the filtered differentially expressed genes of Heat shock samples by our method are well-matched with the genes are listed in Richmond et al. (1999). Recall that in Richmond et al. (1999), the control sample had no differentially expressed genes and IPTG samples had few, which are consistent with our findings.

Table 9. Log-likelihoods for the competitive models.

Microarray	Bivariate Gamma	Gaussian Copula
Control	-28824	-28350
IPTG-A	-28320	-27853
IPTG-B	-28257	-27885
Heat Shock-A	-31936	-30419
Heat Shock-B	-31658	-30282

3.9 CONCLUSIONS

Several methods have been proposed to identify differentially expressed genes in the literature. This chapter develops a Bayesian Gaussian copula model to detect the differentially expressed genes in a cDNA microarray. The accuracy of model parameter estimations is shown with two simulation studies with three different sample sizes. We applied the developed model to the five microarray samples in *E. coli* separately. The experimentally found differentially expressed genes in *E. coli* data have listed in Richmond et al. (1999). The Bayes estimate of the differential expression is used to filter up-regulated and down-regulated genes.

Many of the genes identified as down-regulated by our model are matched with the genes stated in Mav and Chaganty (2004) paper.

However, Mav and Chaganty (2004) have proposed a Bivariate Gamma model for the same purpose on the same *E. coli* data. The larger log-likelihood values under our model with compare to the model of Mav and Chaganty (2004), suggest that our model has an improvement over the Bivariate Gamma model. Our model's main advantage is that it can be applied to any marginal distributions of intensities, while the Biivariate Gamma model is always based on Gamma marginals. In the next chapter we will study the Bayesian Gaussian copula model incorporating a latent Bernoulli variable.

CHAPTER 4

BAYESIAN COPULA MODEL WITH A LATENT VARIABLE

4.1 INTRODUCTION

In Chapter 3 we have developed a Bayesian model that uses a Gaussian copula for the joint distribution for the red and green intensities that arise in a cDNA microarray. Further, we assumed both the marginal and prior distributions are gamma. We have used the posterior estimates of the mean intensities ratios to classify the down and up-regulated genes. In this chapter, we add another layer to the model by introducing a binary latent variable that indicates presence and absence of differential expression. For this extended model, we calculate the posterior probabilities of differential expression and use them to rank order the genes.

4.2 BAYESIAN COPULA MODEL WITH A LATENT VARIABLE

In this section, we start with the model described in Section 3.3 in the previous chapter. Recall that, the marginal distributions of red (R_{1j}) and green (R_{2j}) intensities were assumed to be distributed as gamma with common shape parameter α and different scale parameters $1/\theta_{1j}$ and $1/\theta_{2j}$. In additional to this assumption, we assumed that the prior distributions for θ_{ij}'s are independent gamma distributions with parameters α_0 and $1/\nu$ for $j = 1, \dots, n$ and $i = 1, 2$. Our goal is to extend this model by assuming that there is an unknown proportion p of genes that exhibit differential expression in a microarray. To accomplish this goal we define a latent unobserved Bernoulli variable W_j which indicates whether the j th gene is differentially expressed,

$$W_j = \begin{cases} 0, & \text{if } \theta_{1j} = \theta_{2j} = \theta_j \\ 1, & \text{if } \theta_{1j} \neq \theta_{2j}. \end{cases}$$

If the j th gene is differentially expressed ($W_j = 1$), then the marginal density of (R_{1j}, R_{2j})

is given by

$$\begin{aligned} f_{de}(r_{1j},\, r_{2j}; \boldsymbol{\Upsilon}) &= \int_0^\infty \int_0^\infty f(r_{1j},\, r_{2j}; \theta_{1j}, \theta_{2j}; \boldsymbol{\Upsilon})\; d\theta_{1j}\; d\theta_{2j} \\ &= \left(\frac{\nu^{\alpha_0}}{\Gamma(\alpha)\,\Gamma(\alpha_0)}\right)^2 \int_0^\infty \int_0^\infty c(F_{1j}(r_{1j}), F_{2j}(r_{2j}))\; \times \\ &\qquad \prod_{i=1}^{2} \left[r_{ij}^{\alpha-1}\, \theta_{ij}^{\alpha+\alpha_0-1}\; \exp\left(-\theta_{ij}\left(r_{ij}+\nu\right)\right)\right] d\theta_{1j}\, d\theta_{2j}. \end{aligned} \tag{25}$$

Here $F_{ij}(r_{ij})$ the cumulative distribution function of a gamma distribution with parameters $(\alpha, 1/\theta_{ij})$ for $i = 1, 2$ and $j = 1, 2, \ldots, n$. For a gene j that is not differentially expressed $(W_j = 0)$, the marginal density of $(R_{1j},\, R_{2j})$ is given by

$$\begin{aligned} f_{nde}(r_{1j},\, r_{2j}; \Upsilon) &= \frac{\nu^{\alpha_0}\, (r_{1j}r_{2j})^{\alpha-1}}{\Gamma^2(\alpha)\,\Gamma(\alpha_0)} \int_0^\infty c(F_j(r_{1j}), F_j(r_{2j}))\; \times \\ &\qquad \theta_j^{2\alpha+\alpha_0-1}\; \exp\left[-\theta_j(r_{1j}+r_{2j}+\nu)\right]\, d\theta_j, \end{aligned} \tag{26}$$

where $F_j(.)$ does not depend on i and it is the cumulative distribution function of gamma with parameters $(\alpha, 1/\theta_j)$. Here c is the bivariate Gaussian copula density function given by (14) and $\boldsymbol{\Upsilon} = (\alpha, \alpha_0, \nu, \gamma)$ is the vector of model parameters.

4.3 PARAMETER ESTIMATION PROCEDURE

In this section we discuss maximum likelihood estimation of the parameters $\boldsymbol{\Upsilon}$ and p in the model that we described in Section 4.2. Using (25) and (26), we can write the complete data log-likelihood for a sample of n genes as

$$l(\boldsymbol{\Upsilon}, p) = \sum_{j=1}^{n} \log \left\{ f_{de}(r_{1j},\, r_{2j}; \boldsymbol{\Upsilon})^{w_j} f_{nde}(r_{1j},\, r_{2j}; \boldsymbol{\Upsilon})^{1-w_j} p^{w_j} (1-p)^{1-w_j} \right\}. \tag{27}$$

Recall that w_j's are unobserved latent Bernoulli variables, and therefore, we use expectation maximization (EM) algorithm to maximize the log-likelihood (27) to obtain the maximum likelihood estimates of the parameters. The EM algorithm is an iterative procedure that iterates between an expectation (E) step (to fill the unobserved variables) followed by a maximization (M) step. In a seminal paper Dempster et al. (1977) introduced this method to find the maximum likelihood estimates in the presence of latent variables or missing

data. See McLachlan and Krishnan (1997) for more extensive detailed description of the EM algorithm and several applications of the method.
In summary the EM algorithm to estimate the parameters $\boldsymbol{\Upsilon} = (\alpha, \alpha_0, \nu, \gamma)$ and p goes as follows.

Step 1 Select initial values $\alpha_{\{0\}}, \alpha_{0\{0\}}, \nu_{\{0\}}, \gamma_{\{0\}}, p_{\{0\}}$ for the parameters $\alpha, \alpha_0, \nu, \gamma$ and p respectively.

Step 2 E-step: Calculate $\widehat{w}_j$ using following equation:

$$\widehat{w}_j = E(w_j \,|\, r_{1j},\, r_{2j}) = \frac{p_{\{0\}}\, f_{de}(r_{1j},\, r_{2j}; \boldsymbol{\Upsilon}_{\{0\}})}{p_{\{0\}}\, f_{de}(r_{1j},\, r_{2j}; \boldsymbol{\Upsilon}_{\{0\}}) + (1 - p_{\{0\}})\, f_{nde}(r_{1j},\, r_{2j}; \boldsymbol{\Upsilon}_{\{0\}})}. \tag{28}$$

Step 3 M-step: Maximize (27) and obtain an updated estimates $\alpha_{\{1\}}, \alpha_{0\{1\}}, \nu_{\{1\}}, \gamma_{\{1\}}, p_{\{1\}}$ of the unknown parameters.

Step 4 Repeat the E-step and the M-step until the parameter estimates converge.

Note that in the M-step involves maximizing the likelihood function and this usually done solving the likelihood equation $\partial l(\boldsymbol{\Omega})/\partial \boldsymbol{\Omega} = 0$, where $\boldsymbol{\Omega} = (\boldsymbol{\Upsilon}, p)$. However, analytical expressions for the first order partial derivatives are very complicated and it is no easy task to solve the likelihood equation. An alternative is the method that we have described in Section 3.4.1. We have used the quasi-Newton method "BFGS" in the function optim in the R package stats to obtain the maximum with respect to the parameter $\boldsymbol{\Omega} = (\boldsymbol{\Upsilon}, p) = (\alpha, \alpha_0, \nu, \gamma, p)$ in the M-step.

4.3.1 STANDARD ERRORS FOR ML ESTIMATES

The R routines that we discussed above in Section 4.3 will produce a numerical value for the Hessian Matrix given by

$$\frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \boldsymbol{\Omega} \partial \boldsymbol{\Omega}'} = \begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \alpha^2} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \alpha \partial \alpha_0} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \alpha \partial \nu} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \alpha \partial \gamma} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \alpha \partial p} \\ \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \alpha_0 \partial \alpha} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial {\alpha_0}^2} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \alpha_0 \partial \nu} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \alpha_0 \partial \gamma} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \alpha_0 \partial p} \\ \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \nu \partial \alpha} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \nu \partial \alpha_0} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \nu^2} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \nu \partial \gamma} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \nu \partial p} \\ \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \gamma \partial \alpha} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \gamma \partial \alpha_0} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \gamma \partial \nu} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \gamma^2} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial \gamma \partial p} \\ \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial p \partial \alpha} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial p \partial \alpha_0} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial p \partial \nu} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial p \partial \gamma} & \frac{\partial^2 l(\boldsymbol{\Omega})}{\partial p^2} \end{pmatrix}.$$

The square-root of the diagonal elements of the inverse of this matrix are the standard errors of the parameter estimates.

4.4 DIFFERENTIALLY EXPRESSED GENES

Here we describe our method of identifying differentially expressed genes using the latent variable model that we discussed. We consider a gene is differentially expressed if it has high expected posterior probability or simply the posterior probability of differential expression. Recall that the posterior probability of differential expression of the j th gene is $(\widehat{w_j})$, given by the equation (28). Thus we calculate $(\widehat{w_j})$ for every gene in the microarray and rank them to identify high likely or least likely differentially expressed genes.

4.5 SIMULATION STUDY

We conducted a simulation study to check the parameter estimation method for the Bayesian Gaussian Copula model with a latent variable. For these simulations we took $\mathbf{\Omega} = (\alpha, \alpha_0, \nu, \gamma, p) = (2, 3, 15, 0.8, 0.04)$. Random samples of sizes $n = 100, 500, 3000$ are taken following the steps given below.

Step 1 Generate n pairs of bivariate normal random variables $(x_{1j},\ x_{2j})$ from standard bivariate normal distribution (BVN) with correlation parameter γ.

Step 2 Calculate $(u_{1i}, u_{2i}) = (\Phi(x_{1i}), \Phi(x_{2i}))$ for $j = 1, \ldots, n$ where Φ is the cumulative distribution function of the standard normal.

Step 3 Generate $\theta_{ij} \sim$ gamma $(\alpha_0, 1/\nu)$ for $i = 1, 2$ and $j = 1, \ldots, n_d$ and another set with $\theta_j \sim$ gamma $(\alpha_0, 1/\nu)$ for $j = 1, \ldots, n - n_d$ where $n_d = np$, the number of differentially expressed genes.

Step 4 Calculate $(r_{1i}, r_{2i}) = \left(F_{1j}^{-1}(u_{1i}), F_{2j}^{-1}(u_{2i})\right)$ where $F_{ij}(.)$ is the cumulative distribution function of a gamma distribution with parameters $(\alpha, 1/\theta_{ij})$ for the first n_d of pairs of (u_{1i}, u_{2i}) and with parameters $(\alpha, 1/\theta_j)$ for the remaining $(n - n_d)$ pairs observations.

Recall that for simulating the data we chose the parameter values as $\alpha = 2, \alpha_0 = 3, \nu = 15, \gamma = 0.8$ and $p = 4\%$. Unlike the simulations that we did in Chapter 3, the differentially expressed genes are known in this simulation study. Therefore, the sensitivity of the model can be calculated as the ratio of the correctly identified differentially expressed genes by the model, (say truly identified genes, TI), to the total number of actual differentially expressed genes (AD). Thus the sensitivity measure is defined as,

$$Sensitivity = \frac{TI}{AD}. \tag{29}$$

The results of applying proposed Bayesian Gaussian copula with a latent variable on the simulated sample data are given in the Table 10.

Table 10. Parameter estimates (standard errors) for the simulated data†.

n	α	α_0	ν	γ	$p(\%)$	Sensitivity
100	1.66	2.58	16.12	0.83	3.889	0.50
	(0.208)	(0.347)	(3.508)	(0.034)	(0.148)	
500	1.88	2.87	15.52	0.78	4.102	0.75
	(0.117)	(0.171)	(1.143)	(0.019)	(0.074)	
3000	1.94	3.03	14.82	0.80	3.965	0.83
	(0.042)	(0.075)	(0.113)	(0.007)	(0.034)	

†True parameter values are $\alpha = 2, \alpha_0 = 3, \nu = 15, \gamma = 0.8$ and $p = 4\%$.

An examination of the values in Table 10 shows that the estimate of p is close to the true value for all sample sizes, and the estimates of the other parameters are getting closer to the true values as the sample size increases. Furthermore, the standard errors are getting smaller with increased sample size, for example, the standard error of $\widehat{\alpha}$ is $0.028, 0.117$ and 0.042 for sample sizes of $100, 300$ and 3000, respectively. This implies that the proposed Gaussian copula model is consistently estimating the model parameters. When the sample size is bigger, the sensitivity measure is also increasing towards 1.

Table 11. The mean, MSE and bias of MLE's of the parameters.

		mean	MSE	Bias
	$n = 100$	1.736	0.2762	0.264
$\alpha = 2$	$n = 500$	1.849	0.1757	0.151
	$n = 3000$	1.951	0.0184	0.049
	$n = 100$	2.612	0.4395	0.388
$\alpha_0 = 3$	$n = 500$	2.789	0.2474	0.211
	$n = 3000$	3.052	0.0696	0.052
	$n = 100$	16.074	3.2717	1.074
$\nu = 15$	$n = 500$	15.434	1.3233	0.434
	$n = 3000$	14.834	0.1390	0.166
	$n = 100$	0.830	0.0124	0.030
$\gamma = 0.8$	$n = 500$	0.783	0.0141	0.017
	$n = 3000$	0.797	0.0090	0.003
	$n = 100$	3.880	0.2264	0.120
$p = 4$	$n = 500$	4.094	0.1334	0.094
	$n = 3000$	4.023	0.0210	0.023

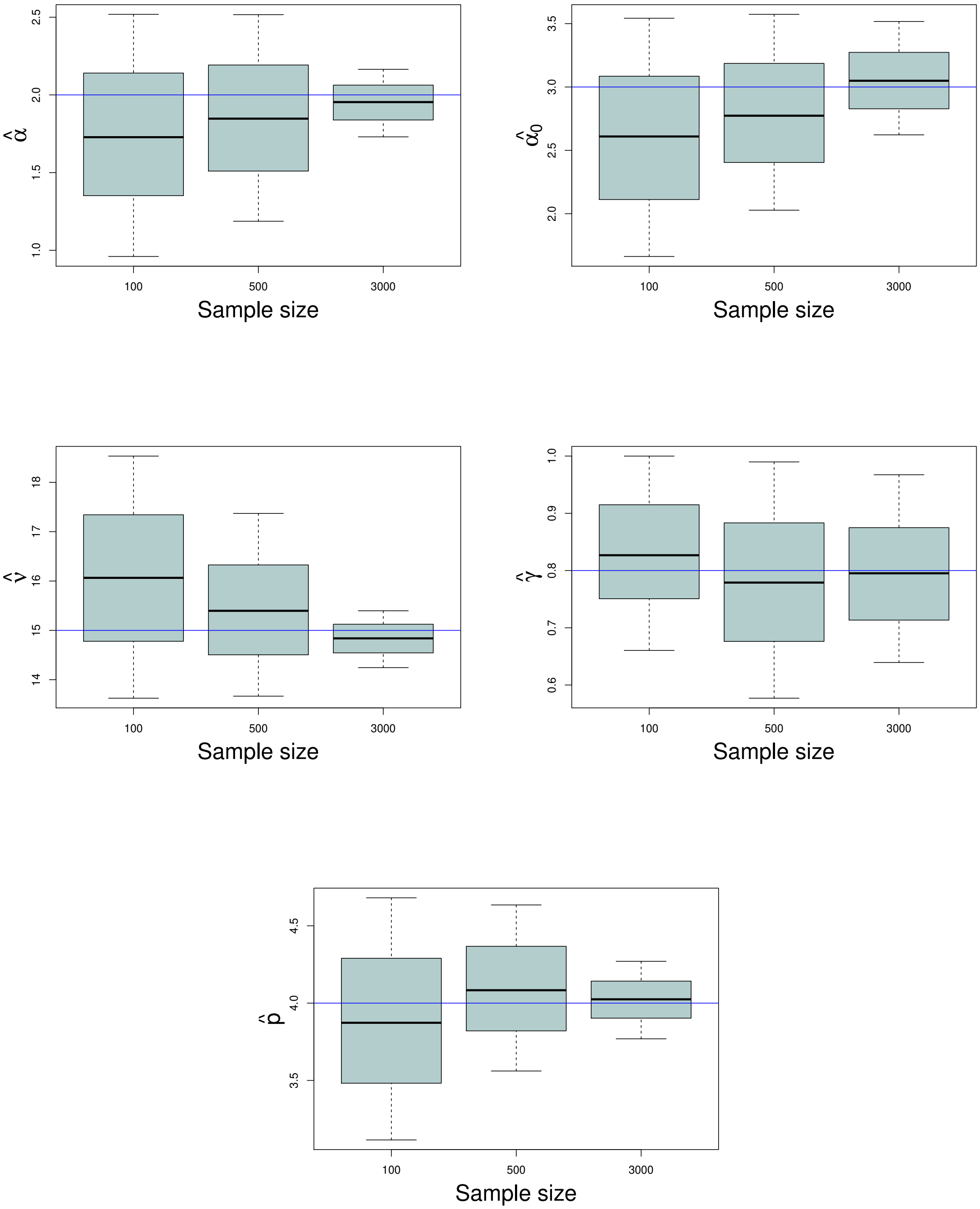


Figure 15. Boxplots of parameter estimates created using bootstrap samples.

To study the behavior of bias and mean square error, we took 1000 samples each of size $n_s = 30$ with replacement from the simulated data. Using these subsamples we computed the bias and the mean squared error (MSE) and bias for each parameter in the model. The results are summarized in Table 11 and the boxplots of parameter estimates are shown in Figure 15. The boxplots are visual evidence to see estimated parameters are getting closer to the true value and the variation getting smaller when the sample size increases. The results in Table 11 also confirmed the same fact. Thus, our simulations suggest that the proposed copula model in this chapter performs better with larger sample sizes.

4.6 ANALYSIS OF *E. COLI* DATA

We apply the developed Bayesian Gaussian copula model with a latent variable to *E. coli* data. This data consists of five samples, control; two IPTG treated samples and two Heat Shock samples. All these data sets were described in Section 3.2. According to Richmond et al. (1999), the control sample has none of the differentially expressed genes, IPTG samples have few, and Heat shock samples have a large number of differentially expressed genes.

Table 12 contains the estimates and the standard errors of those estimates obtained for the five microarray samples in *E. coli* data. As expected, a tiny proportion of genes have differential expression in the control, IPTG-A, and IPTG-B microarrays. In contrast, the proportion of differentially expressed genes is relatively high for the Heat Shock-A and Heat Shock-B microarrays. The standard errors of the estimates of the five microarray samples are relatively small, which suggests the uncertainty associated with each sample statistic is small.

The empirical and estimated marginal densities obtained from estimated parameters with the Gaussian copula incorporate latent variable and the Gaussian copula described in Chapter 3 are superimposed and presented in Figures 16 and 17, separately for red and green intensities. Curves shown in solid lines are going along with the shaded curves more than the dashed curves. Note here, the solid curves are for the Gaussian copula incorporate latent variable, and dashed curves are the Gaussian copula (in Chapter 3). This suggests an improvement of Bayesian Gaussian copula with latent variable over the model without latent variable.

Table 12. Parameter estimates (standard errors) for the *E. coli* data.

Microarray	α	α_0	ν	γ	$p(\%)$
Control	2.06	3.75	21.93	0.94	0.0003
	(0.022)	(0.172)	(1.252)	(0.012)	(<0.001)
IPTG-A	1.26	6.32	67.42	0.95	0.0103
	(0.021)	(0.022)	(2.119)	(0.020)	(0.023)
IPTG-B	1.10	4.95	50.34	0.94	0.0105
	(0.014)	(0.064)	(1.219)	(0.017)	(0.048)
Heat Shock-A	2.07	3.50	14.86	0.44	4.0004
	(0.064)	(0.102)	(0.340)	(0.016)	(0.054)
Heat Shock-B	1.69	2.16	9.97	0.41	3.9908
	(0.034)	(0.061)	(0.355)	(0.015)	(0.038)

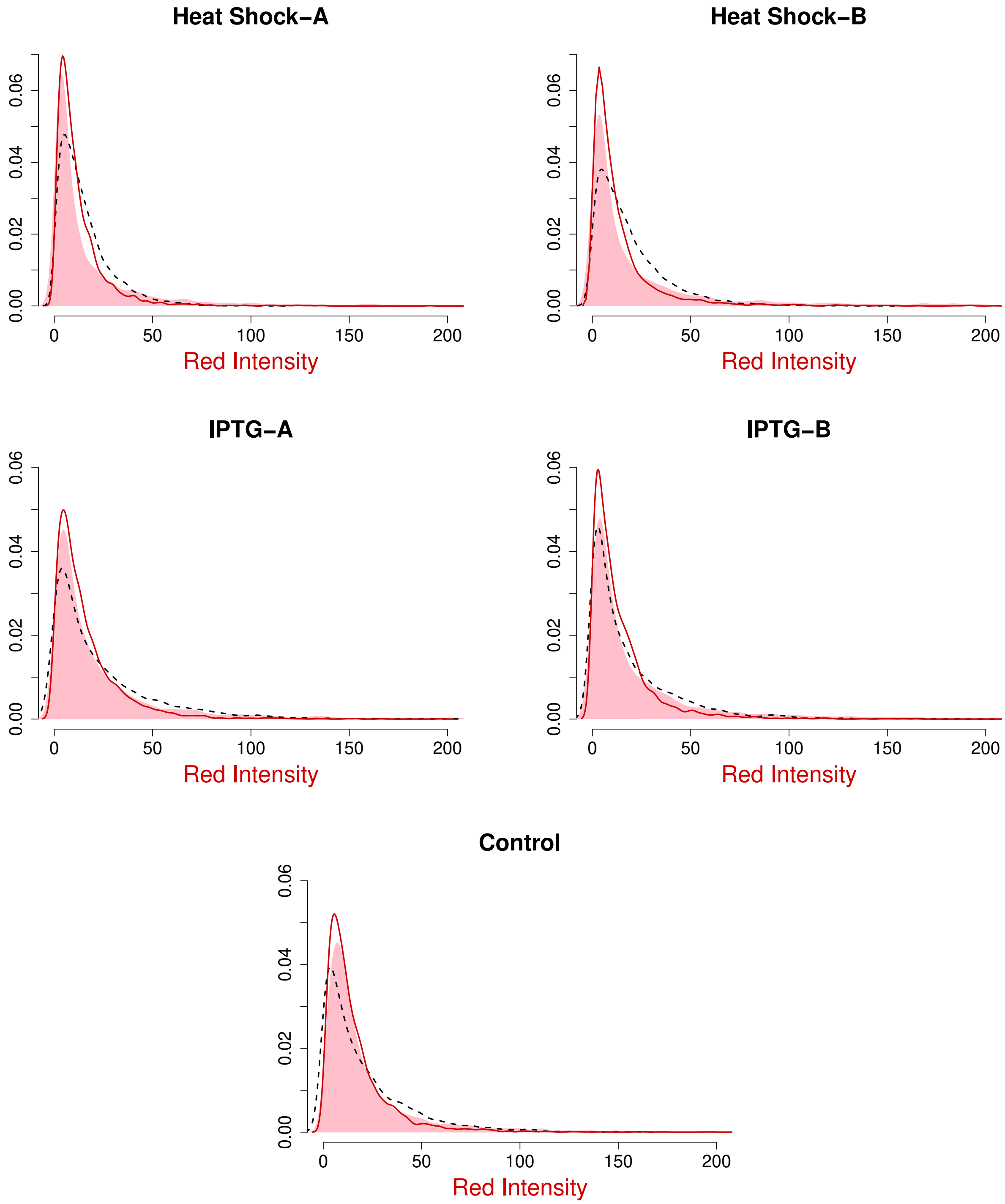


Figure 16. Density plots of red intensities.

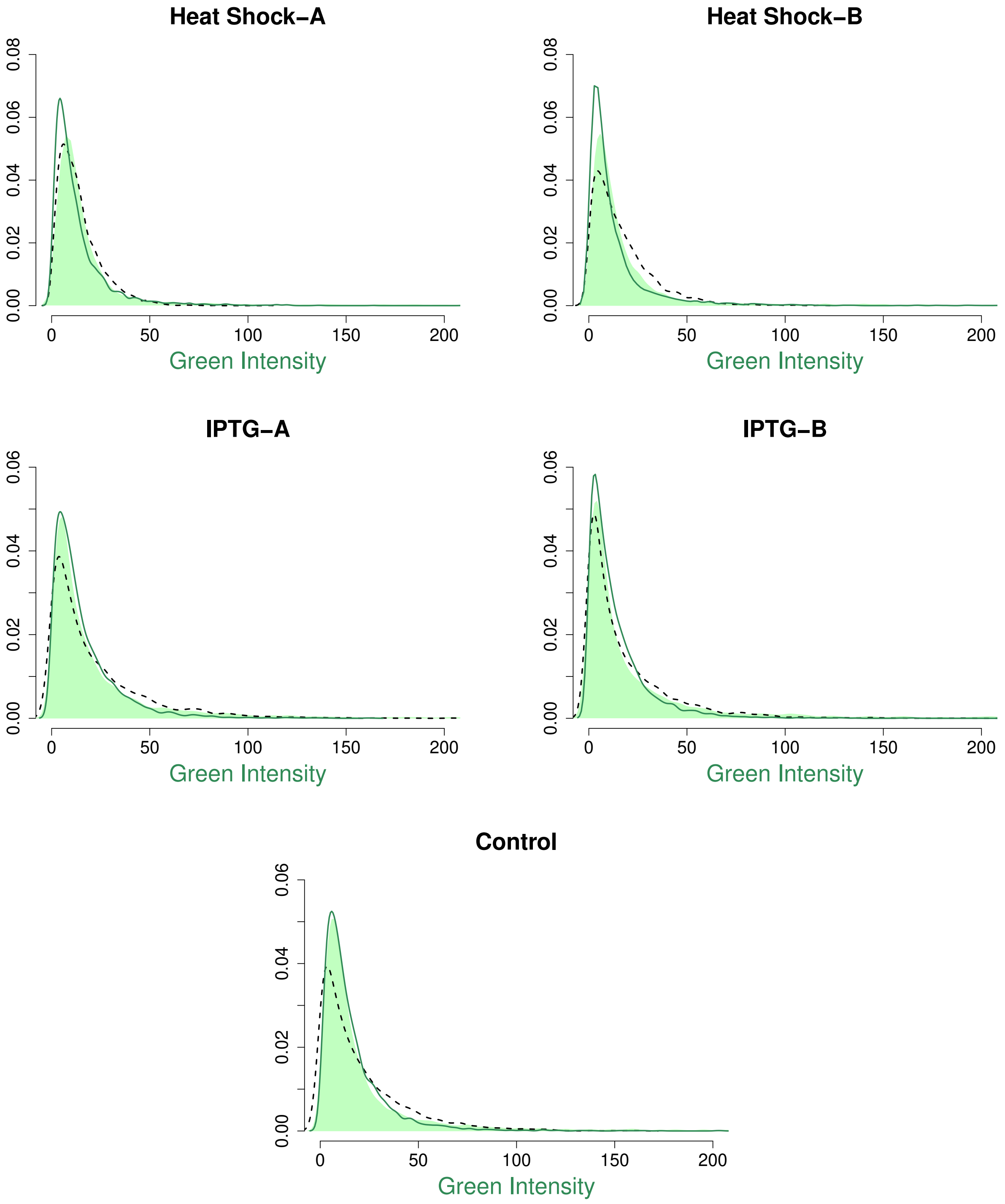


Figure 17. Density plots of green intensities.

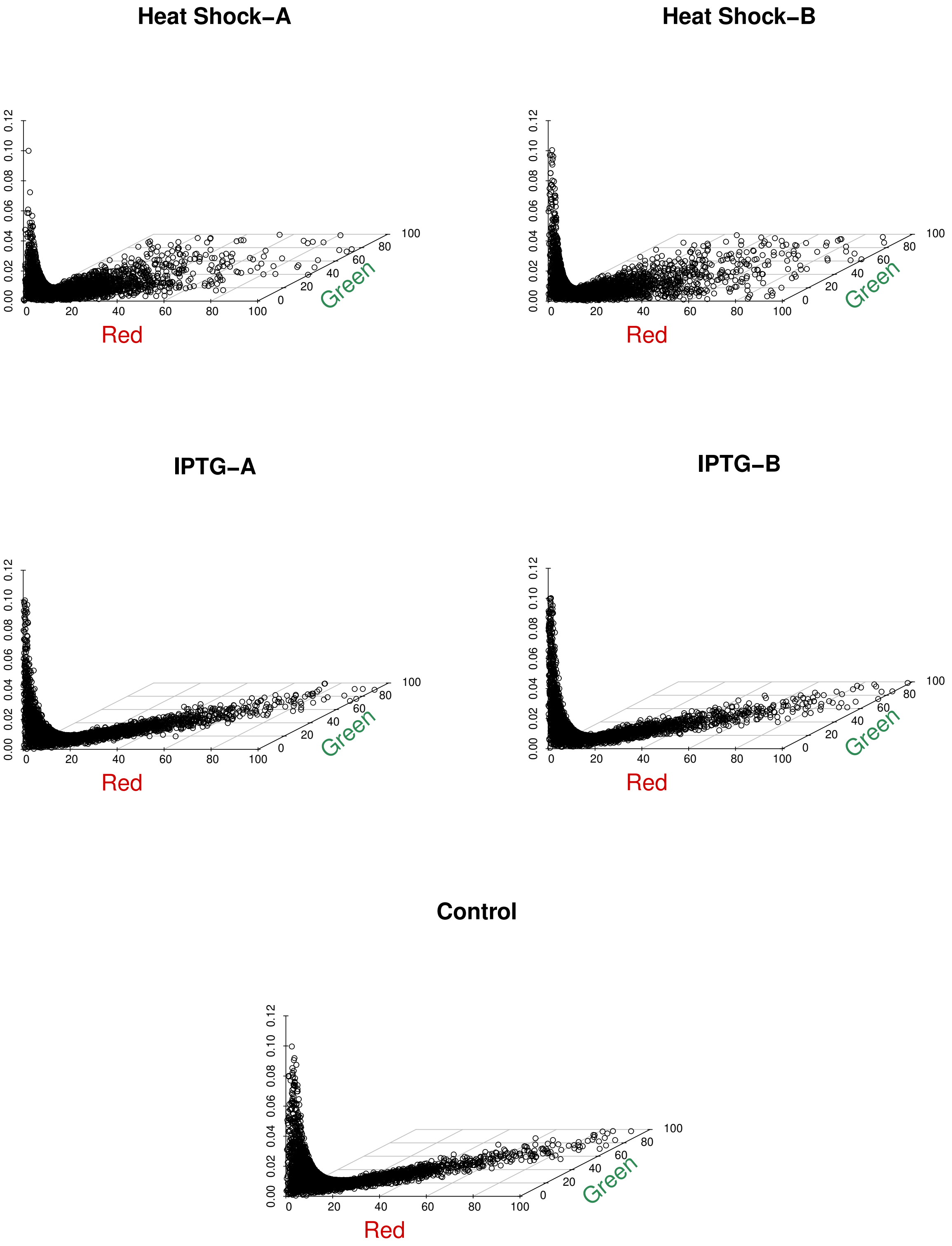
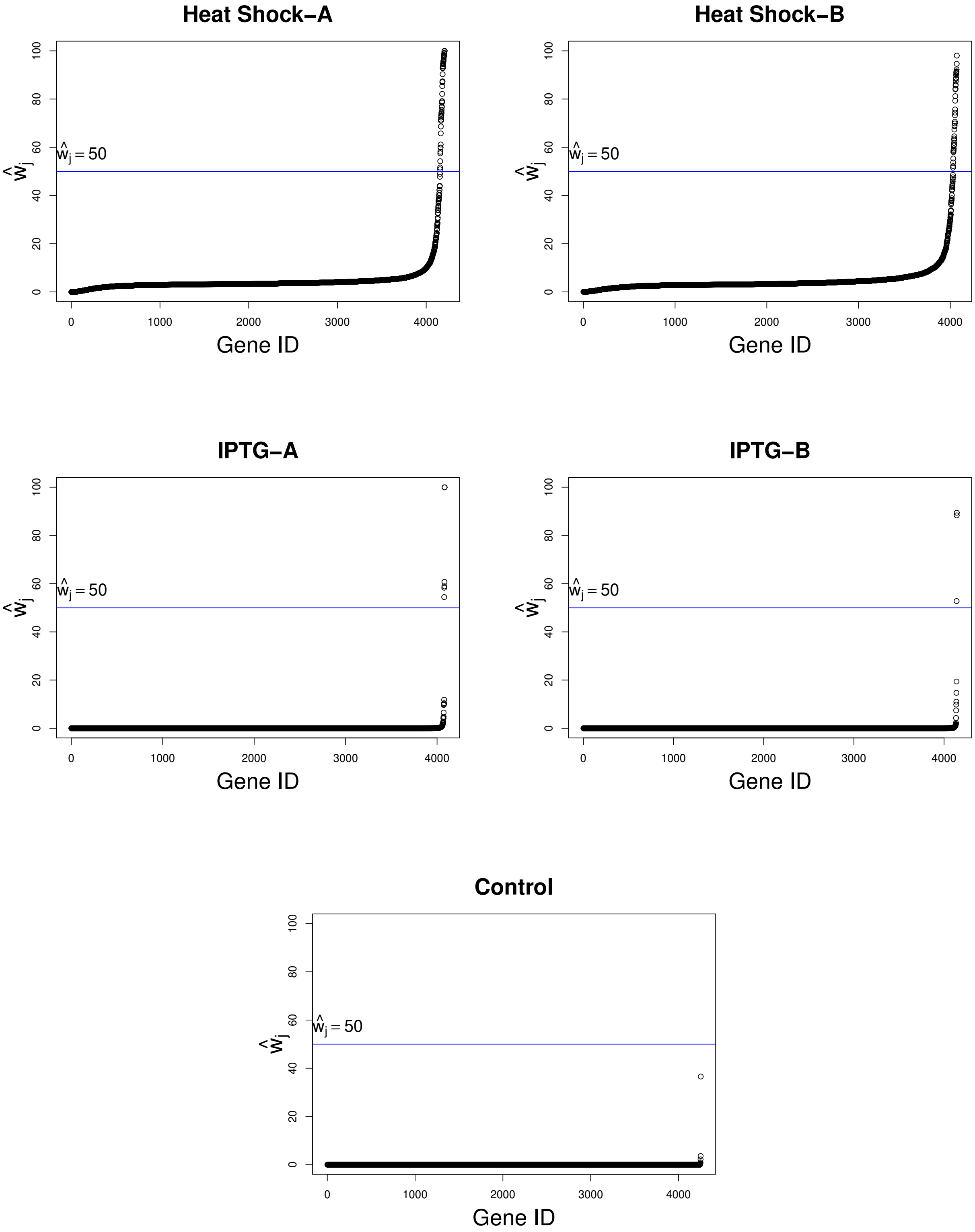


Figure 18. Estimated bivariate density plots of red and green intensities.

The estimated bivariate densities for each microarray are plotted and shown in Figure 18. As mentioned in Chapter 3, the points on the 45^0 line represent equal red and green intensities. The points that are far from the 45^0 line correspond to the differentially expressed genes. Most of the points in control and IPTG samples lie on the 45^0 line, while Heat Shock samples have a relatively large number of points away from the 45^0 line.

Table 13. Top 20 genes with highest posterior probabilities of differentially expression.

	Control		IPTG-A		IPTG-B		Heat Shock-A		Heat Shock-B	
#	Gene	Prob	Gene	Prob	Gene	Prob	Gene	Prob	Gene	Prob
	id	(%)	id	(%)	id	(%)	id	(%)	id	(%)
1	b0233	36.56	b4098	99.97	b4119	89.42	b3686	99.98	b3686	98.01
2	b1325	3.61	b4119	99.92	b4120	88.33	b3687	99.98	b3687	94.62
3	b4325	2.19	b4120	60.76	b4149	52.77	b0014	99.87	b4142	92.31
4	b0558	0.97	b1256	58.89	b1256	19.41	b4142	99.81	b1321	91.52
5	b1319	0.55	b2206	58.30	b2206	14.72	b0015	99.13	b3400	91.32
6	b0542	0.27	b0043	54.44	b4291	11.06	b3400	98.84	b0015	91.11
7	b0657	0.11	b1673	11.86	b0341	9.83	b2592	98.25	b2614	90.78
8	b2740	0.07	b0296	10.43	b0759	7.43	b0582	97.93	b1076	89.91
9	b2051	0.03	b2205	10.03	b1020	4.28	b4143	97.87	b0966	88.79
10	b2129	0.01	b1571	9.87	b1785	2.27	b3401	97.84	b0016	88.46
11	b1447	0.01	b2204	9.69	b0648	1.76	b1967	97.50	b0017	88.36
12	b2358	0.01	b4291	6.52	b1674	1.73	b0473	96.81	b1060	87.35
13	b2418	0.01	b0759	4.81	b2151	1.48	b0016	96.15	b0014	85.95
14	b3818	0.01	b2727	4.30	b2203	1.38	b0439	95.77	b4140	85.64
15	b2387	0.01	b2997	4.13	b2204	1.15	b4171	95.63	b0315	84.26
16	b3834	0.01	b0283	2.71	b2260	1.13	b0399	95.45	b0400	84.00
17	b0185	0.01	b2202	2.60	b0558	1.05	b4140	94.86	b1829	81.21
18	b3616	0.01	b2996	2.60	b2997	1.02	b1321	94.61	b1967	79.31
19	b0295	0.01	b0347	2.08	b2996	0.93	b1829	94.47	b1380	75.67
20	b0548	0.01	b1020	2.06	b2205	0.85	b1076	93.69	b0473	74.38



Heat Shock−A

Heat Shock−B

IPTG−A

IPTG−B

Control

Figure 19. Plots of posterior probability (%).

The five microarray samples were ranked by the $\widehat{w}_j$'s, the posterior probabilities of differential expressions. Table 13 shows the top twenty genes found to be differentially expressed in each microarray. All of the top twenty genes listed under Heat shock samples hold higher posterior probabilities. In comparison, few of the genes have considerably large posterior probabilities for the IPTG samples, and none of the top twenty genes in the control sample exhibits sufficiently large posterior probabilities. These results indicate that the Bayesian Gaussian copula model with a latent variable performs well on *E. coli* data. Moreover, the selection of differentially expressed genes captured by our method is almost identical to those captured from the method suggested by Mav and Chaganty (2004).

The plots of posterior probabilities $\widehat{w}_j$'s for five microarray samples are shown in Figure 19. A reasonable candidate cut-off value of $\widehat{w}_j$ seems to be 50% after considering plots and the fact that the control sample has none of differentially expressed genes. On the other hand, IPTG samples have few, and Heat shock samples have a more significant number of differentially expressed genes.

We present in Table 14 the total number of genes identified as differentially expressed by our Gaussian copula with a latent variable. These numbers are contrasted with the findings of Mav and Chaganty (2004), who have used a bivariate gamma distribution.

Table 14. Total number of differentially expressed genes.

Microarray	# of Genes for which $\widehat{w} > 0.5$	
	Bivariate Gamma with a latent variable	Gaussian Copula with a latent variable
Control	0	0
IPTG-A	1	6
IPTG-B	0	3
Heat Shock-A	60	53
Heat Shock-B	42	42

For the most part all the results are consistent with Richmond et al. (1999)'s hypothesis; the control group has none, IPTG samples have a few, and Heat Shock samples have a large number of differential expressed genes. Also, our Gaussian Copula based model could identify 3 differentially expressed genes in IPTG-B sample, which were not captured by the bivariate gamma based model of Mav and Chaganty (2004).

Table 15. Log-likelihoods for the competitive models.

Microarray	Bivariate Gamma	Gaussian Copula with
	a latent variable	a latent variable
Control	-28273	-27781
IPTG-A	-27881	-27423
IPTG-B	-27929	-27302
Heat Shock-A	-31723	-30170
Heat Shock-B	-31158	-30085

The log-likelihood analysis of competitive models is shown in Table 15. For each microarray sample, the log-likelihoods for the Gaussian copula-based model are larger than that of the bivariate gamma model proposed by Mav and Chaganty (2004). Further, the filtered genes, as differentially expressed, are almost the same genes filtered by Mav and Chaganty (2004)'s method. All together, we can conclude that our method has better performance than Mav and Chaganty (2004)'s method.

4.7 MODEL COMPARISONS

Both log-likelihood analyses in Chapter 3 and in this chapter suggest that the Gaussian copula models outperform the corresponding bivariate gamma models proposed by Mav and Chaganty (2004). In this section, we compare the two Bayesian Gaussian copula models in terms of Akaike Information Criteria (AIC). The AIC is defined as

$$AIC \ = \ 2k - 2 \ \log \ L$$

where k is the number of parameters in the model, log L is the maximized value of the log-likelihood function. The constant $2k$ in AIC, penalizes models which have more parameters as a trick to avoid over-fitting. The model with the least AIC is chosen to be the best model. The AICs for two Bayesian Gaussian copula models are presented in Table 16.

Table 16. AIC for the competitive copula models.

Microarray	Gaussian Copula	Gaussian Copula with
		a latent variable
Control	56708	55572
IPTG-A	55714	54856
IPTG-B	55778	54614
Heat Shock-A	60846	60350
Heat Shock-B	60572	60180

The AIC values under the Bayesian Gaussian copula model with a latent variable are always smaller than that of the Bayesian Gaussian copula. Thus, we can conclude that the Bayesian Gaussian copula model with a latent variable performs better than the Bayesian Gaussian copula model discussed in Chapter 3.

4.8 CONCLUSIONS

In this chapter we proposed another Bayesian Gaussian copula model that includes a latent variable. Using simulations we have shown that our model is estimating consistently the model parameters for large sample sizes. Even the sensitivity, that is defined as the ratio of identified vs true number of differentially expressed genes, was increasing with sample size which shows that our model is good.

We applied our model is applied to *E. coli* samples to capture the differentially expressed genes. The higher posterior probability values reflect the differentially expressed genes in this model. The genes filtered as differentially expressed are well-matched with the genes listed in Richmond et al. (1999)'s study. Finally, we compare our method to the bivariate gamma distribution with latent variable (proposed by Mav and Chaganty (2004)) with the genes filtered and the log-likelihood values. The filtered genes were almost the same in both studies, but the log-likelihood values of our method are larger than that of Mav and Chaganty (2004)'s method. So we can conclude that Bayesian Gaussian copula with a latent variable outperforms.

The lower AICs in the Bayesian Gaussian copula with a latent with compared to that of the Bayesian Gaussian copula (discussed in Chapter 3) confirm the better performance of the Bayesian Gaussian copula with a latent over the other Gaussian copula model. In the next chapter, we will explore the use of Weibull marginals in the Bayesian Gaussian copula with a latent variable.

CHAPTER 5

BAYESIAN COPULA MODEL WITH WEIBULL MARGINALS

5.1 INTRODUCTION

In Chapter 4, we introduced a Bayesian Gaussian copula model with a latent variable that is capable of capturing the differentially expressed genes in a cDNA microarray. However, the discussion was limited to red and green intensities with gamma marginals to compare the results to the models proposed in Mav and Chaganty (2004). Moreover, we found that the model discussed in the previous chapter outperformed the Bayesian Gaussian copula model in Chapter 3. Therefore, in this chapter, we consider the Bayesian Gaussian copula model with a latent variable with Weibull marginals to study the capability of detecting differentially expressed genes.

The Weibull distribution is a continuous probability distribution that can fit a variety of distribution shapes. Its extreme flexibility allows it to model both left- and right-skewed data. Even it can approximate the normal distribution and many other distributions. Examples of different distributional shapes are shown in Figure 20. There are two types of this distribution: the three-parameter Weibull distribution and the two-parameter Weibull distribution. In this chapter, we use the two-parameter Weibull distribution as the marginals of the Gaussian copula model. The formula for the probability density function of the two-parameter general Weibull distribution is:

$$f(r;\alpha,\beta) \;=\; \frac{\alpha}{\beta}\left(\frac{r}{\beta}\right)^{\alpha-1}\exp\left[-\left(\frac{r}{\beta}\right)^{\alpha}\right], \qquad (30)$$

where $\alpha > 0$ is the shape parameter and $\beta > 0$ is the scale parameter of the distribution.

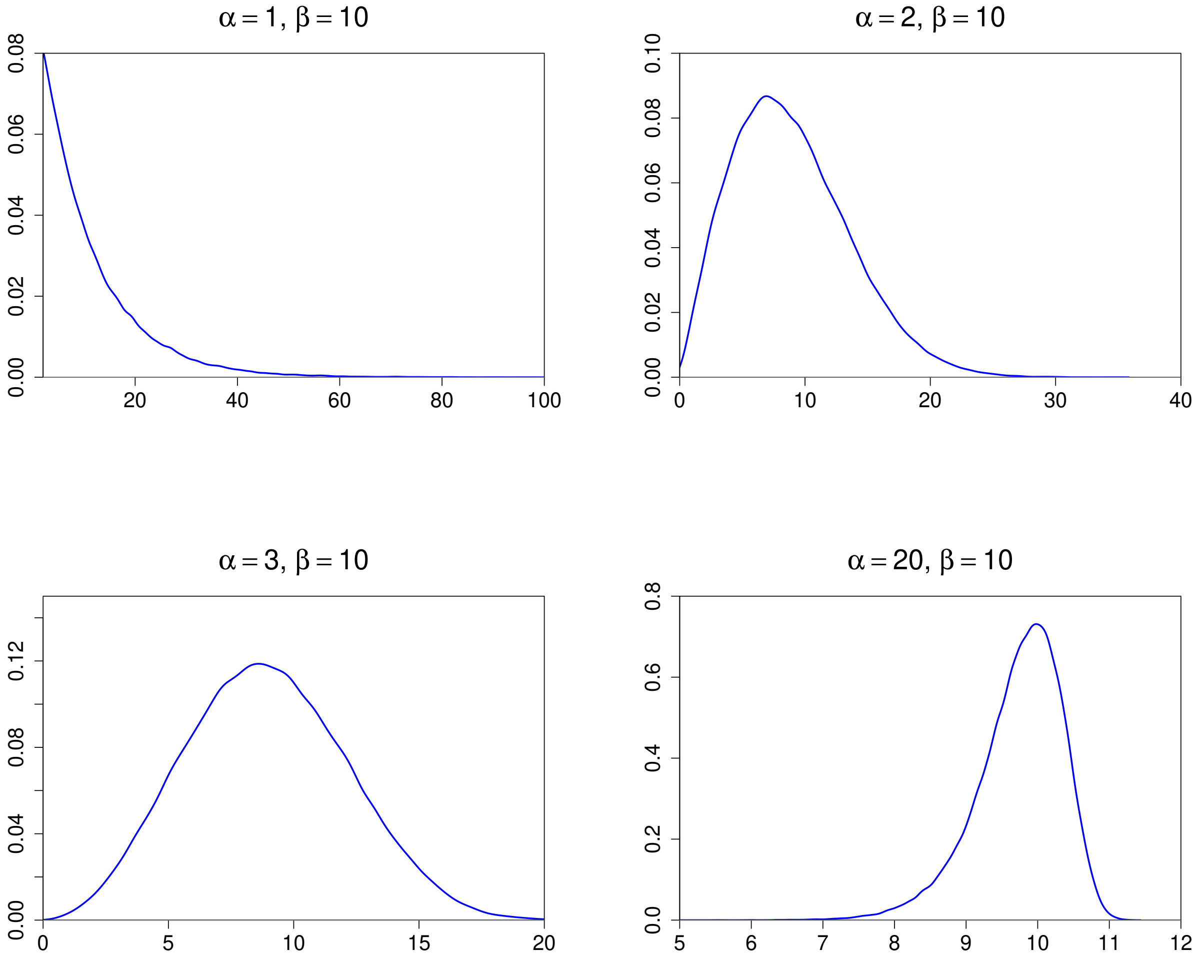


Figure 20. Different distributional shapes of Weibull distribution.

5.2 BAYESIAN COPULA MODEL WITH A LATENT VARIABLE AND WEIBULL MARGINALS

With the usual notations, we assume the marginal distributions of red (R_{1j}) and green (R_{2j}) intensities are distributed as Weibull with common shape parameter α and different scale parameters $1/\theta_{1j}$ and $1/\theta_{2j}$ respectively for $1, 2, \ldots, n$. The probability density function of (R_{ij}) has the following form

$$f_i(r_{ij};\theta_{ij},\alpha) \;=\; \alpha\,\theta_{ij}\,(r_{ij}\theta_{ij})^{\alpha-1}\exp\left[-(\theta_{ij}\,r_{ij})^{\alpha}\right], \quad i=1,2;\; j=1,\ldots,n. \tag{31}$$

We also assume the prior distributions for θ_{ij}'s are independent Weibulls with parameters α_0 and $1/\nu$, and the prior pdf is given by

$$\pi(\theta_{ij};\nu,\alpha_0) \;=\; \alpha_0\,\nu\,(\theta_{ij}\nu)^{\alpha_0-1}\exp\left[-(\nu\,\theta_{ij})^{\alpha_0}\right], \quad i=1,2;\; j=1,\ldots,n. \tag{32}$$

As before we assume there is an unknown proportion p of genes that exhibit differential expression. We introduce for the j th gene an unobserved Bernoulli variable W_j that indicates differential expression as in Section 4.2. With these assumptions for gene j that is differentially expressed ($W_j=1$ and $\theta_{1j}\neq\theta_{2j}$), the joint probability density function of intensities is given by

$$\begin{aligned} f_{de}(r_{1j},\,r_{2j};\boldsymbol{\Upsilon}) \;&=\; (\alpha\,\alpha_0\,\nu^{\alpha_0})^2\int_0^\infty\int_0^\infty c(F_1(r_{1j}),F_2(r_{2j}))\;\times \\ &\prod_{i=1}^{2}\left[r_{ij}^{\alpha-1}\,\theta_{ij}^{\alpha+\alpha_0-1}\,\exp\left[-(r_{ij}\theta_{ij})^{\alpha}-(\nu\theta_{ij})^{\alpha_0}\right]\right]d\theta_{1j}d\theta_{2j}. \end{aligned} \tag{33}$$

Similarly, if the gene j is not differentially expressed ($W_j=0$ and $\theta_{1j}=\theta_{2j}=\theta_j$), then the joint probability density function of intensities is

$$\begin{aligned} f_{nde}(r_{1j},\,r_{2j};\boldsymbol{\Upsilon}) \;&=\; \alpha^2\,\alpha_0\,\nu^{\alpha_0}\,(r_{1j}r_{2j})^{\alpha-1}\int_0^\infty c(F(r_{1j}),F(r_{2j}))\;\times \\ &\theta_j^{2\alpha+\alpha_0-1}\,\exp\left[-(r_{1j}\theta_j)^{\alpha}-(r_{2j}\theta_j)^{\alpha}-(\nu\theta_j)^{\alpha_0}\right]d\theta_j. \end{aligned} \tag{34}$$

Here $F_i(r_{ij})$ and $f_i(r_{ij})$ are the cumulative and densities functions of Weibull$(\alpha,1/\theta_{ij})$. And $F(r_{ij})$ and $f(r_{ij})$ are the cumulative and density functions of Weibull$(\alpha,1/\theta_j)$ for $i=1,2$ and $j=1,2,\ldots,n$ respectively.

5.3 PARAMETER ESTIMATION PROCEDURE

For the model described in Section 5.2, the log-likelihood is

$$l(\boldsymbol{\Upsilon},p)=\sum_{j=1}^{n}\log\left\{f_{de}(r_{1j},\,r_{2j};\boldsymbol{\Upsilon})^{w_j}f_{nde}(r_{1j},\,r_{2j};\boldsymbol{\Upsilon})^{1-w_j}p^{w_j}(1-p)^{1-w_j}\right\}, \tag{35}$$

where $(\boldsymbol{\Upsilon},p)=(\alpha,\alpha_0,\nu,\gamma,p)$ is the parameter vector. Since w_j's are unobserved we use the EM algorithm to obtain the maximum likelihood estimates of the parameters as in Section 4.3.

5.4 DIFFERENTIALLY EXPRESSED GENES

As described in Section 4.4, a gene is considered to be differentially expressed if the posterior probability $\widehat{w_j}$ exceeds a threshold value. Recall, as given in (28) the formula for $(\widehat{w}_j)$ is

$$\widehat{w}_j = E(w_j \,|\, r_{1j},\, r_{2j}) = \frac{\widehat{p}\, f_{de}(r_{1j},\, r_{2j}; \widehat{\boldsymbol{\Upsilon}})}{\widehat{p}\, f_{de}(r_{1j},\, r_{2j}; \widehat{\boldsymbol{\Upsilon}}) + (1-\widehat{p})\, f_{nde}(r_{1j},\, r_{2j}; \widehat{\boldsymbol{\Upsilon}})}.$$

As before we calculate $(\widehat{w}_j)$ for each microarray and rank order them to filter the differentially expressed genes.

5.5 SIMULATION STUDY

In this section we conduct a simulation study to check the parameter estimation for the Bayesian Gaussian Copula model with Weibull marginals. We took the parameter values as $\boldsymbol{\Omega} = (\alpha, \alpha_0, \nu, \gamma, p) = (1.5, 2, 10, 0.7, 0.04)$ and simulated three random sample of sizes $n = 100, 500, 3000$ following the steps outlined below.

Step 1 Generate n pairs of bivariate normal random variables $(x_{1j},\ x_{2j})$ from standard bivariate normal distribution (BVN) with correlation parameter γ.

Step 2 Calculate $(u_{1i}, u_{2i}) = (\Phi(x_{1i}), \Phi(x_{2i}))$ for $j = 1, \ldots, n$ where Φ is the cumulative distribution function of standard normal.

Step 3 Generate $\theta_{ij} \sim$ Weibull$(\alpha_0, 1/\nu)$ for $i = 1, 2$ and $j = 1, \ldots, n_d = np$, and another set with $\theta_j \sim$ Weibull$(\alpha_0, 1/\nu)$ for $j = 1, \ldots, n - n_d$. Note n_d is the number of differentially expressed genes in the sample of size n.

Step 4 Calculate $(r_{1i}, r_{2i}) = \left(F_1^{-1}(u_{1i}), F_2^{-1}(u_{2i})\right)$ where $F_i(.)$ is the cumulative distribution function of a Weibull distribution with parameters $(\alpha, 1/\theta_{ij})$ for the first n_d observations and Weibull with parameters $(\alpha, 1/\theta_j)$ for the remaining $n - n_d$ observations.

The results of our simulation are presented in Table 17.

Table 17. Parameter estimates (standard errors) for the simulated data†.

n	α	α_0	ν	γ	$p(\%)$	Sensitivity
100	1.61	1.86	9.34	0.62	3.847	0.25
	(0.358)	(0.391)	(0.565)	(0.187)	(0.314)	
500	1.51	1.88	9.76	0.65	4.048	0.80
	(0.113)	(0.206)	(0.359)	(0.138)	(0.116)	
3000	1.49	2.08	10.06	0.73	3.984	0.88
	(0.093)	(0.133)	(0.069)	(0.104)	(0.085)	

†True parameter values are $\alpha = 1.5, \alpha_0 = 2, \nu = 10, \gamma = 0.7$ and $p = 4\%$.

The simulation results are similar to what we have observed in Table 10. As the sample size increases, the estimates are getting closer to the true values and the standard errors are becoming smaller. The sensitivity in this model seems higher than that of the model with gamma marginals even for smaller sample size $n = 500$.

To study the bias and mean squared error (MSE) of the parameter estimates, we took 1000 replicates with sub-samples drawn with replacement of size $n_s = 30$ from the simulated data. With these replicates we calculated bias and mean square errors of the parameter estimates. The results are presented in Table 18 and the box plots are in Figure 21. The bias and MSE for all parameter estimates are decreasing as the sample size increases. This establishes consistency of the estimation procedure. And thus this simulation study provides evidence that the Bayesian Gaussian Copula model with Weibull marginals is a good model for large sample sizes.

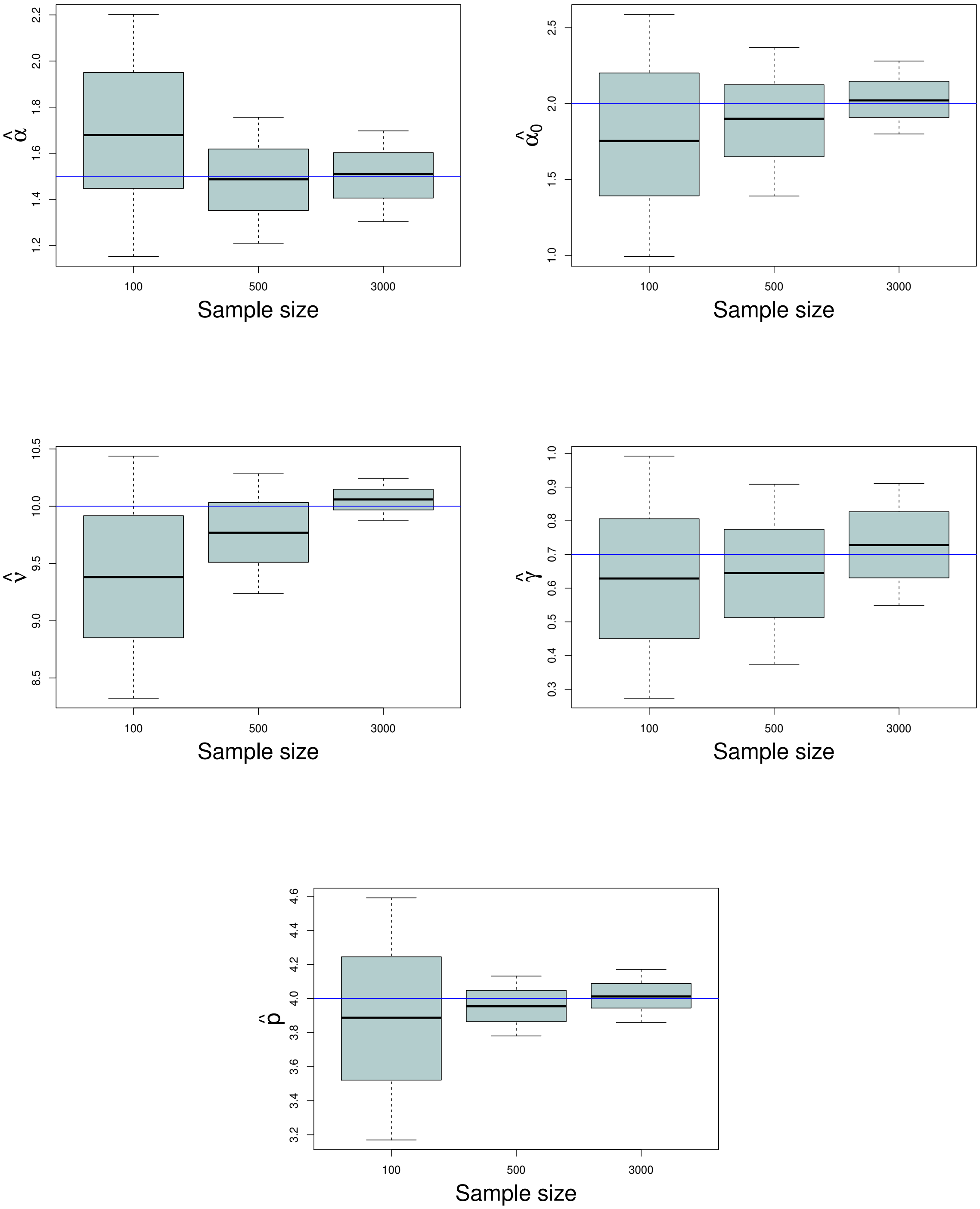


Figure 21. Boxplots of parameter estimates created using bootstrap samples

Table 18. The mean, MSE and bias of mle of parameters.

		Mean	MSE	Bias
	$n =$ 100	1.682	0.3005	0.182
$\alpha = 1.5$	$n =$ 500	1.486	0.1560	0.014
	$n = 3000$	1.504	0.1144	0.004
	$n =$ 100	1.791	0.4667	0.209
$\alpha_0 = 2$	$n =$ 500	1.887	0.2827	0.113
	$n = 3000$	2.029	0.1357	0.029
	$n =$ 100	9.379	0.6138	0.621
$\nu = 10$	$n =$ 500	9.768	0.3029	0.232
	$n = 3000$	10.058	0.1056	0.058
	$n =$ 100	0.631	0.2083	0.069
$\gamma = 0.7$	$n =$ 500	0.643	0.1520	0.057
	$n = 3000$	0.730	0.1073	0.030
	$n =$ 100	3.886	0.4120	0.114
$p = 4$	$n =$ 500	3.956	0.1031	0.044
	$n = 3000$	4.014	0.0873	0.014

5.6 MISSPECIFICATION STUDY

The objective of this misspecification study is to study the robustness of the models that we have discussed. In particular, we would be interested in knowing the effect on identifying differentially expressed genes if the true marginal distributions are gamma but misspecified as Weibull or vice versa. First we consider the simulated data in Section 4.5 with gamma marginals with true parameter values as $\alpha = 2, \alpha_0 = 3, \nu = 15, \gamma = 0.8, p = 4\%$. We misspecify and fit the model with Weibull marginals for this simulated data. The results are summarized in Table 19 and Figure 22. Note here, the solid curves are for the estimated densities from generated data (misspecified) with estimated parameters of the Gaussian Copula in with Weibull marginals, and shaded curves are the empirical densities of simulated data with gamma marginals with $\alpha = 2, \alpha_0 = 3, \nu = 15, \gamma = 0.8$ and $p = 4\%$.

According to the results in Table 19, the standard errors of the misspecified model with Weibull marginals are higher compared to the corresponding standard errors of the model with gamma marginals under each sample size. Moreover, the sensitivity values of the misspecified model are smaller than that of the correctly specified model. However, the misspecified model can identify a considerable amount of truly differentially expressed genes when the sample size is large.

Table 19. The results of misspecification study for the simulated data with gamma marginals with $\alpha = 2, \alpha_0 = 3, \nu = 15, \gamma = 0.8$ and $p = 4\%$.

n		α	α_0	ν	γ	$p(\%)$	Sensitivity
100	Gamma	1.66	2.58	16.12	0.83	3.889	0.50
		(0.208)	(0.347)	(3.508)	(0.034)	(0.148)	
	Weibull	1.41	2.30	9.62	0.75	3.633	0.25
		(0.933)	(0.589)	(3.178)	(0.109)	(0.235)	
500	Gamma	1.88	2.87	15.52	0.78	4.102	0.75
		(0.117)	(0.171)	(1.143)	(0.019)	(0.074)	
	Weibull	1.54	2.42	9.71	0.77	3.873	0.45
		(0.825)	(0.297)	(1.889)	(0.086)	(0.135)	
3000	Gamma	1.94	3.03	14.82	0.80	3.965	0.83
		(0.042)	(0.075)	(0.113)	(0.007)	(0.034)	
	Weibull	1.31	2.63	9.90	0.82	3.952	0.67
		(0.575)	(0.111)	(0.993)	(0.063)	(0.091)	

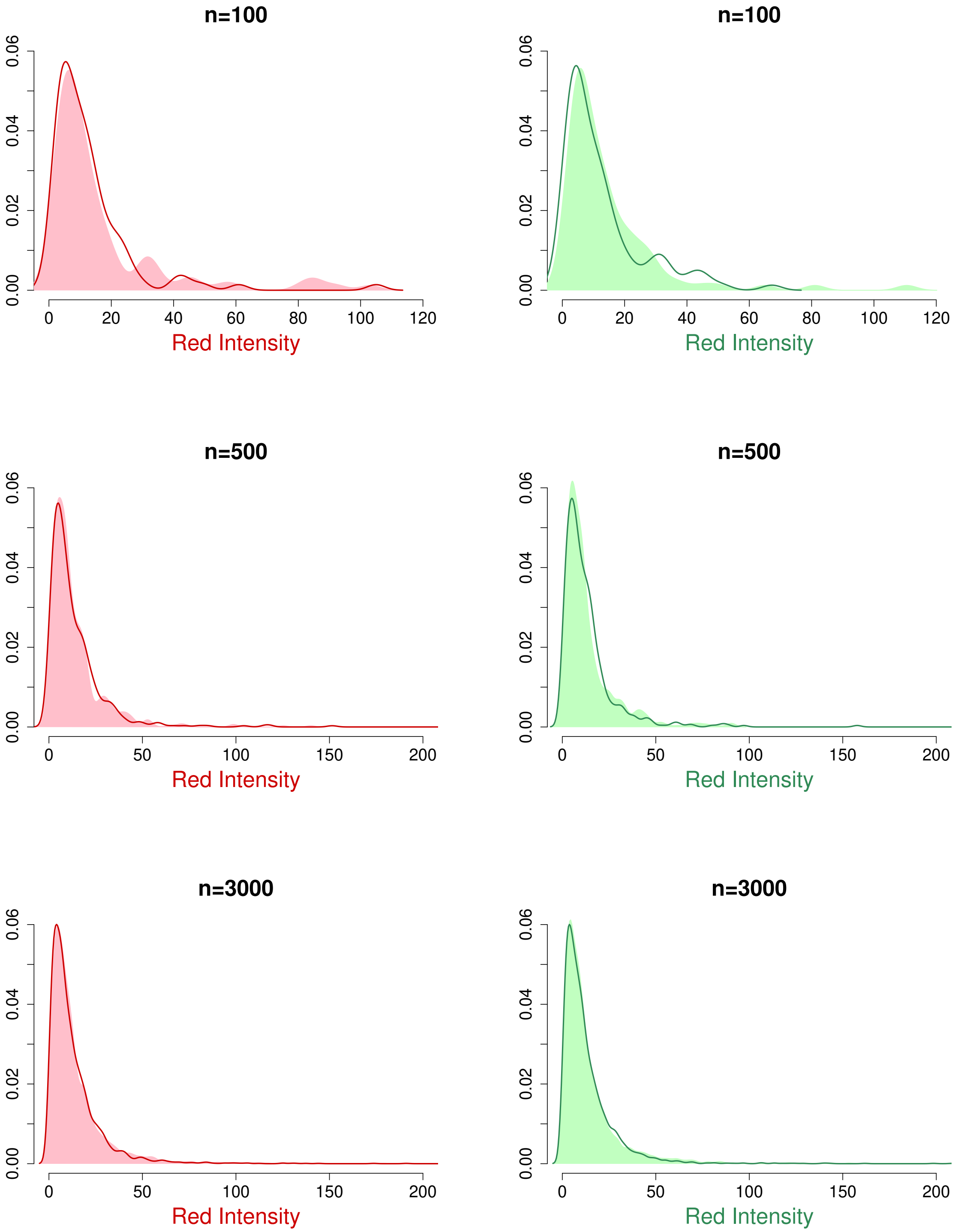


Figure 22. Density plots of simulated data.

For our second misspecification study, we consider the simulated data generated using Weibull marginals in Section 5.5. We misspecify and fit the model with gamma marginals. The results are presented in Table 20 and Figure 23. Note here, the solid curves are for the estimated densities from generated data (misspecified) with estimated parameters of the Gaussian Copula in with gamma marginals, and shaded curves are the empirical densities of simulated data with Weibull marginals with $\alpha = 1.5, \alpha_0 = 2, \nu = 10, \gamma = 0.7$ and $p = 4\%$. From the Table 20, we can observe a similar behavior of standard errors as in the first misspecification study.

Table 20. The results of misspecification study for the simulated data with Weibull marginals with $\alpha = 1.5, \alpha_0 = 2, \nu = 10, \gamma = 0.7$ and $p = 4\%$.

n		α	α_0	ν	γ	$p(\%)$	Sensitivity
100	Weibull	1.61	1.86	9.34	0.62	3.847	0.25
		(0.358)	(0.391)	(0.565)	(0.187)	(0.314)	
	Gamma	2.23	2.34	12.45	0.75	3.699	0.00
		(0.888)	(0.718)	(1.356)	(0.294)	(0.728)	
500	Weibull	1.51	1.88	9.76	0.65	4.048	0.80
		(0.113)	(0.206)	(0.359)	(0.138)	(0.116)	
	Gamma	2.12	2.87	12.30	0.87	3.637	0.15
		(0.738)	(0.685)	(1.007)	(0.186)	(0.645)	
3000	Weibull	1.49	2.08	10.06	0.73	3.984	0.88
		(0.093)	(0.133)	(0.069)	(0.104)	(0.085)	
	Gamma	2.54	3.22	11.71	0.90	3.873	0.34
		(0.266)	(0.435)	(0.854)	(0.119)	(0.321)	

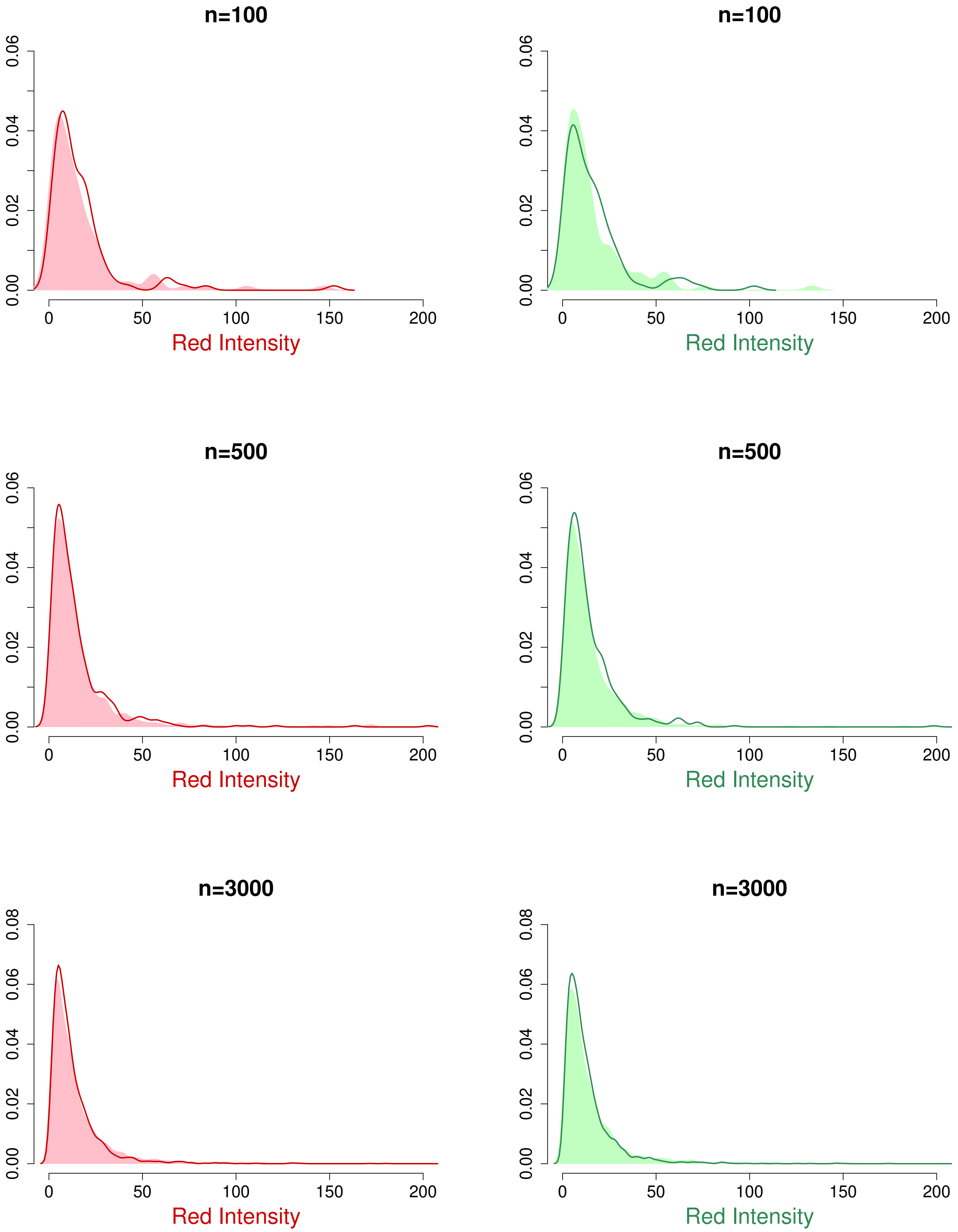


Figure 23. Density plots of simulated data.

Further, we plot the density curves of the fitted misspecified model with gamma marginals and the empirical density of simulated data and present in the Figure 23. Those plots also imply a better fit of the misspecified model with gamma on the simulated data with Weibull marginals. However, the sensitivity values in Table 20 suggest that the misspecified model fails to filter truly differentially expressed genes in the simulated sample even with the higher sample sizes.

In summary, using the Weibull marginals is a robust solution because irrespective of the true marginals, whether gamma or Weibull, the model can correctly identify a large amount of differentially expressed genes. This is a good indication of better performance of Weibull marginals over gamma marginals.

5.7 ANALYSIS OF *E. COLI* DATA

To illustrate the proposed model with Weibull marginals and compare the results to the model with Gamma marginals in the previous chapter, we revisit the *E. coli* data and apply the model. To recap, the data is from Richmond et al. (1999) and consists of five samples, control (with no differentially expressed genes), two IPTG samples (with few differentially expressed genes), and two Heat shock samples (with many differentially expressed genes).

Table 21 provides point estimates and standard errors for the five microarray samples in *E. coli* data. The estimated proportions of genes exhibit differential expression (p) under each microarray sample agree with Richmond et al. (1999)'s findings. Moreover, the standard errors of estimates are also small, similar to what we observed in Chapter 4.

The visual comparison of the estimated density of the proposed model with Weibull marginals, the estimated density of the proposed model with gamma marginals (from Chapter 4) to the empirical density is shown in Figure 24 and Figure 25 for red and green intensities separately. In both sets of density curves, the fitted distributions with Weibull marginals (solid curves) always go alone with the empirical distributions (shaded curves) more than that of the fitted distributions with gamma marginals (black dashed curves). Hence, this is an excellent indication that the copula models with Weibull marginals provide a better fit for the data.

Table 21. Parameter estimates (standard errors) for the *E. coli* data.

Microarray	α	α_0	ν	γ	$p(\%)$
Control	1.70	2.41	13.52	0.95	0.0012
	(0.059)	(0.078)	(0.545)	(0.071)	(0.009)
IPTG-A	1.13	2.52	13.01	0.95	0.1030
	(0.112)	(0.108)	(0.823)	(0.095)	(0.087)
IPTG-B	0.91	2.80	12.73	0.96	0.1050
	(0.153)	(0.076)	(0.754)	(0.102)	(0.124)
Heat Shock-A	1.31	1.49	11.67	0.53	4.1091
	(0.218)	(0.109)	(0.328)	(0.029)	(0.057)
Heat Shock-B	1.54	1.22	8.17	0.49	3.9472
	(0.023)	(0.069)	(0.685)	(0.081)	(0.085)

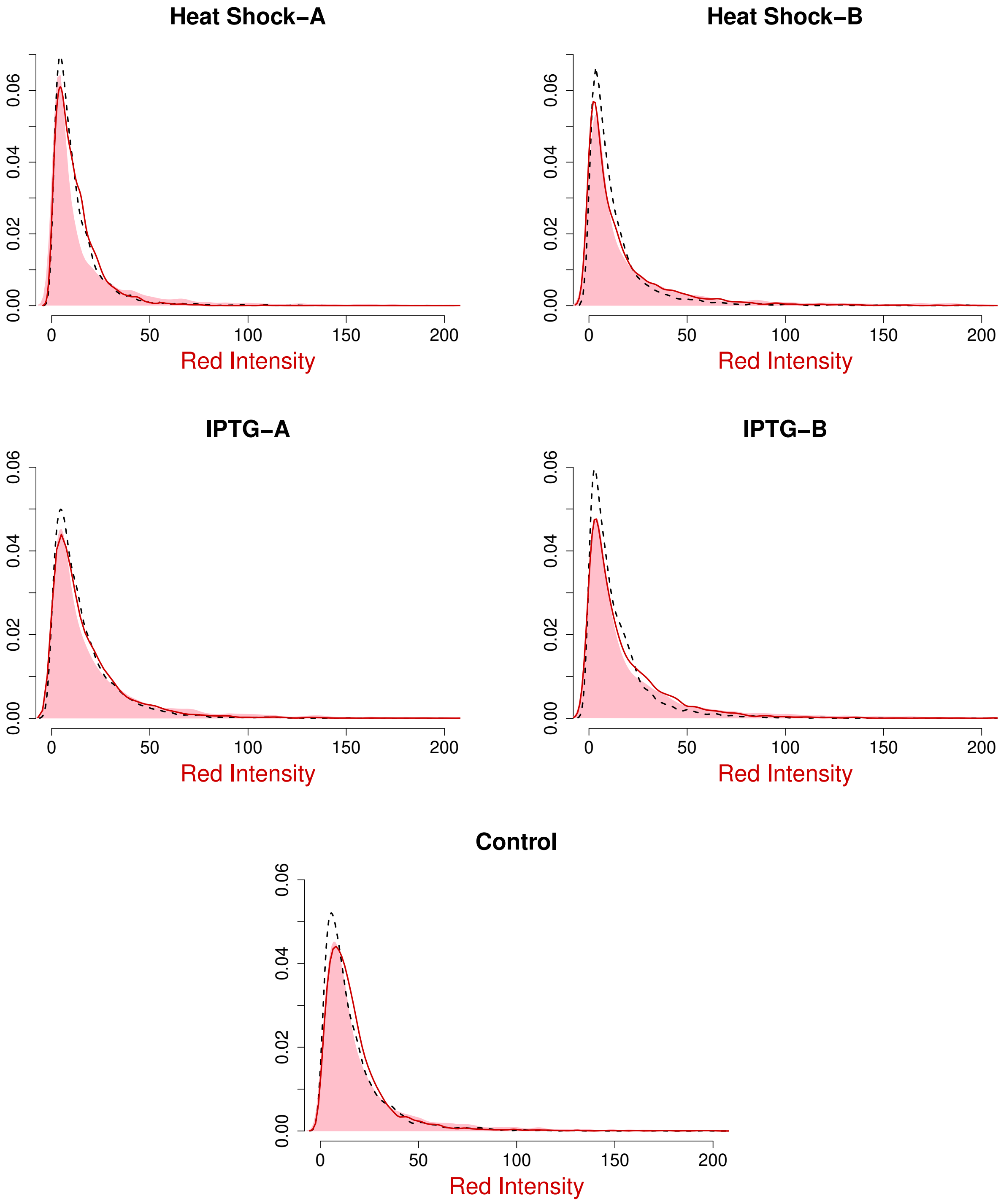


Figure 24. Density plots of red intensities.

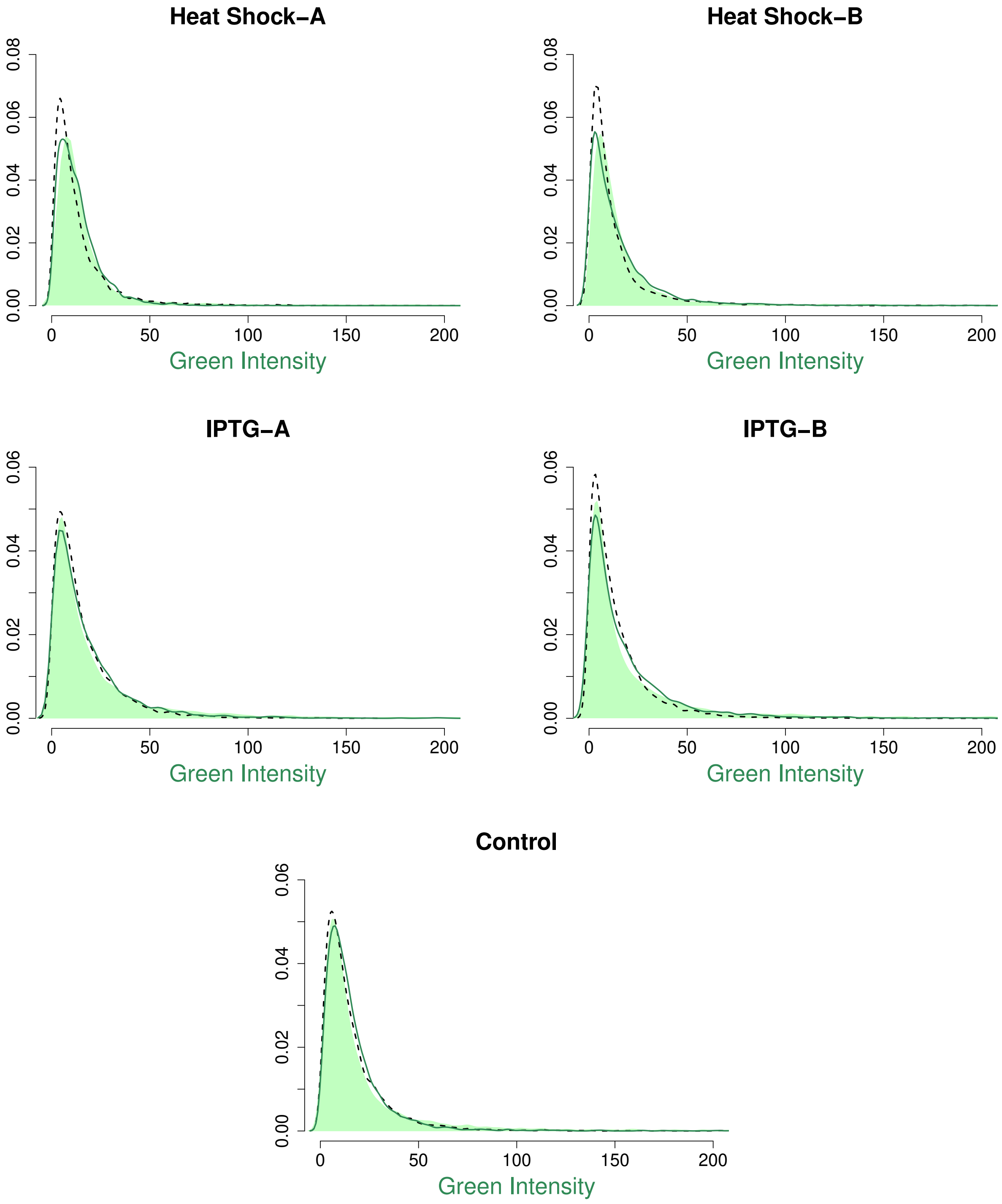


Figure 25. Density plots of green intensities.

Following similar steps in Chapter 4, the posterior probabilities $\widehat{w}_j$'s are plotted and shown in Figure 26. We chose a cut-off value 50% for $\widehat{w}_j$ to identify the differentially expressed genes. The twenty genes with higher posterior probabilities ($\widehat{w}_j$) listed in Table 22. We notice that the captured differentially expressed genes after applying the proposed model on the Heat Shock samples are similar to those captured from the model stated in the previous chapter. However, the order is slightly different. More importantly, this model can identify a differentially expressed gene for IPTG samples that the previous models have failed. When the dataset is double-checked with the genes mentioned as differentially expressed in Richmond et al. (1999)'s paper, we notice that genes are labeled with $b0342, b0343, b0344$, and $b3047$ are missing in the original dataset. Therefore, this might be a reason for capturing fewer differentially expressed genes in IPTG samples for every model proposed through this dissertation. Nevertheless, identifying an additional true differentially expressed gene is a good indication of better performance of Gaussian copula that incorporates a latent Bernoulli variable with Weibull marginals.

By considering the interpretations which are obtained from density plots, posterior probability plots differentially expressed genes listed in Table 22 and the comparison of the total number of genes captured from the Gaussian copula incorporates a latent Bernoulli variable with Weibull marginals, we can conclude that the proposed Gaussian copula includes a latent Bernoulli variable with Weibull marginals provides a better fit and improves the identification of genes.

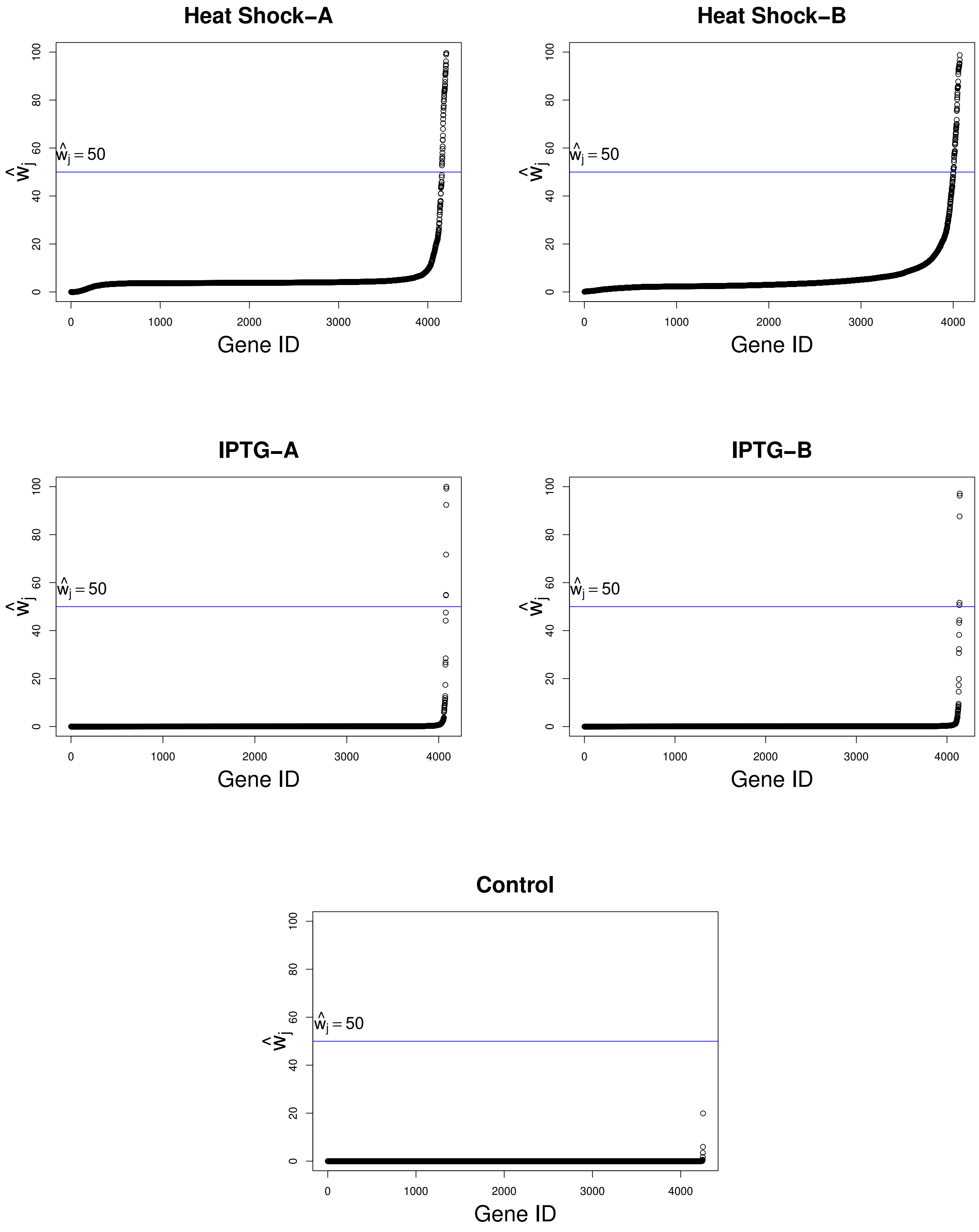


Figure 26. Plots of posterior probability (%).

Table 22. Top 20 genes with highest posterior probabilities of differentially expression.

#	Control		IPTG-A		IPTG-B		Heat Shock-A		Heat Shock-B	
	Gene id	Prob (%)	Gene id	Prob (%)	Gene id	Prob (%)	Gene id	Prob (%)	Gene id	Prob (%)
1	b0233	19.92	b4098	99.99	b4119	97.12	b0014	99.57	b3686	98.75
2	b4325	5.97	b4119	99.20	b4120	96.25	b3687	99.53	b3687	96.70
3	b1325	3.50	b0043	92.42	b4149	87.61	b3686	99.49	b1076	95.30
4	b0558	1.75	b4120	71.69	b1297	51.63	b4142	98.94	b2614	95.11
5	b0657	0.63	b1571	54.90	b1785	50.64	b0015	96.13	b1321	95.07
6	b0542	0.43	b1297	54.69	b2206	44.32	b2592	94.81	b4142	94.83
7	b1319	0.32	b2206	47.51	b0341	43.25	b3400	94.30	b1060	94.04
8	b2740	0.21	b0296	44.13	b4291	38.23	b0582	93.05	b0015	94.03
9	b1447	0.10	b1673	28.44	b0648	32.28	b4143	91.88	b0016	93.98
10	b2129	0.08	b2205	26.75	b0759	30.66	b3401	91.26	b0017	93.85
11	b2418	0.06	b2204	25.81	b1020	19.83	b0016	91.21	b3400	93.42
12	b3616	0.05	b4291	17.39	b4307	17.26	b0473	90.92	b0315	93.17
13	b0185	0.05	b0283	12.71	b0558	14.56	b0399	90.50	b0400	93.06
14	b0679	0.04	b0759	11.88	b1674	9.50	b1967	88.74	b4140	93.02
15	b2628	0.04	b0347	11.15	b2151	8.89	b0439	87.55	b0966	92.72
16	b3834	0.03	b2202	11.03	b2997	8.30	b4171	86.32	b0014	91.80
17	b3818	0.03	b2996	10.10	b2203	8.00	b1829	85.82	b1967	90.73
18	b1064	0.03	b2727	9.76	b2260	7.11	b4140	84.84	b0473	87.70
19	b2051	0.03	b1020	8.91	b0857	7.10	b1321	84.70	b0399	85.82
20	b3147	0.03	b2203	8.81	b2204	6.77	b1076	84.30	b1829	85.61

5.8 MODEL COMPARISONS

This section compares the two Bayesian Gaussian copulas that incorporate a latent Bernoulli variable with Weibull marginals to the same model with gamma marginals described in Chapter 4. Both models have five parameters and exact sample sizes. Therefore, log-likelihood analysis can select the best model among these two candidate Bayesian Gaussian copulas. Table 23 summarizes the results of log-likelihood analysis.

Table 23. Log-likelihoods for the competitive copula models.

Microarray	Gaussian copulas incorporates a latent variable with	
	Gamma marginals	Weibull marginals
Control	-27781	-27136
IPTG-A	-27423	-27025
IPTG-B	-27302	-27153
Heat Shock-A	-30170	-29891
Heat Shock-B	-30085	-29645

For every microarray sample, the differences of log-likelihoods of the competitive models are relatively small. However, the log-likelihoods for the Gaussian copula model incorporate a latent Bernoulli variable with Weibull marginals holding higher values than the Gaussian copula model, including a latent Bernoulli variable with gamma marginals. This implies that the Gaussian copula model incorporates a latent Bernoulli variable with Weibull marginals has better performance than the model with gamma marginals.

5.9 CONCLUSIONS

In summary, in this chapter, we propose a Bayesian Gaussian copula model incorporated with a latent variable which is quite similar to the previous model in Chapter 4, but with a Weibull marginal instead of gamma marginal. The Weibull distribution can fit a variety of distribution shapes like right-skewed, left-skewed, symmetric, and many more. Thus, this Bayesian Gaussian copula model can be applied to many data sets while assuming Weibull marginals. Using a simulation study, we show that this Gaussian copula-based model with Weibull marginals consistently estimates the model parameters for large sample sizes. Further, we conduct a misspecification study to observe the performance of wrongly fitted distribution by using the same simulated data in Sections 4.5 and 5.5. We notice that the misspecified model with Weibull marginals can identify a relatively large amount of truly differentially expressed genes in the simulated data with gamma marginals.

We illustrate the application of our model on samples of *E. coli* data. Comparing the empirical density curve and the fitted density curves of Gaussian copula models with gamma marginals and Weibull marginals suggests that the copula model with Weibull marginals provides a better fit to the data. Furthermore, we notice that this particular model is capable of detecting more differentially expressed genes than the previous model in Chapter 4 with gamma marginals.

The higher log-likelihood values of the model with Weibull marginals than the model with gamma marginals is good evidence to conclude that the Bayesian Gaussian copula incorporates a latent variable with Weibull marginals outperforming and better fit to the data.

CHAPTER 6

SUMMARY

Microarray technology is one of the modern technologies developed to identify differentially expressed from thousands of genes on a DNA molecule. There are two major microarray technologies available for the expression analysis: Spotted cDNA array and oligonucleotide array. This dissertation focuses on the statistical analysis of data from the spotted cDNA, also known as two-channel microarray. Numerous models have been proposed in the literature to identify differentially expressed genes from the red and green intensities measured by the two-channel microarray.

Motivated by the Bayesian models described in Newton et al. (2001) and Mav and Chaganty (2004), we propose two models for the joint distribution of the red and green intensities using a Gaussian copula, which accounts for the dependence. The differentially expressed genes were identified by calculating the Bayes estimates of the differential expression under the first proposed copula model with gamma marginals (in Chapter 3). The accuracy of the model parameter estimations is shown with two simulation studies with three different sample sizes. We applied the model to five microarray samples in *E. coli* data. The genes filtered as differentially expressed are matched with the genes have filtered with the model proposed by Mav and Chaganty (2004). The larger log-likelihood values under our model compare to the model of Mav and Chaganty (2004) suggest that our model has an improvement.

Then we proposed another Bayesian Gaussian copula model incorporated with a latent variable, which indicates differential expression. Here also we considered gamma marginals. The EM algorithm is applied to calculate the posterior probabilities of differential expression for the second model. The posterior probabilities rank the genes. Using simulation studies, we show that our Gaussian copula-based models are an improvement in identifying differential expression over the models given in Newton et al. (2001) and Mav and Chaganty (2004). To select the best model among our Gaussian models, we conducted an AIC study. The lower AICs in the Gaussian copula incorporate a latent variable, which suggests that it is a better fit for the data.

In Chapter 5, we presented our findings of the Gaussian copula incorporated with a latent variable with Weibull marginals. The ability of the Weibull distribution: fitting a variety of distributional shapes allows this certain Gaussian copula combined with a latent variable to have different forms of continuous data. Furthermore, we noticed that this model is capable of capturing a higher number of truly differentially expressed genes in *E. coli* data. In conclusion, the Gaussian copula incorporated with a latent variable with Weibull marginals provides a better fit and improves genes' identification.

REFERENCES

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., Lu, L., Lewis, D. B., and Tibshirani, R. (2000), "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." *Nature*, 403, 503–511.

Alwine, J. C., Kemp, D. J., Richmond, C. S., and Stark, G. R. (1977), "Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes." *Proc Natl Acad Sci U S A.*, 74, 5350–5354.

Baldi, P. and Hatfield, G. W. (2002), *DNA Microarrays and Gene Expression*, Cambridge: Cambridge University Press.

Bao, L., Zhu, Z., and Ye, J. (2009), "Modeling oncology gene pathways network with multiple genotypes and phenotypes via a copula method," in *2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 237–246.

Borchers, H. W. (2021), *pracma: Practical Numerical Math Functions*, r package version 2.3.3.

Broyden, C. G. (1970), "The convergence of a class of double-rank minimization algorithms," *Journal of the Institute of Mathematics and Its Applications*, 6, 76–90.

Chaba, L. A. (2006), "A Copula-based approach to differential gene expression analysis," Ph.D. thesis, Strathmore University, Nairobi, Kenya.

Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997), "Ratio-based decision and the quantitative analysis of cDNA microarray images." *Biomedical Optics*, 2, 364–374.

Cong, R. and Brady, M. (2012), "The interdependence between rainfall and temperature: Copula analyses," *The Scientific World Journal*, 2012.

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B*, 39, 1–38.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996), "Use of a cDNA microarray to analyse gene expression patterns in human cancer." *Nature Genetics*, 14, 457–460.

Di Lascio, F. M. L. (2008), “Analyzing the dependence structure of microarray data: a copula–based approach,” Ph.D. thesis, Università di Bologna, Bologna, Italy.

Efron, B., Tibshirani, R., and Tusher, V. G. (2001), “Empirical bayes analysis of a microarray experiment.” *Journal of the American Statistical Association*, 96, 1151–1160.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), “Cluster analysis and display of genome-wide expression patterns.” *Proc. Natl. Acad. Sci.*, 95, 14863–14868.

Engler, D. and Li, Y. (2009), “Survival analysis with high-dimensional covariates: an application in microarray studies.” *Stat Appl Genet Mol Biol.*, 8, 14.

Escarela, G. and Carriere, J. F. (2003), “Fitting competing risks with an assumed copula,” *Statistical Methods in Medical Research*, 12, 333–349.

Fletcher, R. (1970), “A new approach to variable metric algorithms,” *Computer Journal*, 13, 317–322.

Gilbert, P. and Varadhan, R. (2019), *numDeriv: Accurate Numerical Derivatives*, r package version 2016.8-1.1.

Goldfarb, D. (1970), “A family of variable metric updates derived by variational Means,” *Mathematics of Computation*, 24, 23–26.

Gui, J. and Li, H. (2005), “Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data.” *Bioinformatics*, 21, 3001–3008.

Joe, H. (1997), *Multivariate Models and Multivariate Dependence Concepts*, Chapman and Hall/CRC.

— (2015), *Dependence Modeling with Copulas*, Chapman and Hall/CRC.

Kasa, S. R., Bhattacharya, S., and Rajan, V. (2020), “Gaussian mixture copulas for high-dimensional clustering and dependency-based subtyping,” *Bioinformatics (Oxford, England)*, 12, 621—628.

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., and Mallick, B. K. (2003), “Gene selection: a Bayesian variable selection approach.” *Bioinformatics*, 19, 90–97.

Li, M., Boehnke, M., Abecasis, G. R., and Song, P. X. (2006), “Quantitative trait linkage analysis using Gaussian copulas,” *Genetics*, 173, 2317–2327.

Liu, H., Bebu, I., and Li, X. (2010), “Microarray probes and probe sets,” *Frontiers in Bioscience*, E2, 325–338.

Low, R. K. Y. (2018), “Vine copulas: modelling systemic risk and enhancing higher-moment portfolio optimisation,” *Accounting and Finance*, 58, 423–463.

Ma, S. and Huang, J. (2007), “Additive risk survival model with microarray data.” *BMC Bioinformatics*, 8, 192–201.

Ma, S., Huang, J., and Shen, S. (2009), “Identification of cancer-associated gene clusters and genes within clusters via clustering penalization.” *Stat Interface*, 2, 1–11.

Mav, D. and Chaganty, N. R. (2004), “Bivariate Models for Identifying Differentially Expressed Genes in Microarray Experiments,” *Journal of Statistical Theory and Applications*, 3, 111–124.

McLachlan, G. J. and Krishnan, T. (1997), *The EM algorithm and extensions*, Hoboken, NJ: John Wiley and Sons.

Mesbahzadeh, T., Miglietta, M. M., Mirakbari, M., Sardoo, S., and Abdolhoseini, M. (2019), “Joint modeling of precipitation and temperature Using Copula theory for current and future prediction under climate change scenarios in Arid Lands (Case Study, Kerman Province, Iran),” *Advances in Meteorology*, 2019.

Modiri, S., Belda, S., Heinkelmann, R., Hoseini, M., Ferrándiz, J. M., and Schuh, H. (2018), “Polar motion prediction using the combination of SSA and Copula-based analysis,” *Earth, Planets and Space*, 70.

— (2020), “A new hybrid method to improve the ultra-short-term prediction of LOD,” *Journal of Geodesy*, 94.

Narasimhan, B., Koller, M., Johnson, S. G., Hahn, T., Bouvier, A., Kiêu, K., and Gaure, S. (2021), *cubature: Adaptive Multivariate Integration over Hypercubes*, r package version 2.0.4.2.

Nash, J. C. (1979), *Compact numerical methods for computers: linear algebra and function minimisation*, Bristol and New York: Asam Hilger, 2nd ed.

Nelsen, R. B. (1996), *An Introduction to copulas*, New York: Springer.

— (2006), *An Introduction to copulas: 2nd ed.*, New York: Springer.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001), "On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data," *Journal of Computational Biology*, 8, 37–52.

Owzar, K., Jung, S. H., and Sen, P. K. (2007), "A copula approach for detecting prognostic genes associated with survival outcome in microarray studies," *Biometrics*, 63, 1089–1098.

Qian, D., Wang, B., Qing, X., Zhang, Y.and Wang, M., and Nakamura, M. (2017), "Drowsiness detection by bayesian-bopula discriminant classifier based on EEG signals during daytime short nap," *IEEE Transactions on Biomedical Engineering*, 64, 743–754.

Richmond, C. S., Glasner, J. D., Mau, R., Jin, H., and Blattner, F. (1999), "Genome-wide expression profiling in Escherichia coli K-12," *Nucleic Acids Research*, 27, 3821–3835.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C. F., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. (2000), "Systematic variation in gene expression patterns in human cancer cell lines." *Nature Genetics*, 24, 227–234.

Schena, M. (2003), *Microarray Analysis*, Hoboken, NJ: John Wiley and Sons.

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.

Schena, M., Shalon, D., Davis, R. W., Brown, P. O., and A., C. (1996), "Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes," *Proc. Natl. Acad. Sci. USA*, 93, 10614–10619.

Sebastiani, P., Gussoni, E., Kohane, I. S., and Ramoni, M. F. (2003), "Statistical challenges in functional genomics." *Statistical Science*, 18, 33–70.

Shanno, D. F. (1970), "Conditioning of quasi-Newton methods for function minimization," *Mathematics of Computation*, 24, 647–656.

Sklar, A. (1959), "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

Speed, T. P. (2003), *Statistical Analysis of Gene Expression Microarray Data*, Boca Raton, FL: Chapman and Hall (CRC Press).

Sreekumar, J.and Jose, K. K. (2008), "Statistical tests for identification of differentially expressed genes in cDNA microarray experiments." *Nature Genetics*, 7, 423–436.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999), "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." *Proc. Natl. Acad. Sci.*, 96, 2907–29126.

Tavazoie, S., Hughes, J., Campbell, M., Cho, R., and Church, G. (1999), "Systematic determination of genetic network architecture." *Nature Genetics*, 22, 281–285.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001), "Significance analysis of microarrays applied to the ionizing radiation response." *Proceedings of National Academy of Sciences*, 98, 5116–5121.

Valle, L., Leisen, F., and Rossini, L. (2017), "Bayesian non-parametric conditional copula estimation of twin data," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67.

Yang, S., Liu, T., and Hong, H. (2017), "Reliability of tower and tower-line systems under spatiotemporally varying wind or earthquake loads," *Journal of Structural Engineering*, 143.

Yuan, A., Chen, G., Zhou, G., and Rotimi, C. (2008), "Gene copy number analysis for family data using semiparametric copula model," *Bioinformatics and Biology Insights*, 2, 343–355.

Zhang, A., Fang, J., Calhoun, V. D., and Wang, Y. (2018), "High dimensional latent Gaussian copula model for mixed data in imaging genetics," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 105–109.

Zhang, Y., Beer, M., and Quek, S. T. (2015), "Long-term performance assessment and design of offshore structures," *Computers and Structures*, 154, 101–115.

APPENDIX A

SELECTED R CODES

In this section, we provide some of the important R codes we developed. Brief descriptions of all the important functions are stated below.

A.1 R CODES FOR CHAPTER 3

The following code is used to implement the marginal density of (R_{1j}, R_{2j}) which given in the equation (17). Here the function cop outputs the probability density function of the bivariate Gaussian copula in equation (14).

```
cop<-function(z1,z2,gam)
{
  z1_2=z1^2
  z2_2=z2^2
  gam_2=gam^2
  gam_2[1-gam2<1.e-16]=1-1.e-16
  gam_z=gam*z1*z2
  exp(-(gam_2*(z1_2+z2_2)-2*gam_z)/(2*(1-gam_2)))/sqrt(1-gam_2)
}

f_m<-function(parameter,data)
{
  r1=data[2]
  r2=data[3]
  alpha=parameter[1]
  alpha_0=parameter[2]
  v=parameter[3]
  gam=parameter[4]

  #constraints on parameters
  if(alpha > 0 && alpha_0 >0 && v>0 && gam>0 && gam<1 )
  {
```

```
f_t1<-function(t1) #function of theta1
{
  f_t2<-function(t2) #function of theta2
  {
    u1=pgamma(r1,alpha,rate=t1)  # Gamma CDF of R_1j
    u2=pgamma(r2,alpha,rate=t2)  # Gamma CDF of R_2j
    u1[1-u1<1.e-5]=1-1.e-5
    u2[1-u2<1.e-5]=1-1.e-5
    u1[u1<1.e-5]=1.e-5
    u2[u2<1.e-5]=1.e-5
    z1=qnorm(u1) #Standard Normal Inverse CDF of u1
    z2=qnorm(u2) #Standard Normal Inverse CDF of u2

    cop(z1,z2,gam)*dgamma(r1,alpha,rate=t1)*dgamma(r2,alpha,rate=
       t2)*dgamma(t1,alpha_0,rate=v)*dgamma(t2,alpha_0,rate=v)
  }
  int_t2 <- try(integrate(f_t2, lower=0, upper=Inf), silent =
     TRUE) #integrating w.r.t theta2
  if(inherits(int_t2 ,’try-error’))
  {
    warning(as.vector(int_t2))
    int_t2<- NA_real_
  }
  else
  {
    int_t2 <- int_t2$value
  }
  int_t2
}
f_t1 <- Vectorize(f_t1)
int_t1t2 <- try(integrate(f_t1, lower=0, upper=Inf), silent =
   TRUE) #integrating w.r.t theta1
if(inherits(int_t1t2 ,’try-error’))
{
  warning(as.vector(int_t1t2))
  int_t1t2<- NA_real_
}
else
```

```
    {
      int_t1t2<- int_t1t2$value
    }
  }
  else
  {
    int_t1t2=NA
  }
  int_t1t2
 }
}
```

Function E_RG evaluates the expected value of R_1R_2 in equation (23). Note that, adaptIntegrate is a built-in function in the R package cubature.

```
library(cubature)

E_RG<-function(parameter)
{
  alpha=parameter[1]
  alpha_0=parameter[2]
  v=parameter[3]
  gam=parameter[4]

  frt<-(t)
  {
    t1=t[1] # theta_1j
    t2=t[2] # theta_2j
    r1=t[3] # r_1j
    r2=t[4] # r_2j
    u1=pgamma(r1,alpha,rate=t1)  # Gamma CDF of R_1j
    u2=pgamma(r2,alpha,rate=t2)  # Gamma CDF of R_2j
    u1[1-u1<1.e-5]=1-1.e-5
    u2[1-u2<1.e-5]=1-1.e-5
    u1[u1<1.e-5]=1.e-5
    u2[u2<1.e-5]=1.e-5
    z1=qnorm(u1) #Standard Normal Inverse CDF of u1
    z2=qnorm(u2) #Standard Normal Inverse CDF of u2
    r1*r2*cop(z1,z2,gam)*dgamma(r1,alpha,rate=t1)* dgamma(r2,alpha,
       rate=t2)*dgamma(t1,alpha_0,rate=v)*dgamma(t2,alpha_0,rate=v)
  }
  e_rg=adaptIntegrate(frt, c(0,0,0,0), c(Inf,Inf,Inf,Inf))$integral
  e_rg
}
```

A.2 R CODES FOR CHAPTER 4

The following function calculates the marginal density of (R_{1j}, R_{2j}) for a gene j that is not differentially expressed (equation (26)). Similarly, the joint marginal density of intensities for a gene j that is differentially expressed given in equation (25) can be calculated with the function f_m which stated in A.1, after modifying for five parameters.

```
f_nde<-function(parameter,data)
{
  r1=data[2]
  r2=data[3]
  alpha=parameter[1]
  alpha_0=parameter[2]
  v=parameter[3]
  gam=parameter[4]
  p=parameter[5]

  #constraints on parameters
  if(alpha > 0 && alpha_0 >0 && v>0 && gam>0 && gam<1 && p>0 && p
     <100)
  {
    f_t<-function(t) #function of theta
    {
        u1=pgamma(r1,alpha,rate=t)  # Gamma CDF of R_1j
        u2=pgamma(r2,alpha,rate=t)  # Gamma CDF of R_2j
        u1[1-u1<1.e-5]=1-1.e-5
        u2[1-u2<1.e-5]=1-1.e-5
        u1[u1<1.e-5]=1.e-5
        u2[u2<1.e-5]=1.e-5
        z1=qnorm(u1) #Standard Normal Inverse CDF of u1
        z2=qnorm(u2) #Standard Normal Inverse CDF of u2

        cop(z1,z2,gam)*dgamma(r1,alpha,rate=t)*dgamma(r2,alpha,rate=t
           )*dgamma(t,alpha_0,rate=v)
    }
      int_t <- try(integrate(f_t, lower=0, upper=Inf), silent = TRUE)
          #integrating w.r.t theta
      if(inherits(int_t ,'try-error'))
      {
        warning(as.vector(int_t))
        int_t2<- NA_real_
      }
      else
      {
        int_t <- int_t$value
      }
  int_t
  }
}
```

We use the function logl to obtain the complete data loglikelihood written in equation (27). And also the expectation step of EM algorithm mentioned in Chapter 4 equation (28) is included in this function.

```
logl<-function(parameter,data)
{
  r1=data[2]
  r2=data[3]
  alpha=parameter[1]
  alpha_0=parameter[2]
  v=parameter[3]
  gam=parameter[4]
  p=parameter[5]

  fm=f_m(c(alpha,alpha_0,v,gam,p),data)
  fnde=f_nde(c(alpha,alpha_0,v,gam,p),data)
  if(p>0 && p<100)
  {
    w=(p*fm/100)/((p*fm/100)+((100-p)*f0/100)) #posterior probability
    llik=(w*(log(fm)+log(p)-log(100)))+((1-w)*(log(fnde)+log(100-p)-
       log(100))) #the complete data loglikelihood
  }
  else
  {
    llik=NA
  }
  llik
}
```

VITA

Prasansha Liyanaarachchi
Department of Mathematics and Statistics
Old Dominion University
Norfolk, VA 23529

Education

Ph.D. Old Dominion Universtiy, Norfok, VA. (Anticipated December 2021)
Major: Computational & Applied Mathematics (Statistics).
M.S. Sam Houston State University, Huntsville, TX. (May 2015)
Major: Statistics.
B.S. University Of Peradeniya, Sri Lanka. (September 2011)
Major: Statistics.

Experience

Statistical Analyst, Chesapeake Bay Program, Old Dominion University, Norfolk, VA, (07/2018-12/2021).
Graduate Teaching Assistant, Old Dominion University, Norfolk, VA, (08/2016 - 06/2018).
Asistant Lecturer, Wayamba University of Sri Lanka, Sri Lanka, (09/2015-07/2016).
Graduate Teaching Assistant, Sam Houston State University, Huntsville, TX, (08/2013-05/2015).
Asistant Lecturer, Wayamba University of Sri Lanka, Sri Lanka, (10/2012-06/2013).
Teaching Assistant, University of Peradeniya, Sri Lanka, (11/2011-08/2012).

Typeset using LaTeX.