BAYESIAN MODEL SELECTION FOR HIGH-DIMENSIONAL HIGH-THROUGHPUT DATA

A Dissertation

by

ADARSH JOSHI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2010

Major Subject: Statistics

BAYESIAN MODEL SELECTION FOR HIGH-DIMENSIONAL HIGH-THROUGHPUT DATA

A Dissertation

by

ADARSH JOSHI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Valen E. Johnson
	David B. Dahl
Committee Members,	Jianhua Hu
	Bradley M. Broom
	Ivan Ivanov
Head of Department,	Simon J. Sheather

May 2010

Major Subject: Statistics

ABSTRACT

Bayesian Model Selection for High-dimensional High-throughput Data.

(May 2010)

Adarsh Joshi, B.Tech., Indian Institute of Technology Bombay;

M.S., Texas A&M University

Co-Chairs of Advisory Committee: Dr. Valen E. Johnson
Dr. David B. Dahl

Bayesian methods are often criticized on the grounds of subjectivity. Furthermore, misspecified priors can have a deleterious effect on Bayesian inference. Noting that model selection is effectively a test of many hypotheses, Dr. Valen E. Johnson sought to eliminate the need of prior specification by computing Bayes' factors from frequentist test statistics. In his pioneering work that was published in the year 2005, Dr. Johnson proposed using so-called local priors for computing Bayes' factors from test statistics. Dr. Johnson and Dr. Jianhua Hu used Bayes' factors for model selection in a linear model setting. In an independent work, Dr. Johnson and another colleage, David Rossell, investigated two families of non-local priors for testing the regression parameter in a linear model setting. These non-local priors enable greater separation between the theories of null and alternative hypotheses.

In this dissertation, I extend model selection based on Bayes' factors and use non-local priors to define Bayes' factors based on test statistics. With these priors, I have been able to reduce the problem of prior specification to setting to just one scaling parameter. That scaling parameter can be easily set, for example, on the basis of frequentist operating characteristics of the corresponding Bayes' factors. Furthermore, the loss of information

by basing a Bayes' factors on a test statistic is minimal.

Along with Dr. Johnson and Dr. Hu, I used the Bayes' factors based on the *likelihood ratio statistic* to develop a method for clustering gene expression data. This method has performed well in both simulated examples and real datasets. An outline of that work is also included in this dissertation. Further, I extend the clustering model to a subclass of the decomposable graphical model class, which is more appropriate for genotype data sets, such as single-nucleotide polymorphism (SNP) data. Efficient `FORTRAN` programming has enabled me to apply the methodology to hundreds of nodes.

For problems that produce computationally harder probability landscapes, I propose a modification of the Markov chain Monte Carlo algorithm to extract information regarding the important network structures in the data. This modified algorithm performs well in inferring complex network structures. I use this method to develop a prediction model for disease based on SNP data. My method performs well in cross-validation studies.

Dedicated to my girlfriend Savita, my parents and my sister

ACKNOWLEDGEMENTS

I am indebted to many people for helping me attain my Ph.D. I feel very fortunate to have met people at different times in my life who have constantly encouraged me. I would like this opportunity to express my gratitude to every person who made it possible for me to achieve my goals.

First, I would like to thank my parents and my sister for giving me the support throughout my years of education. Their sacrifices for my sake have always inspired me to achieve more than my percieved capabilities. Next, I wish to express my thanks to my friends from my high school, undergraduate and graduate years, especially, Titu Das, Gunjan Agrawal, Alok Srivastava, Karan Advani and Shahriar Kibriya, for willing to help when needed.

I am greatly indebted to my advisor Dr. Valen Johnson and very thankful to have had an opportunity to work with him. He has been a stellar mentor and a friend. I would also like to thank Dr. Jianhua Hu for being my mentor at the University of Texas M. D. Anderson Cancer Center. Dr. Johnson and Dr. Hu have taught me to become an independent reseacher through their continual encouragement and positive criticism, while giving me the opportunity to seek my individual interests. I shall always be grateful to them for helping me shape my research career.

I would also like to thank my advisor at Texas A&M University, Dr. David Dahl and other members of my committee: Dr. Bradley Broom and Dr. Ivan Ivanov. I appreciate the time they devoted to answering my research questions, to reading my dissertation, and to attending my preliminary examination and dissertation defense.

Among other members of the faculty at the Department of Statistics at Texas A&M University, I would like to thank Dr. Daren Cline, Dr. Michael Longnecker, Dr. Faming Liang, Dr. Randy Eubank and Dr. Clifford Spiegelman. Each one of these individuals has

been instrumental in making my graduate school years a learning experience by giving me the encouragement on a personal level outside the classroom environment. I would also like to thank Soma Dhavala, a graduate student in the Department of Statistics for proof-reading my dissertation.

Lastly and most importantly, I want to express my warmest thanks to my girlfriend Savita for being by my side during the good times and the hard times. I would not have been able to come this far without her support and will always be grateful for her love and for all the personal sacrifices she made for me.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Markov Chain Monte Carlo (MCMC) based Bayesian model selection algorithms are becoming very popular for studying and comparing candidate models from an extremely large space. Bayesian model selection methods have emerged as a strong alternative against stepwise frequentist methods and have inherent advantages over their frequentist counterparts. Within the Bayesian paradigm, model selection is based on posterior probabilities of different models.

Research in this area was initiated by George and McCulloch (1993, 1997), who proposed a "Stochastic Search Variable Selection" procedure (SVSS) to determine promising subsets of predictor variables in the linear model setting. They used latent variables to identify subset choices according to posterior probabilities within the context of a hierarchical Bayesian mixture model. They used a Gibbs sampling algorithm to probe the multinomial posterior distribution of each latent indicator variable. By examining the results of the Gibbs output, they determined subsets of variables that had the highest posterior probabilities of appearing as predictors in the regression equation.

Madigan and Raftery (1994) combined model selection with Bayesian model averaging for prediction. In making predictions for future observations, Bayesian model averaging accounts for model uncertainty, which often represents a major component of prediction uncertainty (e.g., Leamer 1978; Hodges 1987; Raftery 1996; Moulton 1991; Draper 1995).

Subsequently, other researchers have extended Bayesian model selection ideas to more

---

The format and style follow that of *Biometrics*.

complicated model settings. Using latent variable methodology (e.g., Albert and Chib 1993), Lee et al. (2003) cast variable selection for probit regression models into a framework similar to that previously developed for linear models. Chen et al. (2004) proposed a conditional beta-binomial prior density for latent variables in an SVSS-type algorithm to explicitly model the presence of interaction terms. Zhang and Liu (2007) constructed Bayes' factors on groupings of single nucleotide polymorphism (SNP) data. They used a multinomial sampling model and a product Dirichlet prior model, and applied their method to datasets containing several hundred thousand SNP values to identify possible associations between subsets of SNPs.

In probability theory, statistics, and machine learning, a graphical model is a diagram that represents conditional independence relationships among random variables. The nodes of a graph are random variables that can be either continuous or categorical, and missing edges represent conditional independence relationships. The edges can have a direction, implying causal dependency between the "parents" and "children". The information about causal inference is not present in the data, so a researcher has to make modeling assumptions or impose prior information to investigate such relationships. When such relationships are not of interest, one may choose graphs with undirected edges. Graphical models with undirected edges are generally called Markov random fields or Markov networks. In this work, only model selection on undirected graphs is considered.

Graphical models are a useful tool in studying large interaction networks because they divide the set of nodes into "cliques", or maximal sets with no missing edges. One of the fundamental results pertaining to the theory of graphical models is the unpublished Hammersley-Clifford theorem that states that the joint probability distribution of the graph is determined completely after having determined the probability distributions of each clique and their shared nodes or edges. Besag (1974) offers an insightful discussion of the Hammersley-Clifford theorem along with an alternative proof.

Graphical model theory for categorical data was introduced by Darroch, Lauritzen, and Speed (1980). Around the same time, Goodman (1970, 1971) and Haberman (1974) studied a sub-class of log-linear models in which the hypothesis tests for interactions and the estimated cell counts could be computed directly; that is, without employing an iterative scheme, they developed a class of models which they called *decomposable models*. Darroch, Lauritzen, and Speed (1980) showed that graphical model class was a subset of the hierarchical log-linear model class and contains the class of decomposable models.

Madigan and York (1995) proposed a model selection algorithm for graphical model involving categorical random variables. Gaussian graphical models (GGMs; e.g., Lauritzen 1996; Jones and West 2005; Carvalho, Massam, and West 2007) assume a multivariate normal distribution on the nodes in a clique, and have become an increasingly popular class of models in the statistics and artificial intelligence community. GGMs have been used to detect interactions among gene expression levels (Wu, Yong, and Kalpathi 2003 and Dobra et al. 2004).

The model space in Zhang and Liu (2007) is also a subspace of the graphical model class. Verzilli, Stallard, and Whittaker (2006) have also used a subclass of the graphical model class for identifying the important SNPs amongst a hundred thousand of them. Along with other authors, I have devised a methodology for gene clustering (Hu, Joshi, and Johnson 2009) which is discussed in detail in this dissertation. These "cluster" models are also form a subclass of the graphical model class.

Bayesian methods are often criticized on the grounds of subjectivity. The subjectivity arises because of the need for prior specification. The greatest strength of Bayesian methods is their ability to incorporate prior information. However, the lack of prior information in many scenarios turns this greatest strength into the greatest weakness of Bayesian methods. Misspecified priors tend to affect Bayesian inference and Model Selection more than parameter estimation. For example, Zhang and Liu (2007) noted in their paper that

their estimated posterior probabilities depend critically on vaguely specified but very high-dimensional product Dirichlet prior model. Verzilli, Stallard, and Whittaker (2006) also used a similar prior. Bayesware Discoverer (`www.bayesware.com`) is a popular software program for graphical model selection which also uses a hyper-Dirichlet prior. The same is also true for almost all model selection algorithms including the ones that have been referenced in this introduction (except for Hu, Joshi, and Johnson 2009). However, little work has been done so far to address the problem of subjectivity in Bayesian analysis.

I also note that the use of prior information is not the only good reason for using Bayesian methods over frequentist methods. The Bayesian angle of visualizing the data generating mechanism has inherent advantages over the frequentist vision. In the context of model selection, Bayesian methods are better than stepwise methods because they compute the posterior probability for each possible model in a simultaneous fashion. Also, model averaging makes more sense than just concentrating on the one maximum likelihood estimate because we usually find that many models are able to explain the data to a comparable degree. However, to average over comparable models one needs to have a probability distribution over the space of models - something that's absent in the frequentist approach.

In our clustering method, we sought to eliminate the need of prior specification by computing Bayes' factors from frequentist test statistics (Hu, Joshi, and Johnson 2009). Noting that model selection is effectively a test of many hypotheses, we first obtained the likelihood ratio statistic (LRS) for such a test. Asymptotic theory completely specifies the distribution of the LRS under the null and alternative hypotheses. Under the null hypothesis, this distribution is free of any parameters, and under the alternative the distribution has only one non-centrality parameter. We specified a prior on that non-centrality parameter in a manner that produced good results in cross-validation.

For our clustering method, we used a Gaussian prior centered at 0 to specify the dis-

tribution of the non-centrality parameter of the LRS under the alternative hypothesis. Such priors are called local alternative priors and are standard in Bayesian hypothesis testing. Johnson and Rossell (2010) proposed two classes of alternative prior distributions called the *non-local alternative priors* which assign less weight to the region around the null value (which is 0 for the non-centrality parameter) and more weight in the regions away from the null value. Such priors therefore offer a convenient mechanism to separate the theory of the null hypothesis from the alternative hypothesis. By ignoring small deviations from the null hypothesis, they also offer a break from testing for statistical significance to a move towards testing for practical significance.

In this dissertation, I extend the clustering methodology (presented in Hu, Joshi, and Johnson 2009) by replacing local alternative priors with non-local alternative priors. Also, I replace the clustering model by a more general subclass of graphical models that allow the cliques to share nodes and edges. Allowing node and edge sharing is more appropriate in many applications, particularly in genetics problems (e.g., SNP data). Graphical models also allow very large clusters to be broken into component cliques that may share nodes and edges.

An important principle in modeling is known as the Occam's razor. According to the principle of Occam's razor, when competing hypotheses are equal in other respects, one should select the hypothesis that introduces the fewest assumptions and postulates the fewest entities while still sufficiently answering the question. Jefferys and Berger (1992) present a discussion of Occam's razor and Bayes' factors. This has significance in model selection algorithms which tend to select the simpler models due to an implicit Occam's razor inherent to the Bayes' factors. Hence, there's a higher chance of discovering the complex interaction structure in large-scale genetics data when employing the graphical model class than when using the clustering model class, because many clustering models are rejected under the latter's assumptions.

In this dissertation, only graphical models pertaining to categorical data are considered. However, these ideas can be easily extended to graphical models with quantitative nodes. Also, I restrict attention to decomposable graphs. I have done this primarily to make the methodology computationally feasible. Although the class of decomposable models is a subset of the class of graphical models, the decomposable model class still forms a rich class of models with many practical applications.

As noted above, Zhang and Liu (2007), Verzilli, Stallard, and Whittaker (2006) and our clustering method (Hu, Joshi, and Johnson 2009) also employ different subclasses of graphical models in their methods. However, since the models considered here are less restrictive than those models, a new MCMC methodology was developed. Madigan and York (1995) and Bayesware Discoverer employ a more general class of models than those considered in this dissertation. However, their MCMC method requires checking for admissibility of every candidate model in every step, thereby rendering their methods very computationally intensive. I also introduce a novel MCMC scheme to navigate through the proposed model space. In fact, many of the restrictions on the model space used by me were imposed with computational convenience in mind.

The organization of the rest of this dissertation is as follows. Chapter II discusses the theory of Bayesian hypothesis testing using test statistics. Chapter III discusses a methodology for clustering gene expression data as presented in Hu, Joshi, and Johnson (2009). Chapter IV covers a novel methodology for objective Bayesian graphical model selection. An application of this methodology to develop a predictive model for SNP-disease data is discussed in Chapter V. The derivation of Bayes' factors for model selection is given in Appendix A.

CHAPTER II

BAYESIAN HYPOTHESIS TESTING USING TEST STATISTICS

2.1 Introduction

Bayesian model selection can be thought of as a comparison amongst all the models in the model space. Compared to parameter estimation, hypothesis testing tends to be more sensitive to prior misspecification; a bad choice of priors will have a deleterious effect on model selection whereas, in typical estimation problems, the effect of prior misspecification wears off as the sample size increases. In fact, slight changes in the hyperparameter values can result in an entirely different set of models being selected.

Johnson (2005) and Johnson and Rossell (2008) sought to study the problem of hypothesis testing in the Bayesian paradigm from a more theoretical perspective. Taking an unconventional approach, he proposed to use test statistics to define Bayes' factors. Not only does this approach have a nice theoretical justification, it can easily be extended to models beyond the paradigm of linear models. This chapter reviews the recent work of Johnson and his co-authors (Hu and Johnson 2009; Hu, Joshi, and Johnson 2009) in the field of Bayesian hypothesis testing, who have used the LRS and *local alternative prior* to compute Bayes' factors. Johnson and Rossell (2010) have proposed two classes of *non-local alternative priors* for testing hypothesis concerning coefficients of a linear regression model. Later in the chapter, I develop the theory of computing Bayes' factors from LRS and *non-local alternative priors*. Section 2.5 discusses parameter setting for the various priors, while Section 2.6 compares the properties of the Bayes' factors obtained for the various types of priors. Chapter III outlines the methodology developed by Hu, Joshi, and Johnson (2009) for clustering gene expressions that uses *local alternative* priors for computing Bayes' factors.

2.2 Background

For testing hypotheses in the Bayesian paradigm, one needs to specify at least two different prior distributions corresponding to two hypotheses. If interest lies in testing a point null hypothesis versus a composite alternative, only the prior distribution under the alternative is required.

2.2.1 Vague priors in model selection

The primary impediment to the widespread application of objective Bayesian methodology to parametric hypothesis tests has been the requirement to specify proper prior distributions on model parameters. Bayes' factors cannot be defined with improper prior distributions, and the deleterious effects of vaguely specified prior distributions do not diminish even as sample sizes become large. For example, consider the test of hypotheses related to the mean parameter μ of a normal distribution with variance 1 based on n observations:

$$H_1 : \mu = 0 \tag{2.1}$$

$$vs.$$

$$H_2 : \mu \neq 0. \tag{2.2}$$

In a Bayesian setting, hypothesis H_2 must be redefined as a distribution on the parameter μ. Consider, for example:

$$H_2' : \mu \sim N(0, \sigma_\beta^2). \tag{2.3}$$

H_2' is indeed a re-statement of the "spike and slab" prior used by the SVSS algorithm of George and McCulloch (1993, 1997), except that the prior probabilities of individual hypotheses have not been assigned. Since interest lies in the Bayes' factor, the choice of those probabilities is immaterial to this discussion. Let $\phi(y^*, \mu^*, \sigma^{2*})$ denote the normal

density with mean μ^* and variance σ^{2*} evaluated at y^*. The Bayes' factor in favor of H_2 versus H_1 is:

$$BF(2|1) = \frac{\phi(\bar{\boldsymbol{Y}}, 0, \left(\sigma_\beta^2 + 1/n\right))}{\phi(\bar{\boldsymbol{Y}}, 0, 1/n)} = \frac{1}{(1 + n\sigma_\beta^2)^{\frac{1}{2}}} \exp\left(\frac{1}{2}\frac{n\sigma_\beta^2}{1 + n\sigma_\beta^2} Z_n\right). \tag{2.4}$$

Here, Z_n is the observed likelihood ratio of the test. It's evident from equation (2.4) that if for fixed n, $\sigma_\beta^2 \to \infty$, $BF(2|1) \to 0$ under both the null and alternative distributions. From a practical point of view this means that a vague prior distribution is not suitable for moderate sample sizes.

Suppose for instance that data is indeed generated from a $N(0.2, 1)$ distribution. Let $n = 200$. The median value of the distribution of the LRS is then close to 8.0. Setting a vague prior $\sigma_\beta^2 = 100$ when the observed value of the LRS is indeed 8.0, we get a Bayes' factor in favor of the alternative equal to 0.122. In most practical applications, we would like to be able to detect a mean of 0.2, however, in this case the vague prior is supporting the null hypothesis. In terms of model selection, this means that models assigned significant posterior probabilities will not perform well for prediction.

At the same time, letting σ_β^2 be too small can also create problems. If for fixed n we let $\sigma_\beta^2 \to 0$, then $BF(2|1) \to 1$ under both the null and alternative distributions. In terms of model selection, this means that prediction will suffer.

2.2.2 Zellner's g-prior

A popular choice for the alternative distribution in variable selection is Zellner's g-prior (Zellner 1986). A useful feature of this prior is that it assigns a variance covariance matrix to the predictors in a regression model which is proportional to $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, where $\boldsymbol{X}$ is the design matrix. As a result, the marginal density of the data can be obtained analytically, yielding a convenient expression for the Bayes' factor. In the one-dimensional case, the

g-prior corresponds to an alternative distribution of the following form:

$$H_2' : \mu \sim N(0, \frac{c}{n}). \tag{2.5}$$

Hence the expression for the corresponding Bayes' factor can be written as

$$BF(2|1) = \frac{\phi(\bar{\boldsymbol{Y}}, 0, (c+1)/n)}{\phi(\bar{\boldsymbol{Y}}, 0, 1/n)} = \frac{1}{(c+1)^{\frac{1}{2}}} \exp\left(\frac{1}{2}\frac{c}{c+1}Z_n\right). \tag{2.6}$$

In the multi-dimensional setting, if the data model is $\boldsymbol{Y} \sim N(\boldsymbol{X}\beta, I)$ and we test

$$H_1 : \boldsymbol{\beta} = \mathbf{0} \tag{2.7}$$

$$vs.$$

$$H_2' : \boldsymbol{\beta} \sim N(\mathbf{0}, c\,(\boldsymbol{X}'\boldsymbol{X})^{-1}), \tag{2.8}$$

the expression for Bayes' factor is:

$$BF(2|1) = \frac{1}{|\boldsymbol{I} + c\boldsymbol{P}_{\boldsymbol{X}}|^{\frac{1}{2}}} \exp\left(\frac{1}{2}Y'\{\boldsymbol{I} - (\boldsymbol{I} + c\boldsymbol{P}_{\boldsymbol{X}})^{-1}\}Y\right) \tag{2.9}$$

$$= \frac{1}{(c+1)^{\frac{d}{2}}} \exp\left(\frac{1}{2}\frac{c}{c+1}Z_n\right). \tag{2.10}$$

Here, $\boldsymbol{P}_{\boldsymbol{X}}$ is the projection matrix defined by the predictors in $\boldsymbol{X}$. The Bayes' factors in equations (2.6) and (2.10) is not consistent, since it does not go to 0 when the null hypothesis is indeed true.

2.2.3 Nuisance parameters

An additional complication encountered in Bayesian hypothesis testing occurs in the presence of nuisance parameters. Since nuisance parameters require additional modeling, their hyperparameters can have a significant effect on the Bayes' factors for the parameters of interest. Suppose, for example, that model for the data is $Y_i \sim N(\mu, \sigma^2)$, where σ^2 is unknown. If interest lies in testing hypotheses in equation (2.1) versus equation (2.3), the corresponding Bayes' factor is

$$BF(2|1) = \frac{\int \phi(\bar{\boldsymbol{Y}}, 0, \left(\sigma_\beta^2 + 1/n\right))\pi(\sigma^2)d\sigma^2}{\int \phi(\bar{\boldsymbol{Y}}, 0, 1/n)\pi(\sigma^2)d\sigma^2}. \tag{2.11}$$

Here, $\pi(\sigma^2)$ denotes the prior distribution on σ^2. From a practical point of view, the existence of nuisance parameters makes it harder for an investigator to find a good choice of prior parameters. Since, Bayesian models typically involve nuisance parameters, this problem is relevant to Bayesian analysis in general.

2.3 Bayes' factors based on test statistics: testing nested models

Johnson (2005), Johnson and Rossell (2008), and Hu and Johnson (2009) proposed to use test statistics to define Bayes' factors for nested models. This approach has the following advantages:

1. The theoretical framework for defining Bayes' factors is well established.
2. These Bayes' factors require just one hyperparameter which could be easily set to get the desired operational characteristics.
3. The ideas are directly extendable to non-linear models.

The theoretical set up for defining Bayes' factors based on test statistics is as follows. Following Davidson and Lever (1970), consider the test of hypothesis:

$$H_1 : \boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2\right), \tag{2.12}$$

where $\boldsymbol{\theta}_1^0$ is $d-$dimensional, against the sequence of local alternative hypotheses

$$H_{2,n} : \boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^n, \boldsymbol{\theta}_2^*\right). \tag{2.13}$$

where

$$\theta_i^n = \theta_i^0 + \delta_{i,n}/\sqrt{n}, \text{ with } \lim_{n\to\infty} \delta_{i,n} = \delta_i, \quad i = 1, \ldots, d, \tag{2.14}$$

and $\boldsymbol{\theta}_2^*$ is the vector of the true values of the nuisance parameter $\boldsymbol{\theta}_2$ under H_2.

Let $Z_n = -2\log(\lambda_n)$ be the LRS for the above test. Under the regularity conditions of Davidson and Lever (1970), the distribution of the LRS converges to central χ_d^2 when the null is true and a $\chi_d^2(\boldsymbol{\delta}'\bar{\boldsymbol{C}}_{11}\boldsymbol{\delta})$ distribution when the alternative is true. Here, $\bar{\boldsymbol{C}}_{11}$ denotes the upper $d \times d$ submatrix of the inverse of the information matrix $\boldsymbol{\Sigma}^{\boldsymbol{I}}$ based on a single observation, and $\boldsymbol{\delta} = \{\delta_i\}$. Also the symbol χ_ν^2 is used to represent a central chi-squared distribution with ν degrees of freedom and the symbol $\chi_\nu^2(\Delta)$ is used to represent a chi-squared distribution with ν degrees of freedom, and non-centrality parameter Δ. Specifying a prior distribution on $\boldsymbol{\delta}$ thus yields the marginal distribution of the LRS under the alternative hypothesis. The Bayes' factor is obtained as a ratio of the marginal distributions of the LRS under the two hypotheses.

As a further note, the prior on distribution $\boldsymbol{\delta}$ should generate values of δ_i that are $O_p(\sqrt{n}), \forall i = 1, \ldots, d$. Under certain regularity conditions on the design matrix, this ensures that the difference between values of θ drawn under $H_{2,n}$ and H_1 is $O_p(1)$. Only in that case, do we have a meaningful test of hypotheses.

2.3.1 Local alternative prior

Johnson (2005) and Hu and Johnson (2009) proposed using a *local alternative prior* to define the alternative distribution of $\boldsymbol{\delta}$. Specifically, they assumed that $\boldsymbol{\delta}$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $cn\bar{\boldsymbol{C}}_{11}^{-1}$. For this sequence of alternatives, the Bayes' factor in favor of the alternative model for a fixed value of n based on

Z_n is thus:

$$BF_{local}(2|1) = (cn+1)^{-d/2} \exp\left[\frac{cnZ_n}{2(cn+1)}\right]. \tag{2.15}$$

The proof of the result in equation (2.15) can be found in Johnson (2005). The Bayes' factor in equation (2.15) is similar to the term in equation (2.10), except that the scale parameter c is now replaced by cn. Unlike the latter, the Bayes' factor in equation (2.15) is consistent under a true null and a true alternative hypothesis. In the one-dimensional case, the Bayes' factor in equation (2.15) is the same as the Bayes' factor in equation (2.4). Even in the presence of nuisance parameters, the expression for the Bayes' factor computed from the LRS remains the same as in equation (2.15), so the only parameter c can be set easily based on, for example, frequentist operating characteristics.

2.4 Non-local alternative priors

Johnson and Rossell (2010) proposed two classes of alternative prior distributions called *non-local alternative prior distributions.* These priors assign less weight to the region around the null value and more weight to regions away from the null value, hence providing a convenient mechanism to separate the theory of the null hypothesis from the alternative hypothesis. By ignoring small deviations from the null hypothesis, they also offer a means of testing for practical signicance as opposed to statistical significance. Figure 1 shows the two *non-local* alternative prior densities plotted along with the *local* alternative prior density.

Johnson and Rossell (2010) used non-local priors in regression models by assigning non-local prior densities to regression coefficients. In Subsections 2.4.1 and 2.4.2 below,I provide an outline for using non-local prior densities to compute Bayes' factors based on test statistics. Detailed proofs of the expressions for Bayes' factors can be found in Appendix A of this dissertation. The proofs involve connecting the linear model to the LRS

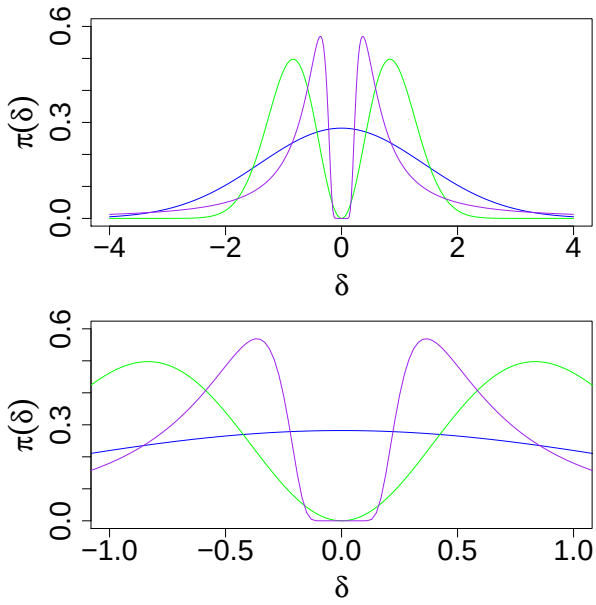

Figure 1: Alternative prior densities. A MoM density is depicted by the green curve, an iMoM density by the purple curve, and a local prior density by the blue curve.

through a vector of random variables which have a joint multivariate Normal distribution of dimension d. The test of hypothesis in equation (2.12) versus (2.13) then becomes equivalent to a test concerning the mean vector of that multivariate normal distribution.

2.4.1 Moment prior

The first class of non-local alternative priors proposed in Johnson and Rossell (2010) is called the Moment prior (MoM) density. Modeling $\boldsymbol{\delta}$ under the alternative hypothesis as a MoM density results in a prior density of the form:

$$\pi_M(\boldsymbol{\delta}) = \frac{1}{\prod_{i=0}^{k-1}(d+2i)} \left[\frac{\boldsymbol{\delta}'\bar{\boldsymbol{C}}_{11}\boldsymbol{\delta}}{n\tau}\right]^k \frac{\left|\bar{\boldsymbol{C}}_{11}\right|^{\frac{1}{2}}}{(2\pi n\tau)^{\frac{d}{2}}} \exp\left[-\frac{\boldsymbol{\delta}'\bar{\boldsymbol{C}}_{11}\boldsymbol{\delta}}{2n\tau}\right], \tag{2.16}$$

which leads to a Bayes' factor for testing hypothesis in equation (2.12) versus (2.13) that can be expressed as

$$BF_M(2|1) = \frac{\mu_k^*}{\prod_{i=0}^{k-1}(d+2i)} \frac{1}{(1+n\tau)^{k+\frac{d}{2}}} \exp\{\frac{1}{2}\frac{n\tau}{(n\tau+1)}Z_n\}. \tag{2.17}$$

2.4.2 Inverse Moment prior

The other class of non-local alternative priors proposed in Johnson and Rossell (2010) is called the Inverse Moment prior (iMoM) density class. Modeling $\boldsymbol{\delta}$ under the alternative hypothesis as a iMoM density results in a prior density expressible as:

$$\pi_I(\boldsymbol{\delta}) = c_I \left[\frac{\boldsymbol{\delta}'\bar{\boldsymbol{C}}_{11}\boldsymbol{\delta}}{n\tau}\right]^{-\frac{\nu+d}{2}} \exp\left[-\left(\frac{\boldsymbol{\delta}'\bar{\boldsymbol{C}}_{11}\boldsymbol{\delta}}{n\tau}\right)^{-k}\right], \tag{2.18}$$

where

$$c_I = \left|\frac{\bar{\boldsymbol{C}}_{11}}{n\tau}\right|^{1/2} \frac{k}{\Gamma(\nu/2k)} \frac{\Gamma(d/2)}{\pi^{d/2}}, \tag{2.19}$$

and the Bayes' factor for testing hypothesis in equation (2.12) versus (2.13) obtained under this alternative prior density is

$$BF_I(2|1) = \left(\frac{2}{n\tau}\right)^{d/2} \frac{k\Gamma(d/2)}{\Gamma\left(\frac{\nu}{2k}\right)} \frac{E_z\left[\left(\frac{n\tau}{z}\right)^{(\nu+d)/2} \exp\{-(n\tau/z)^k\}\right]}{\exp\{-\frac{1}{2}Z_n\}}, \tag{2.20}$$

where $z \sim \chi^2_d(Z_n)$, and $E_z(\cdot)$ represents expectation with respect to the density of z. This value must be computed numerically as a one-dimensional integral.

2.5 Setting prior parameters

2.5.1 Local prior

For the local alternative prior, values of c between 1 and 3 produce models which perform well in cross-validation studies. A method for clustering genes was developed by Hu, Joshi,

and Johnson (2009). For their method, they used values of c in the range of 1-3. The details of that method are discussed in Chapter III.

2.5.2 One dimensional non-local prior

For the non-local priors, the prior parameters obtained by Johnson and Rossell (2010) are for the linear model case. They can be extended to Bayes' factors based on the LRS using the same connection between the linear model and the LRS that was used to derive those Bayes' factors. Specifically, we can view $\frac{\bar{\boldsymbol{C}}_{11}^{\frac{1}{2}}\boldsymbol{\delta}}{\sqrt{n}}$ as a vector of standardized effects, where $\bar{\boldsymbol{C}}_{11}^{\frac{1}{2}}$ is the half-matrix of $\bar{\boldsymbol{C}}_{11}$.

Johnson and Rossell (2010) propose setting $k = 1$, and $\nu = 1$ for the iMoM prior. This leaves only one scaling parameter τ that must be specified. Johnson and Rossell (2010) propose two methods for setting this parameter. In the first, they determine τ by fixing the mode of the prior density at a value deemed most likely under the alternative hypothesis. Let $w = \frac{\boldsymbol{\delta}'\bar{\boldsymbol{C}}_{11}\boldsymbol{\delta}}{n}$. Then the MoM and iMoM prior modes occur at the locus of points for which $w = 2k\tau$ and $w = \tau(\frac{2k}{\nu+d})^{\frac{1}{k}}$, respectively.

Alternatively, τ can be determined so that high prior probability is assigned to the region of high practical significance, and low probability is assigned to regions of less practical significance. For instance, standardized effect sizes of less than 0.2 are often not considered substantively important in the social sciences (e.g., Cohen 1992). Small standardized effects in model selection correspond to those features in data which are not important in prediction and should be excluded by the rule of the Occam's razor.

In the linear model case, Johnson and Rossell (2010) suggested that τ may be determined so that the prior probability assigned to the event that a standardized effect size is less than 0.2, is say, less than 5%. When $k = 1$, and $d = 1$ the probability assigned to the interval (a, a) by MoM prior centered on 0 with scale τ and $n = 1$ is

$$2\left[\Phi\left(\frac{a}{\sqrt{\tau}}\right) - \frac{a}{\sqrt{2\pi\tau}}\exp\left(-\frac{a^2}{2\tau}\right) - \frac{1}{2}\right]. \tag{2.21}$$

For an iMoM prior, the corresponding probability is

$$1 - G\left[(a/\sqrt{\tau})^{-2k}; \frac{\nu}{2k}, 1\right], \tag{2.22}$$

where $G(\cdot; s_1, s_2)$ denotes a gamma distribution function with shape s_1 and scale s_2. For $\tau = 0.114$ and 0.348, the probabilities assigned by the MoM prior to the interval (0.2, 0.2) are 0.05 and 0.01, respectively. The corresponding values of τ for the iMoM prior are 0.077 and 0.133 respectively.

2.5.3 Multi-dimensional non-local prior

In this subsection I demonstrate how the one-dimensional parameter recommendations can be used to obtain parameter settings for multi-dimensional non-local alternative prior densities. When $d > 1$, the LRS has a χ^2_d when the null is true and a $\chi^2_d(\boldsymbol{\delta}'\bar{\boldsymbol{C}}_{11}\boldsymbol{\delta})$ distribution when the alternative is true. Since χ^2_d and $\chi^2_d(\boldsymbol{\delta}'\bar{\boldsymbol{C}}_{11}\boldsymbol{\delta})$ are obtained as sums of d independent χ^2_1 's (central and non-central respectively), the problem of parameter tuning of the d-dimensional non-local alternative prior distribution could be interpreted as either of the following:

1. The hypersphere in the d-dimensional space that is centered around the null value and has a radius of $0.2 * \sqrt{d}$ has a low probability, say 0.05 or 0.01.

2. The mode of the d-dimensional distribution lies on a hyper-sphere whose radius is proportional to $\sqrt{d}$.

It turns out that both approaches are easy to implement for the MoM prior, by choosing $\tau = \tau_1 * d$, where τ_1 = 0.114 or 0.348. However, the iMoM prior is more difficult. If the first approach is taken, then τ is equal to $\tau_1 * d$. However, doing this keeps the mode at roughly the same distance from the null as in the one-dimensional prior. If we instead take the second approach then the probability in the hypershphere in point 1 above will go to 0.

This is not surprising because increasing d will cause the prior to become flatter near the null value and increase sharply near the mode. Taking the view that it is more important for the mode to move away from the null distribution than assigning a non-diminishing probability to the region around the null distribution, τ should be set as equal to $\tau_1 * d * \frac{d+1}{2}$, where τ_1 = 0.077 or 0.133. The practical implication of the foregoing discussion is that only one parameter is required for all Likelihood Ratio tests on all degrees of freedom.

2.6 Comparison of Bayes' factors

To compare the performance of the Bayes' factors based on the local and non-local alternative prior densities, I simulated the values from the distribution of the LRS under true *null* and true *alternative* hypotheses in a test of the mean vector of a multivariate normal distribution with a covariance matrix equal to the identity matrix of size d. Formally speaking, I assumed that $Y \sim N(\boldsymbol{\mu}, \boldsymbol{I}_d)$. I am interested in the distributions of the LRS involving the test of point hypotheses:

$$H_0 : \boldsymbol{\mu} = \mathbf{0} \tag{2.23}$$

$$vs.$$

$$H_1 : \boldsymbol{\mu} = \delta^* \underline{\mathbf{1}}. \tag{2.24}$$

Here $\underline{\mathbf{1}}$ represents a vector of 1's. Let n denote the sample size. Then the distributions of the LRS are under H_0 and H_1 are, respectively, χ^2_d and $\chi^2_d((\sqrt{n}\delta^*)^2)$. To get the distributions of the Bayes' factors, I first simulated the LRS and then use equations (2.15, 2.17, 2.20) to obtain the corresponding Bayes' factors. I set $d = 1$ and $d = 9$. For the alternative hypotheses, I set $\delta^* = 0.4$ for the case $d = 1$, and $\delta^* = 0.3$ for the case $d = 9$.

The empirical distributions of the Bayes' factors thus obtained are plotted in Figure 2. Compared to the Bayes' factors based on the *local* alternative prior density, the Bayes'

factors based on the *non-local* alternative prior densities tend to be much smaller when the data is generated from true *null* hypothesis. However, when the data is generated from true *alternative* hypotheses, Bayes' factors based on the local and non-local alternative prior densities are comparable in value.

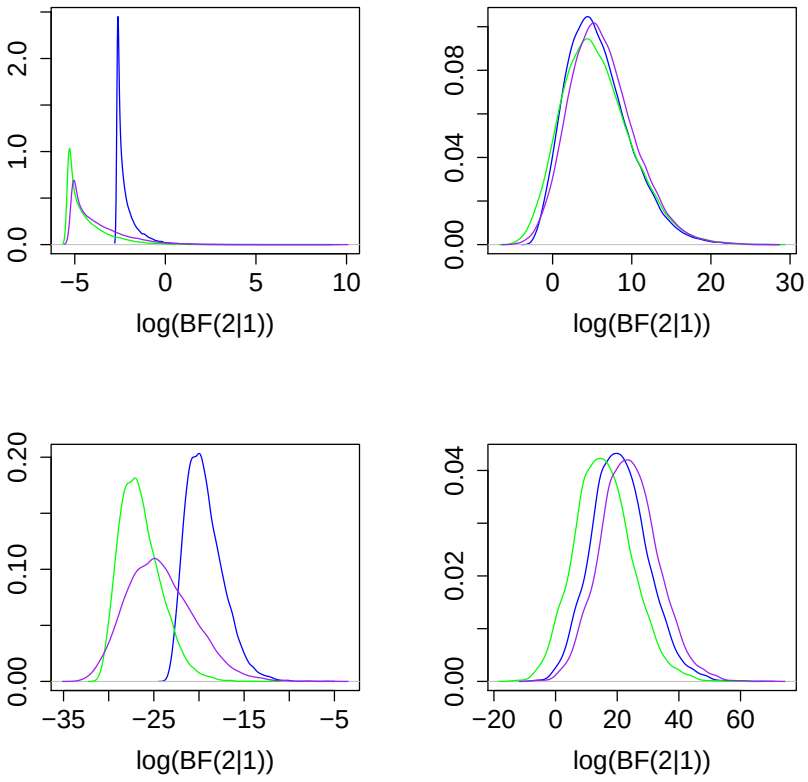

Figure 2: Distribution of Bayes' factors. These figures show the empirical densities of Bayes' factors based on test statistics for the local (blue), MoM(green) and iMoM(purple) alternative prior densities. The left panels correspond to scenarios when the data were generated from the true *null* distribution, and the right panels correspond to scenarios when the data were generated from the true *alternative* distributions. The top panels correspond to the scenarios when the test of the null versus the alternative involves only 1 degree of freedom and the bottom panels correspond to the scenarios when the test of the null versus the alternative involved 9 degrees of freedom. All the figures correspond to a sample size of 100.

CHAPTER III

CLUSTERING GENE EXPRESSIONS

3.1 Introduction

Using the Bayes' factors based on the local alternative prior in equation (2.15), we developed a method for clustering genes based on their co-expression values. We were able to successfully apply the method to data on breast cancer patients and controls and obtained gene clusters whose existence is corroborated by scientific literature. This section reviews the methodology presented in Hu, Joshi, and Johnson (2009).

Our methodology consists of four innovations. The first, of course, was the use of Bayes' factors based on the LRS in model selection. Secondly, observed data structures were converted to contingency tables by discretizing the raw gene expression data. With this reduction, we were able to use classical log-linear models to explore interactions between arbitrary subsets of expression profiles. Unlike other methods for discretizing gene expression data (e.g., Potamias, Koumakis, and Moustakis 2004), we discretize based on within-subject rankings of expression levels. The obtained ranked values are invariant to monotonic transformations applied to expression values at the subject level, hence this method of discretization eliminates many of the normalization artifacts often associated with gene expression experiments. In particular, the expression levels measured on a single microarray produce ranked values that are invariant to monotonic transformations applied to expression levels measured from that chip.

The third innovation of our methodology was the specification of a simple prior density on the space of cluster configurations - the Ewens' sampling distribution. Analytic expressions are readily available for prior expectations on the number of clusters for this density, and density requires only one parameter to be set. Such expressions provide a

convenient way to set that single parameter.

We restricted our attention to hierarchical log-linear models. Finally, we constrained the variables in the model to be included in at most one cluster. This restriction implies that a variable A cannot be assumed to interact with both variables B and C unless there exists a multi-way interaction between the three variables. This assumption greatly improves the computational efficiency of our algorithm by eliminating the need to implement iterative estimation procedures to compute the LRS, hence allowing the method to be applied to high-throughput genomic data comprising of potentially thousands of genes.

3.2 Discretization of gene expression data

Briefly, the discretization method can be explained in the steps below:

1. For each subject, rank the expression scores for all genes for that subject.

2. For each gene, discretize the expression levels based on the ranks obtained in the previous step.

We converted each gene expression to a categorical random variable with only two categories; however a discretization into more than 2 categories can be achieved in a similar way. Table 1 shows the discretization process for a hypothetical example, each subject corresponds to a specific row of the table and each gene corresponds to a specific column. A disadvantage of our method is that the discretization of continuous variables will result in the loss of information. However, the alternative strategy will be to model the shape of the interactions, which is unnecessary and the additional parameters required to model those interactions will affect the inference on the interaction structures. We have compared our method to methods for continuous gene expression data and have obtained similar results. This suggest that the loss of information associated with the above discretization is small compared to the gains through model simplification.

Table 1: Discretization of gene expression data. This table shows the discretization method applied to gene expression data. The left most column contains the binary phenotype information for each subject. The columns marked $G_1 - G_4$ show the hypothetical gene expression data (columns 2-5), within column ranks of gene expressions (columns 6-9) and the discretized scores (columns 10-13) for 4 genes.

	Raw Expression data				*Within-subject ranks*				*Discretization within columns*			
Cancer	G_1	G_2	G_3	G_4	G_1	G_2	G_3	G_4	G_1	G_2	G_3	G_4
0	3	6	9	0	3 2	6 3	9 4	0 1	3 2 0	6 3 1	9 4 1	0 1 0
0	2	9	8	1	2 2	9 4	8 3	1 1	2 2 0	9 4 1	8 3 1	1 1 0
1	2	1	0	3	2 3	1 2	0 1	3 4	2 3 1	1 2 0	0 1 0	3 4 1
1	6	1	4	3	6 4	1 1	4 2	5 3	6 4 1	1 1 0	4 2 0	5 3 1

3.3 Model

We model contingency tables of discretized gene expression data as a log-linear model. To explain the model and the constraints, first consider a saturated log-linear model comprising of 3 variables. Following the notation of Bishop, Fienberg, and Holland (1975), such a model can be expressed as:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}, \qquad (3.1)$$

where m_{ijk} is the expected cell count for level i of variable 1, level j of variable 2, and level k of variable 3. The parameters of the model are the intercept u; main effects of variables 1, 2, and 3, $u_{1(i)}$, $u_{2(j)}$, $u_{3(k)}$, respectively; two-way interaction terms $u_{12(ij)}$, $u_{23(jk)}$, $u_{13(ik)}$; and the three-way interaction term $u_{123(ijk)}$. Each of the seven subscripted u terms sum to 0 over each lettered subscript (i, j, or k). So if the number of levels of each variable are n_1, n_2, and n_3 respectively, $u_{1(i)}$ consists of $(n_1 - 1)$ independent parameters (degrees of freedom), $u_{12(ij)}$ consists of $(n_1 - 1) * (n_2 - 1)$ independent parameters, and so on. The total number of independent parameters in the model equals the number of elementary cells.

The hierarchical model constraint implies that a lower order term will be 0 only if all the higher order terms containing the lower order term are also 0. With reference to the

model in equation (3.1), this means that if $u_{12(ij)} = 0$, then $u_{123(ijk)} = 0$. The constraint that variables in the model be included in at most one cluster imply that we cannot have a model which includes the interaction term $u_{12(ij)}$ and $u_{13(ik)}$, unless the model also includes $u_{23(jk)}$. So an unsaturated model like

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)}, \tag{3.2}$$

is inadmissible in the model space, even though it is a hierarchical model. However, a model like

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)}, \tag{3.3}$$

is admissible. The model constraints provide two nice properties which enable efficient sampling algorithms:

1. Interactions within a cluster can be estimated and tested using only the marginal table of that cluster.

2. Each of the models in the model space are *directly estimable*, that is, the cell estimates can be obtained without employing an iterative scheme.

3.4 Ewens' prior

We imposed a prior based on the Ewens' sampling distribution (Ewens 1972) on the model space. Tavare and Ewens (1997) provide analytic expressions for the expected number of singleton clusters a and the total number of clusters K as a function of number of variables p and a parameter θ as below:

$$E(a_1) = \frac{p\theta}{p + \theta - 1}, \text{and,} \tag{3.4}$$

$$E(K) = \sum_{i=0}^{p-1} \frac{\theta}{\theta + i}. \tag{3.5}$$

From the above equations, we selected values of θ to constrain the number of nonsingleton clusters expected a priori or, equivalently, to specify the prior expected proportion of singleton clusters. For our analyses, we set θ so that $r = E(a_1)/E(K) \approx 0.8$.

3.5 Sampling scheme

Our interest was in obtaining an "average" model for gene clustering. A complete enumeration of posterior probabilities of the models in the model space is not feasible even for a moderate number of variables p. We therefore resorted to a Markov chain Monte Carlo algorithm as described in the steps below.

1. Initialize the chain to a state in which each variable forms its own cluster.

2. (a) With probability P_s, select a cluster c random in the current model, say M_k, to potentially split.

   (b) With probability $1 - P_s$, select two clusters c_1 and c_2 randomly to potentially merge.

   Let M' denote the proposed model state.

3. If the cluster chosen to split contains only one variable, or if the two clusters considered for merge contain more than the allowed number of variables, record the current state and return to Step 2.

4. When a split is proposed, define the proposal density $Q(M'|M_k)$ to be

$$Q(M'|M_k) = \frac{P_s}{|M_k|2^{1-|c|}}, \tag{3.6}$$

   and the (reverse) proposal density to be

$$Q(M_k|M') = \frac{1 - P_s}{\binom{|M_k|+1}{2}}. \tag{3.7}$$

These proposal densities correspond to an equiprobable random split of a cluster c containing $|c|$ variables in model configuration M_k, which contains $|M_k|$ clusters, into one of $2^{|c|-1}$ configurations of two subclusters. When a merge of clusters c_1 and c_2 is proposed, the proposal densities are

$$Q(M'|M_k) = \frac{1 - P_s}{\binom{|M_k|}{2}}, \tag{3.8}$$

and

$$Q(M_k|M') = \frac{P_s}{(|M_k| + 1)2^{1-|c_1 \cup c_2|}}. \tag{3.9}$$

5. Compute the Bayes' factor BF_{local} based on the LRS between the nested log-linear models defined by the current and candidate states according to equation (2.15).

6. Let the Ewens' prior probabilities for the current and candidate states be denoted by π_k and π', respectively. The Metropolis-Hastings ratio for transition to the candidate configuration is obtained as

$$\min\left(1, BF \times \frac{Q(M_k|M')}{Q(M'|M_k)} \times \frac{\pi'}{\pi^t}\right), \tag{3.10}$$

if a merge is proposed, or,

$$\min\left(1, \frac{1}{BF} \times \frac{Q(M_k|M')}{Q(M'|M_k)} \times \frac{\pi'}{\pi^t}\right). \tag{3.11}$$

if a split is proposed.

7. Record the current cluster configuration and return to Step 2 until a sufficient number of cluster configurations have been sampled.

3.6 Simulated gene expression data example

To test the algorithm described above, we designed a simulation study in which a few strongly interacting variables were seeded into a contingency table containing a total of

Table 2: True clusters in the simulated gene expression data

Cluster (5, 6)

	$x_6 = 0$	$x_6 = 1$
$x_5 = 0$	50	10
1	10	50

Cluster (7, 8, 9)

	$x_9 = 0$		$x_9 = 1$	
	$x_8 = 0$	$x_8 = 1$	$x_8 = 0$	$x_8 = 1$
$x_7 = 0$	54	2	2	2
1	2	2	2	54

Cluster (1, 2, 3, 4)

	$x_3 = 0, x_4 = 0$		$x_3 = 1, x_4 = 0$	
	$x_2 = 0$	$x_2 = 1$	$x_2 = 0$	$x_2 = 1$
$x_1 = 0$	43	1	1	6
1	1	1	6	

	$x_3 = 0, x_4 = 1$		$x_3 = 1, x_4 = 1$	
	$x_2 = 0$	$x_2 = 1$	$x_2 = 0$	$x_2 = 1$
$x_1 = 0$	6	1	1	1
1	6	1	43	

1000 binary variables with 120 joint observations on each. Interacting variables were divided into three clusters of different sizes: $\{$(1, 2, 3, 4), (5,6), (7, 8, 9)$\}$. The marginal tables corresponding to the true clusters are in Table 2. No structure was imposed on the remaining variables, which were generated randomly. The parameter θ of the Ewens' prior density was set to 1,700 to limit the expected proportion of singleton clusters to 0.8. Ten million iterations of the MCMC algorithm were performed to obtain the estimated posterior probabilities of the clusters, as depicted in Table 3. The three true clusters were obtained with high posterior probabilities, although three spurious triples are assigned higher posterior probability than the true cluster (5,6).

Table 3: Clustering simulated gene expression data. This table presents the results of the gene clustering method applied to simulated gene expression data with 1000 "genes". The "true gene" clusters are: $\{$(1, 2, 3, 4), (5,6), (7, 8, 9)$\}$. The rightmost column shows the posterior probability estimates for analysis with a local prior ($c = 2$).

1000 Genes				
7	8	9		0.982
1	2	3	4	0.975
465	737	871		0.955
587	743	809		0.920
500	503	877		0.882
5	6			0.757
35	347	859		0.725
230	885			0.646
49	89			0.629
189	380			0.625

3.7 Analysis of breast cancer data

Our purpose in developing a statistical model for cluster configurations is to facilitate the analysis of gene expression data. We therefore applied our method to the data presented in West et al. (2001). Those data set consist of gene expression measurements taken on 7,129 genes sampled from 49 patients. Twenty-five of these patients were diagnosed as estrogen receptor positive (ER+), whereas the remaining 24 were diagnosed as estrogen receptor negative (ER-). The raw data were pre-processed using the robust multi-array analysis (RMA) procedure proposed by Irizarry et al. (2003).

There are only a few previous studies of gene-gene interactions among these genes, and one motivation for the current analysis is the exploration of interactions between previously implicated biomarkers of breast cancer. We therefore restricted our study to 71 genes in which previous studies have indicated an association with some form of breast cancer (e.g., Cooper 2001, Spurdle et al. 2007).

To proceed, we discretized the gene expression data using the method described above. Since only 49 samples were available for this study, we restricted our investigation to four-

way interactions. Marginal tables associated with four-way interactions thus had expected cell counts of approximately 3 under the independence model; such models are on the borderline in terms of the chi-squared approximation to the LRS.

For our analysis, the Ewens' prior parameter was set to $\theta = 119$. Thus, the prior expectation of the proportion of non-singleton clusters was $r \approx 0.2$. Results from this analysis are displayed in Table 4, where the 10 clusters estimated to have the highest posterior probabilities are displayed. The MCMC algorithm described above was run to obtain 10 million samples from the posterior distribution on the clusters.

Table 4: Breast cancer data results with 71 genes

Local ($c = 1$)		Local ($c = 2$)		Local ($c = 3$)	
BRCA1,JNK	0.597	BRCA1,JNK	0.694	BRCA1,JNK	0.763
HSF1,P53a	0.575	HSF1,P53a	0.635	HSF1,P53a	0.645
GSTP1b,AKT2	0.432	ESR1,GATA3	0.532	KRAS,CYR61,ATF	0.479
ESR1,GATA3	0.422	GSTP1b,AKT2	0.500	AR,FOXA1	0.476
CDKN2B,TSG101	0.304	AR,FOXA1	0.393	GSTP1b,AKT2	0.474
AR,FOXA1	0.287	GSTP1a,AR,FOXA1	0.389	ESR1,GATA3	0.457
ESRRA,ERBB2	0.283	KRAS,CYR61,ATF	0.345	GSTP1a,ESR1,GATA3	0.438
P53b,WNT10b	0.283	P53b,WNT10b	0.332	P53b,WNT10b	0.385
ERK2,NRAS	0.264	CDKN2B,TSG101	0.323	GSTP1a,AR,FOXA1	0.350
CYR61,ATF	0.263	ERK2,NRAS	0.314	ERK2,NRAS	0.348

In this table, we see that the two-gene clusters (BRCA1,JNK) and (HSF1, P53a) were detected with high posterior probabilities. In addition, cluster (ESR1, GATA3) exhibits a high posterior probability for c = 1, 2. For c = 3, both (ESR1, GATA3), and the cluster (GSTP1a, ESR1, GATA3) are assigned fairly high posterior probabilities. For c = 1, 2, the three-gene cluster (GSTP1a, AR, FOXA1) and the subcluster (AR, FOXA1) are detected amongst the best clusters. Taken together, these observations indicate the existence of a more complex interaction structure between the five genes GSTP1a, ESR1, AR, GATA3, and FOXA1.

Marginal three-way contingency tables for the clusters (GSTP1a, ESR1, GATA3) and

Table 5: Clusters in the breast cancer expression data

Cluster (GSTP1a, ESR1, GATA3)					
	$x_{GSTP1a}=0$			$x_{GSTP1a}=1$	
	$x_{GATA3}=0$	1		$x_{GATA3}=0$	1
$x_{ESR1}=0$	4	1	$x_{ESR1}=0$	20	0
1	1	19	1	0	4
Cluster (GSTP1a, AR, FOXA1)					
	$x_{GSTP1a}=0$			$x_{GSTP1a}=1$	
	$x_{FOXA1}=0$	1		$x_{FOXA1}=0$	1
$x_{AR}=0$	4	2	$x_{AR}=0$	19	0
1	1	18	1	1	4

(GSTP1a, AR, FOXA1) are shown in Table 5. In these tables, it appears that the two cluster configurations are similar. Hence, the two three-gene clusters compete for gene GSTP1a, which under our model is restricted to participate in at most one cluster.

Interactions of genes in the cluster (ESR1, GATA3, FOXA1) is well known in the bioinformatics literature. There is an extensive literature (e.g., Bertucci et al. 2000, Lacroix and Leclercq 2004) confirming the strong interaction among these three genes. To examine the posterior probability that genes ESR1, GATA3, and FOXA1 form an interaction cluster, we first removed the genes GSTP1a and AR from our dataset. We then reran the MCMC algorithm on the remaining 69 genes, now taking $\theta = 116$, (again making the prior expectation of the proportion of non-singleton clusters equal to about 0.2. Posterior estimates of cluster presence are displayed in Table 6.

As expected, the cluster (ESR1, GATA3, FOXA1) appears with high posterior probability for all settings of the prior. The three-way contingency table displayed in Table 7 indicates that both genes ESR1 and GATA3 coexpress at their low expression levels when gene FOXA1 is downregulated, and coexpress at their high expression levels when gene FOXA1 is upregulated.

Table 6: Breast cancer data with 69 genes

Local ($c = 1$)		Local ($c = 2$)		Local ($c = 3$)	
ESR1,GATA3,FOXA1	0.742	ESR1,GATA3,FOXA1	0.899	ESR1,GATA3,FOXA1	0.911
BRCA1,JNK	0.559	BRCA1,JNK	0.679	BRCA1,JNK	0.753
HSF1,P53a	0.550	HSF1,P53a	0.662	HSF1,P53a	0.673
GSTP1b,AKT2	0.468	GSTP1b,AKT2	0.493	KRAS,CYR61,ATF	0.517
CDKN2B,TSG101	0.338	KRAS,CYR61,ATF	0.389	GSTP1b,AKT2	0.460
ERK2,NRAS	0.318	P53b,WNT10b	0.356	P53b,WNT10b	0.398
ESRRA,ERBB2	0.301	ESRRA,ERBB2	0.328	ESRRA,ERBB2	0.359
P53b,WNT10b	0.282	ERK2,NRAS	0.323	ERK2,NRAS	0.341
CYR61,ATF	0.281	CDKN2B,TSG101	0.303	CDKN2B,TSG101	0.277
IFI27,TNFa	0.235	CYR61,ATF	0.265	IFI27,TNFa	0.268

Table 7: More clusters in the breast cancer expression data

Cluster (ESR1, GATA3, FOXA1)					
$x_{ESR1} = 0$			$x_{ESR1} = 1$		
	$x_{FOXA1} = 0$	1		$x_{FOXA1} = 0$	1
$x_{GATA3} = 0$	20	1	$x_{GATA3} = 0$	4	0
1	1	3	1	0	20
Cluster (KRAS, CYR61, ATF)					
$x_{KRAS} = 0$			$x_{KRAS} = 1$		
	$x_{ATF} = 0$	1		$x_{ATF} = 0$	1
$x_{CYR61} = 0$	17	0	$x_{CYR61} = 0$	4	4
1	0	8	1	4	12

Our model also detected the three-gene cluster (KRAS, CYR61, ATF) with relatively high posterior probability. The interaction pattern between these genes can be better understood by an examination of the corresponding marginal table as displayed in Table 7. This table suggests that the association between genes CYR61 and ATF is strong when gene KRAS is not highly expressed, and diminishes as the expression level of gene KRAS increases. This kind of three-way interaction structure cannot be represented by sub-models containing only lower order interaction terms.

To determine if the gene clusters identified by our method have any biological signifi-

cance, we used the PubMed search engine (`www.ncbi.nlm.nih.gov/sites/entrez`) to identify functions associated with identified gene clusters. As it happens, most of the interactions detected by our method are well supported in the bioinformatics literature.

BRCA1 is a gene whose association with breast cancer is well-known. Harkin et al. (1999) pointed out that expression of JNK is induced by activation of transcription factor BRCA1. The strong association between HSF1 and P53a expressions has been described in Quenneville et al. (2002).

As discussed above, the results in Table 4 indicate the existence of a more complex interaction structure between the five genes GSTP1a, ESR1, AR, GATA3, and FOXA1. We were indeed able to find literature supporting the three-gene sub-cluster (GSTP1a, ESR1, AR). A significant interaction between ESR1 and AR is reported in Panet-Raymond et al. (2000). Both the genes GSTP1a and ESR1 participate in a biological pathway involved in carcinogen/estrogen metabolism. A more thorough search through the Gene Network Database (`http://humgen.med.uu.nl/lude/genenetwork`) suggests that the two genes are indirectly correlated through ESR1 $\rightleftarrows$ MNAT1 $\rightleftarrows$ PSMC4 $\rightleftarrows$ GSTP1.

The three-gene cluster (KRAS, CYR61, ATF) detected by our method is also interesting. Chien et al. (2004) pointed out that low levels of CYR61 can be induced by aberrant expression levels of transcription factors such as KRAS (e.g., during carcinogenesis of endometrial adenocarcinomas). The two genes CYR61 and ATF3 are indirectly connected through interactions CYR61 $\rightleftarrows$ FGFR1 $\rightleftarrows$ MAPK10 $\rightleftarrows$ TP53 $\rightleftarrows$ ATF3.

To compare our method to other methods, we also applied the Gaussian graphical models (GGM) algorithm of Schafer and Strimmer (2005) to this data. GGM is implemented in the `R` package *GeneNet.* Results extracted from the GeneNet algorithm are displayed in Table 8. The pairwise associations reflected in this table are largely consistent with the clusters detected using our algorithm. For example, GGM identifies (CYR61, ATF) as a potential important association. However, it fails to identify the association of

Table 8: Analysis of breast cancer data with GGM. The second column shows the pairs of nodes that are estimated to have high probability. The third column provides the p-value for the test of zero partial correlation. The fourth column provides the FDR-based q-value, which accounts for multiplicity in testing. The final column provides an estimate of the empirical posterior probability that the indicated nodes are connected in the GGM.

Rank	Doubleton	p-Value	q-Value	Empirical posterior probability
1	ESR1,GATA3	$<$0.0001	0.089	0.825
2	AR,FOXA1	0.0001	0.139	0.825
3	ESRRA,ERBB2	0.0002	0.147	0.635
4	GSTM1,GADD45	0.0006	0.245	0.635
5	BRCA1,BCL2	0.0006	0.256	0.635
6	CYR61,ATF	0.0007	0.269	0.635
7	BARD1,ERBB2	0.0010	0.287	0.635
8	HSF1,P53a	0.0010	0.291	0.635
9	FOLH1,IL10	0.0011	0.292	0.426
10	FOLH1,DCC	0.0016	0.3504	0.426

these genes with KRAS. Similarly, GGM detects the pairwise associations (ESR1, GATA3) and (AR, FOXA1), but gives no indication that these two pairs of variables might be related to each other and gene GSTP1a. At the same time, several pairwise associations identified by the GGM were not detected by the cluster model. We were not able to identify these pairwise correlations in a cursory search of PubMed, although this does not necessarily mean that these interactions are not biologically significant.

As a concluding note, the similarity of results obtained using our method and GGM indicates that little information was lost in the discretization procedure used in the preprocessing step of our method.

CHAPTER IV

OBJECTIVE BAYESIAN GRAPHICAL MODEL SELECTION

4.1 Introduction

In Chapter III I reviewed a method of clustering variables. When the interaction structure in the data is simple, we can use that method and combine it with a post-hoc analysis to learn about the true interaction structure between variables. However, with an increasing number of variables it is reasonable to expect that the interaction structure in the data will also become more complex.

The most serious disadvantage of the model in Chapter III is the restriction that any variable cannot be included in more than one cluster. Thus, detecting an important interaction between variables A and B may not be possible when there is an even stronger interaction between variables B and C. However, from the point of view of prediction of variable B, it will be necessary to include both those interactions. This is true for many high-throughput data, including the SNP data that is analyzed in the next chapter. I return to this point when I present the results of the analysis of SNP data in the next chapter.

As noted in the Introduction of this dissertation, the model proposed here is motivated by my desire to analyze high-dimensional high-throughput data. There is a clear need to move beyond clustering-type models towards models that allow for a general interaction structure. The log-linear model framework in equation (3.1) does allow for any arbitrary interaction amongst the variables. However, estimation of such models can become very hard even when there are very few variables in the model (~ 20). The primary impediment to employing a general log-linear model to high-throughput data is that testing an interaction between two variables requires fitting a model to the entire $P-$ way contingency table, where P is the total number of variables. Furthermore, even for small number of variables,

fitting a model to the entire contingency table requires a computationally-intensive algorithm.

Goodman (1970, 1971) and Haberman (1974) studied a sub-class of log-linear models in which the hypotheses tests for interactions and the estimated cell counts could be computed directly; that is, without employing an iterative scheme. They developed a class of models called *decomposable models.* Like the clustering models of Chapter III, decomposable models are *directly estimable*. Furthermore, interactions involving two variables can be tested in a marginal table that involves only a small subset of all the variables.

For a few decades, scientists have been studying Markov random fields to understand complex interaction structures. Markov random fields divide the set of variables into *cliques*, or the maximal sets with no missing edges. Unlike clustering-type models, these cliques can share variables. One of the fundamental results pertaining to the theory of Markov fields is the Hammersley-Clifford theorem (Besag 1974). Darroch, Lauritzen, and Speed (1980) utilized some close connections between the theory of Markov random fields and log-linear interaction models to define a new class of models for multidimensional contingency tables: graphical models. They further established that the class of *decomposable models* is a proper subset of the class of graphical models, and so decomposable models inherit all the nice properties of the graphical model class. They also established that the graphical model class is a subset of the hierarchical log-linear model class.

Another important property of graphical models is that they are "dense" in the space of log-linear models. This means that for any general log-linear model, one can find a graphical model which includes a few more interactions and includes it. Alternatively, a graphical model can be obtained by removing few interactions from a general log-linear model.

In graphical model terminology, variables are called *vertices*, or *nodes. Nodes* of a graph are joined by *edges.* In terms of the log-linear model of equation (3.1), edges

correspond to interaction terms - deleting an edge is equivalent to removing corresponding interaction terms from the log-linear model equation. Furthermore, the largest interaction (maximal subset) containing a particular subset of the *nodes* are called *cliques.* Henceforth, I use the terminology of graphical models and that of log-linear models interchangeably.

4.2 Limitation of log-linear model: need for collapsibility

Even the simplest log-linear models can be hard to estimate. Consider the model:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}. \tag{4.1}$$

There exists no closed form for the cell estimates of this model. Bishop, Fienberg, and Holland (1975) present an iterative algorithm for estimating this model that involves fitting the marginal table of variables $\{1,2\}$, $\{2,3\}$, and $\{1,3\}$ iteratively. The largest model involving 3 variables which does not need an iterative fitting algorithm is the saturated model presented in equation (3.1), for which the cell estimates are the observed cell counts. The largest unsaturated model which does not need an iterative fitting algorithm is a model which contains only two of the three two-way interactions, for example,

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)}. \tag{4.2}$$

The estimated cell counts of the above model can be expressed in closed form as

$$\hat{m}_{123(ijk)} = \frac{x_{12+(ij)} x_{1+3(ik)}}{x_{1++(i)}}. \tag{4.3}$$

In this equation, $x_{12+(ij)}$ represents the marginal cell counts in the table of variables $\{1,2\}$, and $x_{1+3(ik)}$ represents the marginal cell counts in the table of variables $\{1,3\}$, and $x_{1++(i)}$

represent the marginal counts of variable 1. The LRS for testing any log-linear model containing 3 variables with respect to the saturated model in equation (3.1) is obtained as:

$$Z_n = \sum_{i,j,k} x_{123(ijk)} \log\left(\hat{m}_{123(ijk)}\right). \tag{4.4}$$

Hence, if a model is directly estimable, there exists a closed form solution for testing it against the saturated model. Note that this is not the case with the model in equation (4.1). Now consider the model

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)}. \tag{4.5}$$

The model in equation (4.5) is directly estimable. Let's say we are interested in testing the significance of the $\{1,2\}$ interaction in the model in equation (4.2). We would compute the LRS for that model with respect to the saturated model, and the LRS of the model in equation (4.5). The LRS for the test of $\{1,2\}$ interaction can then be computed by taking the difference of the two quantities, both of which exist in closed form. Hence the LRS for the test of interest also exists in a closed form. Note that this is not the case if we were interested in testing the significance of the $\{2,3\}$ interaction in the model in equation (4.1), which would correspond to a test of model (4.1) versus the model (4.2). Now consider two models for the marginal table of variables $\{1,2\}$:

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}, \text{ and,} \tag{4.6}$$

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)}. \tag{4.7}$$

It can be verified that testing the model in equation (4.2) vs. the model in equation (4.5) is equivalent to testing the model in equation (4.6) vs. the model in equation (4.7). In the

language of Bishop, Fienberg, and Holland (1975), we can *collapse* the three-way table over variable 3 to obtain the test for the interaction $\{1,2\}$ in the model in equation (4.2).

Collapsibility is a very desirable property if we were to use log-linear models with many variables. Unfortunately, most tests in most log-linear models do not permit collapsibility. For example, consider testing the model in equation (4.1) versus the model in equation (4.2). In this case, we cannot collapse the three-way table over variable 1 to obtain the test of the $\{2,3\}$ interaction.

Lack of collapsibility restricts the usefulness of log-linear models to only a small set of variables. Consider a log-linear model with 35 variables which only contains only two-way interactions. Testing for any of those two-way interactions requires that we obtain the cell estimates from a table of 35 variables. Even if we ignore the fact that we would need an iterative algorithm, every step of which would try to fit $\binom{35}{2}$ marginal configurations, the problem of storing a 35-way table cannot be ignored and is not feasible in software packages like `R` and `MATLAB`.

Graphical models provide a convenient mechanism for achieving collapsibility. This is discussed in the next section.

4.3 Graphical models in log-linear model notation

Christensen (1997) provides an informal definition for graphical models: a graphical model is a hierarchical log-linear model for which if all the two-way interactions corresponding to a higher order interaction are included, then that higher-order interaction must also be included in the model. In graph theory, such highest-order interactions are called *cliques.* For example, let's consider a model of 5 variables and the following interactions:

$$\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\},\{2,5\},\{4,5\}. \tag{4.8}$$

Let G_1 denote the smallest graphical model that includes all the interactions in equation

(4.8). Then G_1 must also include the interactions $\{1,2,3,4\}$ and $\{2,4,5\}$. We call these highest interactions as *cliques* of the model G_1, and denote G_1 as below:

$$G_1 :\equiv [1\ 2\ 3\ 4][2\ 4\ 5]. \tag{4.9}$$

Note that since G_1 is also necessarily a hierarchical log-linear model, it must include all the three-way interactions corresponding the variables in the $\{1,2,3,4\}$ interaction. Note that there are lot of log-linear models that do not include all the interactions in model G_1, but include all the interactions in equation (4.8). Let n be the number of all such models and let they be denoted by $\mathcal{M}_1, \ldots, \mathcal{M}_n$. Then, G_1 also happens to be the smallest graphical model that includes *all* the interactions of models $\mathcal{M}_1, \ldots, \mathcal{M}_n$. In graph theory, G_1 is called the *Independence Map (I-Map)* of each of the models $\mathcal{M}_1, \ldots, \mathcal{M}_n$. In graph theory, G_1 corresponds to Figure 3a.

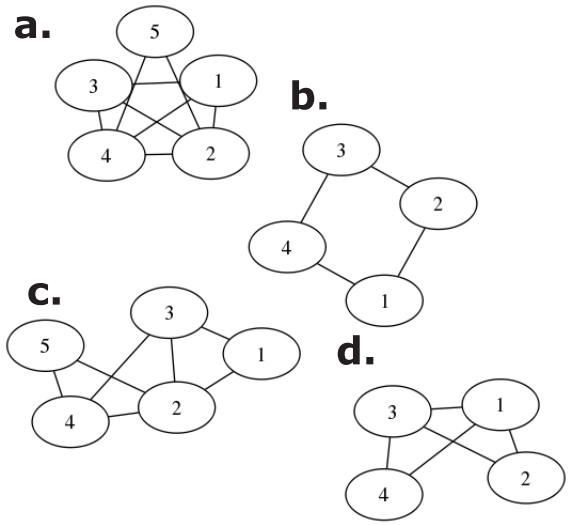

Figure 3: Graphical models

Now consider the smallest graphical model G_2 that includes all the interactions in

equation (4.8) except for {1,4}. In the notation of graphical models, we can write:

$$G_2 :\equiv [1\ 2\ 3][2\ 3\ 4][2\ 4\ 5]. \tag{4.10}$$

G_2 also corresponds to Figure 3c. In the diagrammatic representation, G_1, and G_2 differ only by the *edge* joining the *nodes* 1 and 4. We say G_2 is obtained by breaking the edge {1,4} of model G_2. Since both the models G_1 and G_2 are necessarily hierarchical, a test of hypothesis like below:

$$H_0 : G_2 \tag{4.11}$$

$$vs.$$

$$H_1 : G_1, \tag{4.12}$$

will translate in log-linear model notation to a test of alternative model G_1 against a *nested* null model which is obtained by setting the interaction terms u_{14}, u_{124}, u_{134}, and u_{1234} to 0. Hence each edge corresponds to all the interactions of the log-linear model that contain both the nodes joining that edge.

4.3.1 Implication of the Hammersley-Clifford theorem

The connection between log-linear models and graphical models can also be understood from the Hammersley-Clifford theorem as presented in Besag (1974). For example, with reference to graph G_1, let X_i denote the random variable at node i, for $i = 1, \ldots, 5$. Then the Hammersley-Clifford theorem requires that

$$P(X_1, \ldots, X_5) = \frac{P(X_1, X_2, X_3, X_4)\, P(X_1, X_4, X_5)}{P(X_1, X_4)}, \tag{4.13}$$

where the function $P(\cdot)$ above is the joint probability distribution of the random variables in

the argument list. Assuming a *Dirichlet* distribution over the cells of the 5-way contingency table, we have

$$m^{(5)}_{ijklo} \propto P\left(X_1 = i, X_2 = j, X_3 = k, X_4 = l, X_5 = o\right) \tag{4.14}$$

$$\propto \frac{m^{(4)}_{ijkl} m^{(3)}_{ilo}}{m^{(2)}_{il}}. \tag{4.15}$$

In this equation, $m^{(p)}$ represents the expected cell counts of the $p-$way table involving the variables whose indices appear in the subscript. The equation (4.15) is a direct consequence of the equation (4.13). Hence the Hammersley-Clifford theorem implies the equivalence of the cliques and the highest interactions of the log-linear equation corresponding to the graph.

4.4 Collapsibility in graphical models

Birch (1963, 1964) provided results for computing the maximum likelihood estimates of cell counts for contingency tables. Those results can also be found in Bishop, Fienberg, and Holland (1975). According to Birch's results, any maximum likelihood solution for any hierarchical log-linear model must satisfy all the maximal interaction constraints corresponding to that model. Furthermore the maximum likelihood solution for which cell estimates are non-negative is unique. Hence, Birch's results imply that any model of 5-variables satisfying equation (4.15), will have the expected cell counts $\hat{m}^{(p)}$ satisfy the following relationship:

$$\hat{m}^{(5)}_{ijklo} = \frac{\hat{m}^{(4)}_{ijkl} \hat{m}^{(3)}_{ilo}}{\hat{m}^{(2)}_{il}} \tag{4.16}$$

$$= \frac{x_{1234+(ijkl)} x_{1++45(ilo)}}{x_{1++4+(il)}}. \tag{4.17}$$

The result in equation (4.16) is similar to the results in Goodman (1970, 1971). It means that estimated cell counts for a graphical model are obtained as a product of the estimated

cell counts of individual cliques divided by the estimated cell counts of the clique *separators,* or variables that are common between different cliques. When individual cliques correspond to log-linear models that are directly estimable, the entire model is directly estimable as in equation (4.17). Equation (4.16) then implies collapsibility in graphical models: it can be easily verified that if the interest is in a test of hypothesis corresponding to any edge of a graphical model, then all the variables not included in the cliques containing that edge can be collapsed. For example, when testing

$$H_0 : G_2 \tag{4.18}$$

$$vs.$$

$$H_1 : G_1, \tag{4.19}$$

we can collapse the 5-way table over the node labeled 5.

4.4.1 Decomposable graphical models

A class of log linear models called *decomposable* models were first introduced by Haberman (1974). The decomposable model class is a subclass of hierarchical log linear models, in which all models are directly estimable. Darroch, Lauritzen, and Speed (1980) later showed that the decomposable model class is a strict subset of the graphical model class. Decomposable models do not have any closed loops involving four or more cliques. For example, graph G_3 in Figure 3b can be written in clique notation as $G_3 :\equiv [1\ 2][2\ 3][3\ 4][1\ 4]$. Since, the cliques form a closed-loop, this model is not directly estimable. Adding the diagonal edge, $\{1,3\}$ yields the model $G_4 :\equiv [1\ 2\ 3][1\ 3\ 4]$, shown in Figure 3d, which is directly estimable. Since *decomposability* requires that a graph have certain diagonal edges, decomposable graphical models are also called *triangulated* graphs.

4.5 Proposed model space for graphical model selection

The decomposable model class is directly estimable and provides a convenient mechanism for collapsibility. For these two reasons, they are an attractive choice of model class for modeling complex interaction networks. However, the general decomposable model class is still very large, and it might not be possible to navigate through the entire model space conveniently. Moreover for any moderate number of nodes ($\sim 50 - 500$), a subclass of the decomposable model class can be used to understand the interaction network. For my applications, I propose the following additional restrictions on the model space. In Section 4.6, I propose a Markov chain Monte Carlo algorithm to sample from that model space.

4.5.1 Model constraints

R1. Cliques are restricted to have size of 4 or less.

R2. A clique of size 3 may share *at most* one edge with *at most* one other clique.

R3. Cliques of size 4 may *not* share edges with other cliques.

Restrictions (R2) and (R3) above ensure that the computations in the proposed MCMC are done entirely within one clique, and restriction (R1) restricts the size of that clique to at most 4. Also, as long as the chain is initialized with a state satisfying restrictions (R1)-(R3), any state proposed in the following MCMC scheme will automatically satisfy those restrictions. Since the visited models are always decomposable graphs, the LRS for the hypothesis test between two nested models is obtained in closed form.

4.5.2 Prior distribution

I assume a prior distribution that is uniform over the space of models.

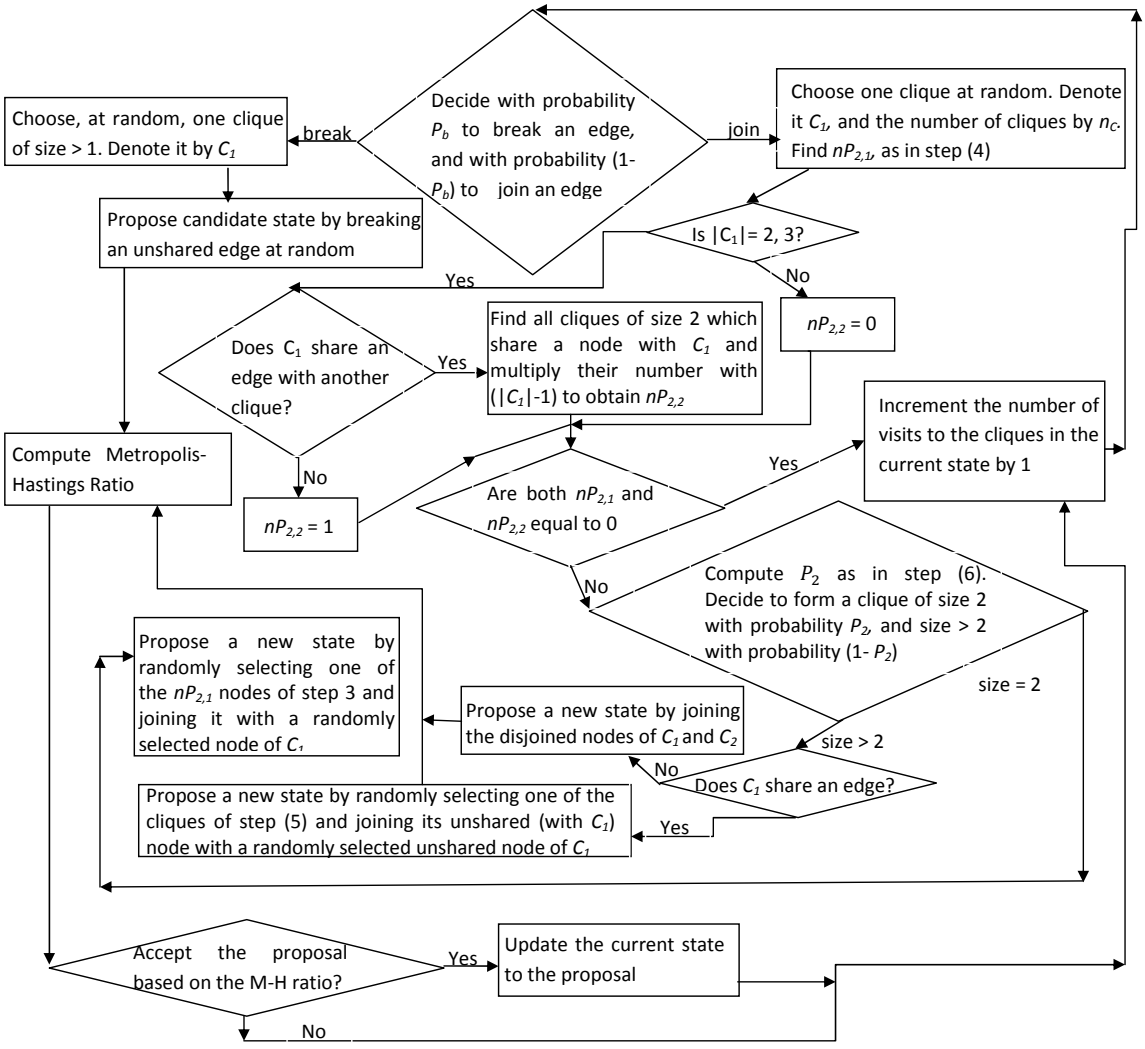

Figure 4: Diagrammatic representation of the proposed MCMC methodology

4.6 MCMC

4.6.1 Algorithm

The sampling algorithm described below ensures that both the current and proposed states of the Markov chain are decomposable graphs. Importantly, this alleviates the need for checking restrictions (R1)-(R3). In the "join" step, it is proposed to join nodes of cliques to form newer edges. In the "break" step, it is proposed to separate nodes connected by an edge. Let $|A|$ denote the number of nodes in clique A. The iterative algorithm is described in steps as follows. A diagrammatic representation of the algorithm above is in Figure 4.

1. Initialize the Markov chain with a graph that has no edges between any pair of nodes.

2. Go to step (3) with probability P_b and step (4) with probability $(1 - P_b)$.

3. Choose at random clique amongst the set of cliques $\{C : |C| > 1\}$, and denote it as C_1. If the current state does not include any clique which includes more than one node, go to step (11); otherwise, go to step (7).

4. Of *all* the cliques in the current state, choose one at random. Denote this clique as C_1 and let the total number of cliques be denoted by n_C. Find all nodes which are *not* connected to C_1, neighbors of C_1, neighbors of neighbors of C_1 and so on, and multiply their number by $|C_1|$ to obtain $nP_{2,1}$. If $|C_1|$ is 2 or 3, go to step (5); otherwise, set $nP_{2,2} = 0$, and go to step (6).

5. If C_1 shares an edge with another clique, denote the other clique by C_2, set $nP_{2,2} = 1$, and go to step (6); otherwise, find all cliques that belong to the set $\{C : |C| = 2\}$, and share a node with C_1 and multiply their number by $(|C_1| - 1)$ to obtain $nP_{2,2}$. Go to step (6).

6. If both $nP_{2,1}$ and $nP_{2,2}$ are 0, skip to step (11); otherwise, compute $P_2 = \frac{nP_{2,1}}{nP_{2,1}+nP_{2,2}}$. Go to step (8) with probability P_2 and step (9) with probability $(1 - P_2)$

7. Propose a new state by breaking a randomly chosen edge of C_1 which is *not* shared with other cliques. Go to step (10).

8. Propose a new state by randomly selecting one of the $nP_{2,1}/|C_1|$ nodes of step (4) and joining it with a randomly selected node of C_1. Go to step (10).

9. If C_1 shares an edge with another clique, propose a new state which contains a new clique containing 4 nodes formed by joining the unjoined nodes of C_1 and C_2; otherwise, propose a new state by randomly selecting one of the cliques of step (5) and joining its unshared (with C_1) node with a randomly selected unshared node of C_1. Go to step (10).

10. Calculate the Metropolis-Hastings ratio as a product of the Bayes' factor and the ratio of the transition probabilities as tabulated in Table 9. If the proposal is accepted, update the clique configurations of the current state. Go to step (11).

11. Increment the number of visits to the cliques in the current state by 1. Go back to step (2).

I used $P_b = 0.5$, for all my examples and is working well in the sense that the acceptance is generally similar for "break" and "join" moves. This choice of P_b is also required to maintain symmetricity in the moves involving breaking and joining of edges.

4.6.2 Transition probabilities

The transition probability ratios for the various steps of MCMC are as in Table 9. Below, n_C^1 is the number of cliques in the larger model which include more than 1 node and n_C^0 denotes the number of *all* the cliques in the smaller model.

Table 9: Transition probability ratios for the proposed MCMC methodology. The numerator of the terms in the last column is the probability of proposing the smaller model when the current state is the bigger model, the denominator is the probability of proposing the bigger model when the current state is the smaller model. The two models only differ by an edge. The first column shows examples of the corresponding moves.

$B \rightleftharpoons S$	Description	$\frac{T(B \to S)}{T(S \to B)} =$
$[1\ 2\ 3\ 4] \rightleftharpoons [1\ 3\ 4]\,[2\ 3\ 4]$	$\rightharpoonup$:An edge is broken in a clique that includes 4 nodes $\rightharpoondown$:Unshared nodes are joined for two cliques, both including 3 nodes, that share a node amongst each other P_2 is the same for both the cliques that share an edge	$\frac{P_b * \frac{1}{n_C^1} * \frac{1}{6}}{(1-P_b) * \frac{1}{n_C^0} * (1-P_2) * 2}$
$[1\ 2\ 3]\,[1\ 2\ 4] \rightleftharpoons [1\ 3]\,[1\ 2\ 4]$	$\rightharpoonup$:An edge is broken in a clique that includes 3 nodes and shares another edge $\rightharpoondown$:Unshared nodes are joined for a clique that includes 3 nodes with a clique that shares a node with it and includes one other node $nP_{2,2}$ is the number in step (5) above for clique $[1\ 2\ 4]$ in the smaller model	$\frac{P_b * \frac{1}{n_C^1} * \frac{1}{2}}{(1-P_b) * \frac{1}{n_C^0} * (1-P_2) * \frac{1}{nP_{2,2}}}$
$[1\ 2\ 3] \rightleftharpoons [1\ 3]\,[2\ 3]$	$\rightharpoonup$:An edge is broken in a clique that includes 3 nodes and does *not* share any edge $\rightharpoondown$:Unshared nodes are joined for two cliques, both including 2 nodes, that share a node with each other $nP^1_{2,2}$ and $nP^2_{2,2}$ are respectively, the numbers in step (5) above for cliques $[1\ 3]$ and $[2\ 3]$	$\frac{P_b * \frac{1}{n_C^1} * \frac{1}{3}}{(1-P_b) * \frac{1}{n_C^0} * \left(\sum_{i=1,2} \frac{(1-P_2^i)}{nP^i_{2,2}} \right)}$
$[1\ 2] \rightleftharpoons [1]\,[2]$	$\rightharpoonup$:An edge is broken in a clique that includes 2 nodes $\rightharpoondown$:Nodes are joined for two cliques which are not connected either directly or through other cliques Index i ranges to include *all* cliques in the smaller model that contain either the first node or the second node Variables 1 and/or 2 might already be included in other cliques, in which case the smaller model will have 1 or 0 new cliques	$\frac{P_b * \frac{1}{n_C^1}}{(1-P_b) * \frac{1}{n_C^0} * \left(\sum_{i_1=1} \frac{P_2^i}{nPsC^i} \right)}$ $\frac{1}{nPsC^i} = \frac{1}{nP^i_{2,1} * \lvert C_i \rvert}$

4.7 Examples

4.7.1 Simulated gene expression data

I applied the algorithm described above to the simulated example in Section 3.6. I used only 100 variables from the actual dataset containing the 1000 variables. As before, interacting variables were divided into three clusters of different sizes: {(1, 2, 3, 4), (5,6), (7, 8, 9)}. The interaction cluster (7, 8, 9) in Table 2 fits the models $[7\ 8][9]$, $[7\ 9][8]$, and $[8\ 9][7]$ equally well and each of these models, when compared to the saturated model, produces an LRS value of 13.18 with 2 degrees of freedom. I therefore changed that interaction cluster to another in which the three-way interaction is stronger. The modified cluster (7, 8, 9) is in Table 10. The remaining part of the dataset was the same as the dataset used to report the results of Section 3.6.

Table 10: Modified true cluster in the simulated gene expression data

Cluster (7, 8, 9)					
	$x_9 = 0$			$x_9 = 1$	
	$x_8 = 0$	$x_8 = 1$		$x_8 = 0$	$x_8 = 1$
$x_7 = 0$	18	17	$x_7 = 0$	3	22
1	7	18	1	32	33

The computation times for a `FORTRAN` routine is about 4-5 hours for every 10 million updates of the MCMC algorithm using either the local or the MoM and iMoM prior densities. Summary statistics are tabulated in Table 11. The MCMC algorithm is able to detect the cliques $[5\ 6]$ and $[7\ 8\ 9]$ with high probability. Instead of detecting the the 4-way clique $[1\ 2\ 3\ 4]$, it detected the existence of cliques $[1\ 2\ 4][3\ 4]$. In a test of model $H_1 : [1\ 2\ 4][3\ 4]$ versus model $H_2 : [1\ 2\ 3\ 4]$, the p-value obtained for the LRS is 0.58, which suggests that this particular observation of data under this model supports both structures approximately equally well.

Table 11: Graphical model analysis of the simulated gene expression data. The posterior probability estimates are based on the number of times each interaction appeared in 10 million updates of the MCMC algorithms for each prior.

Local ($c = 2$)				MoM ($\tau_1 = 0.348$)				iMoM ($\tau_1 = 0.133$)			
7	8	9	0.9947*	7	8	9	>0.9999*	7	8	9	>0.9999*
3	4		0.9579*	3	4		>0.9999*	3	4		>0.9999*
1	2	4	0.9204*	1	2	4	0.9763*	1	2	4	0.9805*
5	6		0.8064*	5	6		0.9425*	5	6		0.9527*
54	94		0.7819	54	94		0.9404	54	94		0.9464
18	91		0.6564	18	91		0.9286	18	91		0.9211
16	98		0.6325	16	98		0.8509	16	98		0.9131
11	90		0.6323	11	90		0.8486	11	90		0.8656
12	69		0.6036	12	69		0.8453	12	69		0.8653
34	94		0.5830	34	94		0.8188	34	94		0.8619
6	78		0.5821	84	88		0.8186	2	86		0.8613
2	86		0.5702	2	86		0.8086	4	61		0.8355
4	61		0.4437	4	61		0.8072	84	88		0.8281
84	88		0.4075	47	53		0.8047	6	78		0.8276
47	53		0.3977	6	78		0.8032	47	53		0.8246

The results in Table 11 are based on 10 million updates. Multiple runs of the MCMC algorithm with different starting values produced similar results. However, the non-local prior densities were unable to detect clique $[1\ 2\ 4]$ for up to 50 million updates when the Markov chain is initialized with a graph with no edges between the nodes. The reason for this behavior is not hard to comprehend. For the MCMC algorithm to detect the clique $[1\ 2\ 4]$, at least two of its edges should form easily. However, the LRS values corresponding to the tests $H_1 : [1][4]$ versus $H_2 : [1\ 4]$, and $H_1 : [2][4]$ versus $H_2 : [2\ 4]$ are both 0. This value of the LRS corresponds to Bayes' factor values in favor of the H_2 versus H_1 of 0.064, 0.003, and 0.003 for the local, MoM and iMoM prior densities. In other words, it is about 21 times harder for each of those edges to form with the non-local priors as compared to the local prior.

Therefore, I used the output from the run with the local prior density to generate initial values for runs with the non-local prior densities. Results based on this initialization

scheme are reported in Table 11. I used multiple starting models but forced the clique $[1\ 2\ 3\ 4]$ to be included in all the starting models.

As a precautionary note, higher order interactions are less likely to be picked in the absence of lower order interactions under the non-local versus local priors. Hence, I advise the user to run the MCMC with the different priors. If there are higher order interactions that are not detected by the runs for non-local prior densities, but are indeed assigned high posterior probability by the local prior density, then they must be because of reasons similar to the one discussed here. There is no guarantee, however, that running the MCMC with the local prior will always be able to detect such higher order interactions.

With reference to the results in Table 11, the "best" 15 cliques obtained under the different alternative prior densities appear to be similar. However, the posterior probabilities assigned to those cliques by the non-local alternatives are much higher than the local alternative prior, and among the non-local priors they are consistently higher for the iMoM density. It should be noted that these "best" cliques indeed correspond to the moderate to strong interactions in the data.

4.7.2 Distribution of posterior probabilities

At first glance it might appear from Table 11 that the *non-local* priors assign higher posterior probabilities to all the visted cliques than the local priors do. However, that is not the case. Since it is not computationally feasible to keep track of all the cliques possible with 100 nodes, I considered a smaller dataset with only 20 binary variables formed by taking the variables 1 through 20 in the example above. With 20 variables, there are 6195 cliques possible. I ran the algorithm in Section 4.6 to obtain 10 million updates of the posterior distributions elicited by the local and non-local priors.

The local prior visited 1228 cliques in the 10 million updates of the algorithm; the MoM prior visited 424 cliques and the iMoM prior visited 476 cliques. I then computed

the summary statistics for the posterior probabilities of all the cliques that were ranked between 51 through 300 for each of the alternative prior densities. I chose the upper rank cut-off of 51 to demonstrate the posterior probability distibution of the cliques that have been assigned ranks higher than 50. I chose the lower rank cut-off of 300 to ensure that we have reliable estimates of the posterior probabilities for the *non-local* priors. The density plots of the posterior probabilities are shown in Figure 5. The *non-local* priors assigned much smaller posterior probabilities to those cliques when compared to the *local* prior. This behavior is expected because the *non-local* priors assign much lower probabilities to small effect sizes *a priori*.

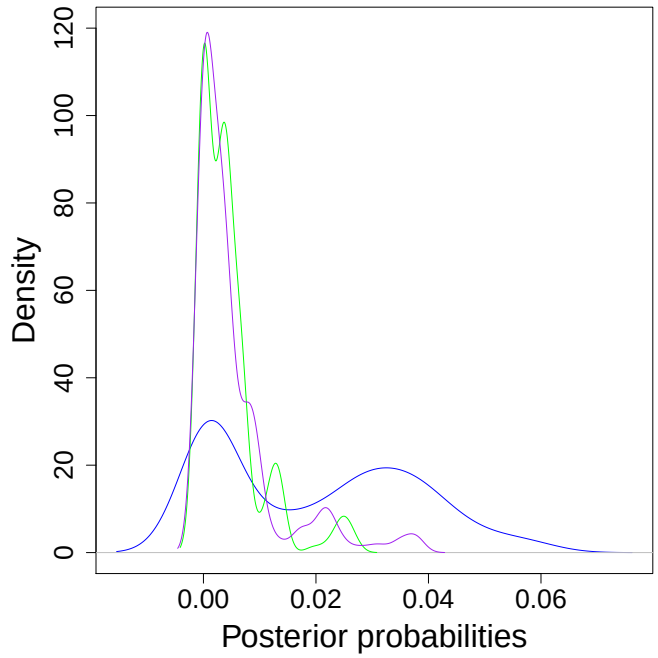

Figure 5: Distribution of posterior probabilities. The figure shows empirical density estimates of posterior probabilities of the cliques ranked between 51 and 300 according to the different alternative prior densities in an example consisting of 20 nodes. The blue curve shows the density plot for the local prior. The corresponding plots for the MoM and iMoM priors are, respectively, in green and purple.

4.7.3 Comparison to clustering method

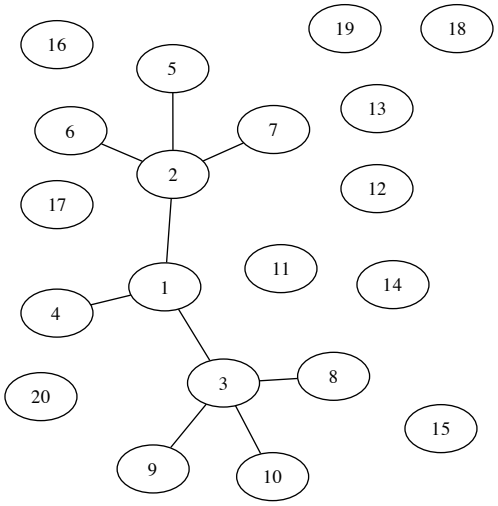

Figure 6: Simulation example comparing graphical model to clustering model

To demonstrate the utility of graphical model algorithm of Section 4.6 over the clustering method of Chapter III, I constructed another example with 20 binary variables and 120 observations. The true interactions in the data have varying strengths and correspond to graph cliques: $[1\ 2][1\ 3][1\ 4][2\ 5][2\ 6][2\ 7][3\ 8][3\ 9][3\ 10]$. Variables 11 through 20 were then generated randomly. The *true* network is shown as in Figure 6. A comparison of the two methods with the local prior $(c = 2)$ is presented in Table 12.

While the graphical model method is able to identify the true 9 interactions and assign them high posterior probabilities, the clustering model is able to identify only 3 of the strongest interactions with high posterior probabilities. The graphical model method also identifies 3 spurious interactions with posterior probability of at least 0.40. In general, and as can also be seen in Table 12, the clustering model does not tend to assign high posterior probabilities to interactions involving any randomly generated variables (false positives). The reason for such behavior could be easily explained. With 20 variables there are 190

Table 12: Comparison of the graphical model to the clustering model. The posterior probability estimates are based on the number of times each interaction appeared in 10 million updates of the MCMC algorithms for the two methods. The last column shows the LRS values corresponding to each interaction detected by the graphical model method. Each of those LRS values correspond to a test with 1 degree of freedom.

Clustering model				Graphical model			
1	4		0.9162*	1	4	0.9637*	21.49
2	5		0.909*	2	5	0.9624*	24.53
3	10		0.6381*	3	10	0.9322*	15.13
6	13		0.2574	1	3	0.8561*	13.83
3	9	10	0.1427	2	7	0.8296*	10.75
12	19		0.1235	3	9	0.7769*	11.42
3	9		0.1068	1	14	0.6736	9.77
14	17		0.0606	1	2	0.6596*	10.94
2	5	7	0.0538	3	8	0.6576*	8.80
1	4	14	0.0447	1	20	0.5074	7.58
3	8		0.0008	6	13	0.4455	6.72
8	19		0.0007	2	6	0.407*	7.32
9	10		0.0003	12	19	0.3141	4.83
1	14		0.0002	2	15	0.1757	4.07
7	9		0.0001	3	17	0.1574	4.08

possible 2-way interactions, so at least some of them are likely to be moderately significant false positives. These spurious cliques will be detected by the graphical model method with moderately high posterior probability. On the other hand, each randomly generated variable is expected to have *several* interactions with comparable strength. The clustering method tends to distribute the posterior probability over these several spurious interactions, and hence the posterior probability assigned to each one of them individually tends to become small.

At this juncture I note that neither of the methods is expected to perform better than the other in every possible scenario. Whereas the clustering model is better if there exist only *few* interactions between the variables, the graphical model method will outperform the clustering model if the data is generated by a complex interaction network.

4.7.4 Breast cancer data

I next applied these methods to the breast cancer data described in Section 3.7. A comparison of the gene clustering method of Chapter III to other methods is presented in Section 3.7. Although the graphical model method is expected to detect more false positives, it is also expected to detect more true positives. The results of the analysis with the graphical model method are presented in Table 13.

Table 13: Analysis of the breast cancer data with the graphical model method. The posterior probability estimates are based on the number of times each interaction appeared in 10 million updates of the MCMC algorithms for each prior.

Local ($c = 2$)		MoM ($\tau_1 = 0.348$)		iMoM ($\tau_1 = 0.133$)	
BRCA1,JNK	0.8963	BRCA1,JNK	0.978	BRCA1,JNK	0.9192
ESR1,GATA3	0.8752	ESR1,GATA3	0.9758	ESR1,GATA3	0.8829
AR,FOXA1	0.8658	AR,FOXA1	0.9715	AR,FOXA1	0.8670
HSF1,P53a	0.7649	HSF1,P53a	0.9288	HSF1,P53a	0.8183
P53b,WNt10b	0.6681	P53b,WNt10b	0.8460	GSTP1a,IKBKE	0.7371
GSTP1a,IKBKE	0.6099	ESRRA,ERBB2	0.8411	P53b,WNt10b	0.7065
CDKN2B,TSG101	0.5836	GSTP1a,IKBKE	0.8069	CDKN2B,TSG101	0.7045
EGR1,CYR61	0.5787	FOSL1,CYR61	0.8057	EGR1,CYR61	0.6992
ESRRA,ERBB2	0.5591	EGR1,CYR61	0.7989	ESRRA,ERBB2	0.6819
FOSL1,CYR61	0.5535	CDKN2B,TSG101	0.7601	GSTP1b,FOSL1,AKT2	0.6497
GSTP1b,FOSL1,AKT2	0.5426	AR,ERBB2	0.7485	FOSL1,CYR61	0.6332
CYR61,ATF	0.5287	ERK2,NRAS	0.7411	AR,ERBB2	0.6313
ERK2,NRAS	0.5220	ESR2,IL10	0.7387	CYR61,LSP1	0.581
ESR2,IL10	0.5168	CYR61,ATF	0.7336	ESR2,IL10	0.5707
AR,ERBB2	0.5041	CYR61,LSP1	0.6660	ERK2,NRAS	0.5697
CYR61,LSP1	0.4690	FOSL1,IL-13	0.6399	KRAS,CYR61,ATF	0.5529
Brca1-delta11b,ERCC2	0.4522	ERBB2 V-erb-b2,PLGL	0.6369	ERBB2 V-erb-b2,PLGL	0.5086
FOSL1,IL-13	0.4439	Brca1-delta11b,ERCC2	0.6368	KIAA0272,CDKN2B	0.5042
ERBB2 V-erb-b2,PLGL	0.4376	Brca1-delta11b,ESR2	0.6090	P450 XVIIA1,WNt10b	0.5034
Brca1-delta11b,ESR2	0.4252	IFI27,TNFa	0.5976	IFI27,TNFa	0.4882
IFI27,TNFa	0.4083	P450 XVIIA1,WNt10b	0.5653	Brca1-delta11b,ERCC2	0.4807
KRAS,CYR61,ATF	0.3892	KIAA0272,CDKN2B	0.5439	FOSL1,IL-13	0.4616

The graphical model method identified all biologically significant gene interactions that were detected by the gene clustering method of Section 3.7. In addition, there are a few more interactions detected by the graphical model method that are supported in the bioinformatics literature. The association between AR and ERBB2 pathways is documented in Naderi and Hughes-Davies (2008), and also has been reported in Sanga et al. (2009). More interesting are the additional interactions of protein CYR61. The interac-

tions {CYR61,ATF} and {KRAS,CYR61,ATF} have been detected by both the clustering model and the graphical model method. In addition, the graphical model method has detected {EGR1,CYR61}, {FOSL1,CYR61} and {CYR61,LSP1}. Note that one could not have detected all of these interactions of CYR61 with the clustering method, because that method forces one gene to be in at most one cluster.

To investigate whether these additional interactions of CYR61 were actually biologically significant findings, I searched through the SA Biosciences website (`http://www.sabiosciences.com`). Cysteine-rich 61 (CYR61/CCN1) is a secreted, extracellular matrix-associated signaling molecule that belongs to the CCN gene family. The other members of the CCN family of genes are CTGF/CCN2, NOV/CCN3, and the recently identified Wnt1-induced secreted proteins WISP-1/CCN4, WISP-2/CCN5, and WISP-3/CCN6.

I found associations between CTGF and other genes. First, CTGF is suspected of associating with the Transforming growth factor beta 1 (TGFB1) polypeptide by Weston, Wahab, and M. (2003). TGFB1 increases the stability of the protein CDKN1A (Gong et al. 2003). CDKN1A is modulated by EGR1 (Ragione et al. 2003). Hence, I could deduce the indirect association CYR61 (CCN1) $\leftrightarrows$ CTGF (CCN2) $\leftrightarrows$ TGFB1 $\leftrightarrows$ CDKN1A (p21Cip1) $\leftrightarrows$ EGR1.

Luo et al. (2005) indicate that TGFB1 is down-regulated by the Transforming growth factor, beta receptor II (TGFBR2) gene. As an AP1 site has been identified in the TGFBR2 promoter, upregulation of FOSL1 (also called FRA-1) may stimulate TGFBR2 transcription (Lombaerts et al. 2006). These observations suggest the indirect association CYR61 (CCN1) $\leftrightarrows$ CTGF (CCN2) $\leftrightarrows$ TGFB1 $\leftrightarrows$ TGFBR2 $\leftrightarrows$ FOSL1 (FRA-1).

A comparison of the clustering method to GGM (Schafer and Strimmer 2005) is presented in Section 3.7. The graphical model method has identified more biologically significant gene interactions than both the clustering method and the GGM. In addition, most of the additional interactions detected by the graphical model method agree with the results

of the clustering model method presented in Table 4. Note that even though I have not been able to verify the existence of the other interactions in the available literature, it does not necessarily mean that those interactions are not biologically significant. A complete analysis would require independent laboratory verifications of the results presented herein.

4.8 Computational issues

The alarm network described in Section 4.9 and the SNP data in Chapter V present a unique computational problem. The probability landscape for these problems is extremely "jagged", so the algorithm in Section 4.6 produces a Markov chain which has an acceptance rate of less than 1%. This problem occurs because the Markov chain reaches a state in which it is unable to form any new edges given the model constraints, and any move involving breaking an already existing edge involves a very large Bayes' factor. Thus, the Metropolis-Hastings update probability for any move is essentially 0. Moreover, when the Markov Chain reaches such a state, none of the "neighbors" of the current state form good candidate proposals, so algorithms like the *shotgun stochastic search* (SSS, e.g., Jones et al. 2005) don't work either.

For these reasons, I designed an approximate MCMC algorithm that was able to escape sharp local maxima of the posterior distribution. In my examples, I observed that small subsets of the set of all graph variables would tend to fit *one*, and sometimes, *two* models, with very high posterior probabilities. Since the cliques of the graph are obtained by fitting models to such small subsets, the probability dispersion over the model space appears to be very low. Each individual run of the algorithm in Section 4.6 tends to stop updating after reaching a "local maximum" in the model space. However, if I generate multiple "local maxima" with multiple runs of the algorithm in Section 4.6 and multiple initial values, it is not unreasonable to expect that the cliques corresponding to the such highest probability model(s) will appear in several of such "local maxima."

To obtain the most important cliques in the data, I modified the algorithm in Section 4.6 in the following way. After every n_1 iterations, I paused the (*main*) chain described in Section 4.6 and ran an *auxiliary* chain for n_2 iterations. At the end of the n_2 iterations, I started a new (*main*) chain initialized at the last state of the *auxiliary* chain. The *auxiliary* chain is constructed using a similar algorithm as the *main* chain, however, the pseudo-Bayes' factor of the *auxiliary* chain is arbitrarily defined to be

$$BF_{aux} = (BF_{main})^{1/f(c_{aux},iter)} . \tag{4.20}$$

The function $f(c_{aux}, iter)$ is similar to the "temperature" parameter of simulated annealing algorithm (Kirkpatrick, Gelatt, and Vecchi 1983; Cerny 1985). In effect, it produces a probability distribution over the model space which can be easily explored by the sampling scheme described in Section 4.6. However, since I do not maintain the detailed balance condition while transitioning between the *main* and *auxiliary* chains, the modified algorithm does not provide samples from the target distribution even as the number of updates becomes large. I note that it is unlikely that the higher temperature chain of a simulated annealing algorithm will produce a state that would be improvement over the "local maximum" at the end of the *main* chain.

4.8.1 Choice of function $f(\cdot, \cdot)$

My purpose of employing the *auxiliary* chain is merely to take the *main* chain to a different part of model space and re-initialize it. While doing this, I want to make sure that the *auxiliary* chain does not break the strong interactions that have already been detected. I enforce this notion by letting the distribution of the *auxiliary* chain depend on the iteration number i of the modified algorithm. Defining n_{cycles} to be the total number of cycles of the *main* and *auxiliary* chains, I defined the function $f(c_{aux}, i)$ to be

$$f(c_{aux}, i) = \left(c_{aux} - (c_{aux} - \log(4.0)) * \left\lfloor \frac{i}{n_1 + n_2} \right\rfloor * n_{cycles} \right) * \frac{1}{\log(4.0)}, \tag{4.21}$$

where $\lfloor x \rfloor$ denotes the largest integer smaller than the real argument x. During the first run of the *auxiliary* chain, $f(c_{aux}, i)$ is defined as $f(c_{aux}, i) = c_{aux}/\log(4.0)$. Therefore, any interaction that yields a Bayes' factor of $BF_{main} = e^{c_{aux}}$, will correspond to a Bayes' factor in the *auxiliary* chain with value $BF_{aux} = 4.0$. Therefore if $e^{c_{aux}}$ corresponds to the Bayes' factor of the strongest "edge" in the *main* chain, that strongest edge will have a chance of $\frac{1}{5}$ of breaking in the first run of the *auxiliary* chain.

With increasing iteration number i of the modified algorithm, $f(c_{aux}, i)$ gets closer to 1, so the distribution of the *auxiliary* chain gets closer to the distribution of the *main* chain. This is equivalent to saying that with increasing number of alternative runs of the *main* and *auxiliary* chains, our belief in the obtained strong cliques is increasing, so we will be less inclined to change their structure.

I have only considered a form the function $f(\cdot, \cdot)$ which has a linear gradient in its second argument. Since the distribution of absolute values of Bayes' factors obtained in any run is expected to be heavily right skewed, one might also consider a form of the function $f(\cdot, \cdot)$ which is super-linear in its second argument, so that there are more runs of the *auxiliary* chain for smaller values of $f(c_{aux}, i)$ and less runs of the *auxiliary* chain for larger values of $f(c_{aux}, i)$. However, I have not explored that option, and the linear gradient form appears to be sufficient for my examples.

4.8.2 Setting simulation parameters

To obtain c_{aux}, I performed a preliminary run of the modified algorithm while artificially setting $BF_{main} = BF_{aux} = 1.0$ and recording the actual BF_{main} for all of the proposed moves corresponding to breaking and joining edges. After a sufficient number of iterations, I set c_{aux} to be 1.1 times of the maximum value of the recorded BF_{main}.

To set the length of the *main* and *auxiliary* chains, n_1 and n_2 respectively, I first ran a short chain from a randomly generated starting value and let it find the nearest "local

maximum." I then recorded the number of cliques in it. Denote this number by n_C^*. Any length of the *auxiliary* chain should be sufficient to break an edge if the Bayes' factor BF_{aux} for such a move is larger than a particular value, say 100. Since I set $P_b = 0.5$ in all my examples and the cliques are selected at random in the algorithm described in Section 4.6, I set the length of *auxiliary* chain as $n_C^* * 200$. I set the length of the *main* chain to be twice the length of the *auxiliary* chain.

The final simulation parameter that needs to be set is the number of cycles, n_{cycles}. I set this parameter so that the slope of the function $f(c_{aux}, i)$ with respect to the iteration number i is 10. This value was chosen because any interaction that yields a Bayes' factor of $BF_{main} = e^{c_{aux}-10}$ will correspond to a Bayes' factor in the *auxiliary* chain $BF_{aux} = 4.0$ in the second run of the *auxiliary* chain, and any interaction that yields a Bayes' factor of $BF_{main} = e^{c_{aux}-20}$ will correspond to a Bayes' factor in the *auxiliary* chain $BF_{aux} = 4.0$ in the third run of the *auxiliary* chain, and so on.

4.8.3 Aggregating multiple runs

There's no guarantee, however, that individual runs of the modified algorithm will yield the highest probability models. For this reason, I ran the modified algorithm multiple times and recorded the best models in individual runs. Then I counted the number of times each clique was obtained in the best models over multiple runs. The cliques that appeared with the highest frequencies were then used to build the model estimate.

4.9 Alarm network

The alarm network is shown in Figure 7. The *true* network has 37 nodes and 46 edges. Table 14 provides the variable index versus the actual names of the variables in my dataset.

I applied my method to a dataset of 10,000 observations generated from this network. I obtained the best cliques by aggregating the output from 100 runs of the modified algorithm

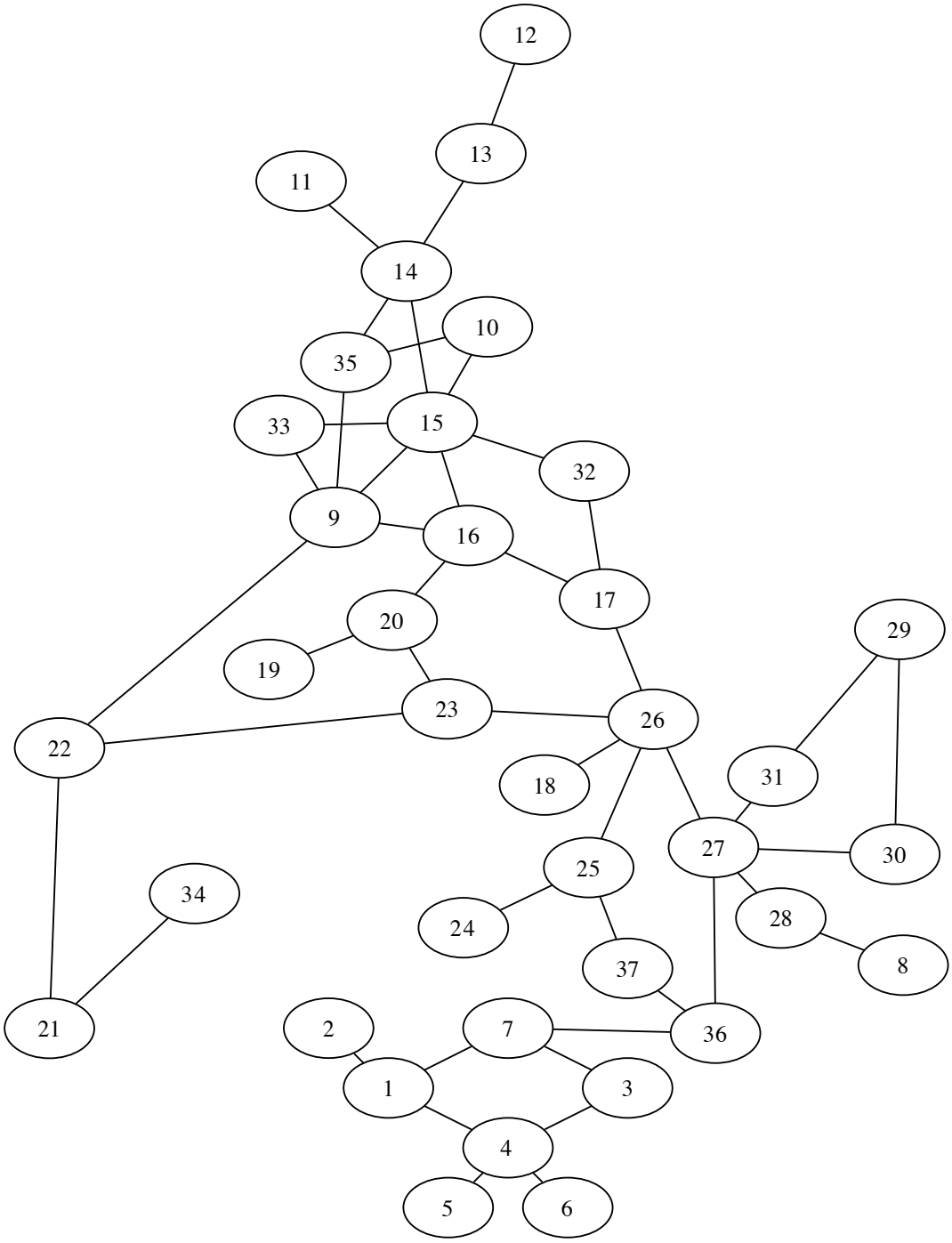

Figure 7: Alarm network

Table 14: Nodes in the alarm network

1	LVFAILURE	11	DISCONNECT	21	PULMEMBOLUS	31	HRSAT
2	HISTORY	12	MINVOLSET	22	SHUNT	32	EXPCO2
3	HYPOVOLEMIA	13	VENTMACH	23	SAO2	33	MINVOL
4	LVEDVOLUME	14	VENTTUBE	24	ANAPHYLAXIS	34	PAP
5	CVP	15	VENTLUNG	25	TPR	35	PRESS
6	PCWP	16	VENTALV	26	CATECHOL	36	CO
7	STROKEVOLUME	17	ARTCO2	27	HR	37	BP
8	ERRLOWOUTPUT	18	INSUFFANESTH	28	HRBP		
9	INTUBATION	19	FIO2	29	ERRCAUTER		
10	KINKEDTUBE	20	PVSAT	30	HREKG		

in Section 4.8. I ran the algorithm in Section 4.8 for 700 cycles and 1400 cycles. The cliques that were obtained with the highest frequency in the 100 runs are tabulated in Table 15.

The results from the modified algorithm in Section 4.8 generally appear satisfactory. In regions of the true model graph, where the true model satisfies the model constraints in Section 4.5, my method produces a network structure that agrees with the true model. Differences between the true model and the structure determined by my method in those regions occur because the data is supports those cliques more than the true model cliques. For example, when the true model cliques are $[11\ 14][13\ 14]$, my method fits the clique $[11\ 13\ 14]$. The test of hypothesis $H_1 : [11\ 14][13\ 14]$ vs. $H_2 : [11\ 13\ 14]$ yielded a LRS value of 1611.398 with 12 degrees of freedom. For the same reasons, my method fit $[7\ 27\ 36]$ where the true model cliques are $[7\ 36][27\ 36]$, $[16\ 19\ 20]$ where the true model cliques are $[16\ 20][19\ 20]$, and $[8\ 27\ 28]$ where the true model cliques are $[8\ 28][27\ 28]$. For a similar reason my method does not detect clique $[18\ 26]$, and variable 18 shows up as singleton clique. For a test of hypotheses which corresponds to the true model structure $H_1 : [9\ 22][21\ 22]$ versus $H_2 : [9\ 21\ 22]$, the Bayes' factors in favor of H_2 over H_1 were 4.9219 and 0.0085 for the local and MoM priors, respectively. As expected, my method us-

Table 15: Graphical model analysis of the alarm network data

Local ($c = 2$)					MoM ($\tau_1 = 0.348$)				
			700 cycles	1400 cycles				700 cycles	1400 cycles
11	13	14	100	100	11	13	14	100	100
1	2		100	100	1	2		100	100
12	13		100	100	12	13		100	100
1	3	4	100	100	18			100	100
1	3	7	100	100	21	22		100	100
21	34		100	100	21	34		100	100
25	36	37	100	100	25	36	37	100	100
27	29	30	100	100	27	29	30	100	100
27	29	31	100	100	27	29	31	100	100
4	5		100	100	4	5		100	100
4	6		100	100	4	6		100	100
7	27	36	100	100	7	27	36	100	100
8	27	28	100	100	8	27	28	100	100
9	15	16	97	100	16	17	32	98	99
9	15	33	96	100	9	15	16	98	99
16	17	32	95	100	9	15	33	98	99
24	25		92	94	24	25		95	100
26	27		91	94	17	26		90	94
18			84	88	26	27		88	93
17	26		81	92	1	3	4	84	91
14	16	35	63	64	1	3	7	84	91
21	22		62	63	14	16	35	58	62
10	16	35	59	59	20	22	23	58	57
20	22	23	56	62	10	16	35	55	55
16	20	22	52	62	16	20	22	52	55
16	19	20	47	38	16	19	20	48	45
19	20		46	57	20	23		42	43
9	21	22	38	38	9	22		42	43
20	23		32	32	19	20		38	45
14	33	35	29	29	14	33	35	33	35
10	33	35	25	25	10	33	35	27	33
16	20	23	12	5	1	4	7	16	9
25	26	36	9	6	3	4	7	16	9

ing the local prior detects both $[9\ 21\ 22]$ and $[21\ 22]$, whereas the MoM prior detects $[9\ 21]$ and $[21\ 22]$.

In the regions of the true model graph which involve a loop of 4 variables, my method tends to *triangulate* the loop by fitting two three-way cliques which share an edge. For example, when the true model cliques are $[1\ 4][1\ 7][3\ 7][3\ 4]$, my method fit the cliques $[1\ 3\ 4][1\ 3\ 7]$, and when the true model cliques are $[27\ 30][29\ 30][29\ 31][27\ 31]$, my method fits the cliques $[27\ 29\ 30][27\ 29\ 31]$.

Even in the regions of the true model graph which involve a more complex interaction

structure, my method fit can yield valuable information about the true model. For example, consider the (ordered) loop involving 5 variables {25,26,27,36,37}. My method fits the cliques [25 36 37] and [26 27] and the edge {27,36} as part of the clique [7 27 36]. Forming the edge {25,26} would then violate the model restrictions in Section 4.5; hence my method does not fit that edge. Similarly, consider the ordered loop involving 5 variables {16,20,23,26,17}. My method fits the cliques [20 23] and [17 26] and the edges {16,17} and {16,20}as part of the cliques [16 17 32], and [16 19 20][16 20 22], respectively. Forming the edge {23,26} would violate the model restrictions in Section 4.5; hence my method does not fit that edge. Also, consider the region involving variables {9,15,16,17,32}. This structure is formed by adding the edge {15,16} to the loop involving variables {9,16,17,32,15}. My method fits the cliques [9 15 16] and [16 17 32]. Any attempt to join the edge {15,32} will violate the model restrictions in Section 4.5, and hence that edge is not formed.

Finally, consider the region of the true model graph that is the hardest to fit. This region involves variables 9, 10, 14, 15, 33 and 35. In the true model, the edge {9,15} is included in the cliques [9 15 33]. That edge is also included in the loop {9,15,10,35}, and the interaction structure formed by adding the edge {15,16} to the loop involving variables {9,16,17,32,15}. There's another loop involving variables {10,35,14,15}, so the edge {10,35} is included in two loops involving 4 variables each. Clearly, this configuration violates many of the model restrictions in Section 4.5. My method fits the cliques [9 15 16] and [9 15 33], so the edge {9,15} cannot be included in any other clique. The variables 10, 14 and 35 form two three-way cliques along with variable 16: [10 16 35] and [14 16 35]. Since, the true model graph has a loop involving variables {10,35,14,15}, it might appear that either the model or the modified algorithm in Section 4.8 has failed in this region of the graph. However, that is not the case.

I first ran the modified algorithm in Section 4.8 with the Bayes' factor based on the local prior as in equation (2.15) with just 4 variables to find the sub-graph that fits the best

to those 4 variables. The 4 variables correspond to variables 10,14,15, and 35 of the alarm network. The best fitting model was obtained as $[10\ 15\ 35][14\ 15\ 35]$, and the sum of the posterior probabilities of all the other models is essentially 0. In the dataset that I obtained, I then tested the model $H_1 : [9][10][14][15][16][35]$ versus $H_2 : [9\ 15\ 16][10\ 15\ 35][14\ 15\ 35]$ and H_1 versus $H_3 : [9\ 15\ 16][10\ 16\ 35][14\ 16\ 35]$. The former test yielded an LRS value of 23695.67 with 108 degrees of freedom and the latter yielded an LRS value of 24368.28 with 108 degrees of freedom. Therefore, the model under H_3 is more likely than the model under H_2. Hence my method found the fit that best explains this particular realization of the data.

4.10 Implication of model inadequacy on prediction

The alarm network is a small network, yet it poses a problem for my model because the model restrictions are not congruent with the true network. However, I emphasize that my interest is in building a prediction model for a disease based on the genotype or expression data, rather than obtaining an estimate of the exact network structure. Genomic applications typically involve at least hundreds of nodes, so there is a trade-off between modeling complexity and its implementation in large datasets.

Moreover, I only need a model that can discover the edges that contain the most information about a node, given its neighbors. In this regard, it is sufficient if my method can predict individual nodes. Consider, for example, the node 26 in the alarm network. In the true network, this node is included in the loops {26,25,37,36,27} and {26,17,16,20,23}, and joins an edge with variable 18. However, my method only fit the cliques $[17\ 26]$ and $[26\ 27]$.

To test the predictive ability of my method versus the true model in predicting variable 26, I divided my data set of 10,000 observations into a *training* sample of size 8,000 and a *validation* sample of 2,000 observations. I then fit the model $[17\ 26][26\ 27]$ to the

contingency table involving the variables 17, 26 and 27, and obtained the estimated cell probabilities. I used these estimated cell probabilities to obtain the conditional probabilities for the levels of variable 26 for each individual combination of the levels of the other two variables. I then classified each observation in the *validation* sample for different threshold values of these conditional probabilities.

I then repeated this procedure to classify the observations in the *validation* sample based on the conditional probabilities obtained from the *training* sample by fitting the model $[17\ 26][16\ 17][16\ 20][20\ 23][23\ 26][26\ 27][25\ 26][25\ 37][36\ 37][27\ 36][26\ 27]$. Note that these are the cliques that belong to the loops $\{26,25,37,36,27\}$ and $\{26,17,16,20,23\}$ in the true model. I excluded the edge $\{18,26\}$ because it was not significant in the observed data. The results of this out of sample cross-validation are shown in Table 16.

Table 16: Prediction in alarm network. Cross-validation results for the prediction of node 26 in the alarm network data: The first column shows the threshold probability for out of sample classification. Columns 2 & 4 show the % of observations that could be classified for that threshold, and columns 3 & 5 show the % of misclassifications for that threshold.

	My method		True model	
Threshold %	% Classified	Misclassification rate (%)	% Classified	Misclassification rate (%)
99	65.80	0.15	79.15	0.13
95	81.85	0.55	86.95	0.29
90	81.85	0.55	92.85	1.02
80	82.85	0.66	94.85	1.32
70	90.15	2.50	98.25	2.90
60	99.95	5.35	99.05	3.23
50	100	5.40	100	3.50

As expected, the true model performs slightly better in out of sample prediction. However, the performance of my method is not far from optimal if misclassification error rates are agreed to be the evaluation criteria. Overall, my method does not perform significantly worse in predicting variable 26.

CHAPTER V

ANALYSIS OF SNP DATA

A single-nucleotide polymorphism (SNP) is a variation in the DNA sequence, a difference in a single nucleotide, A, T, C or G (see Figure 8). This can be a difference between members of a species or within a chromosome pair in the same individual. For example, consider two sequenced DNA fragments from different individuals, AAGCCTAG to AAGCTTAG. The two sequences contain a difference in a single nucleotide. We say that there are two alleles, C and T. Most common SNPs have only two alleles (description borrowed from `http://www.wikipedia.org`).

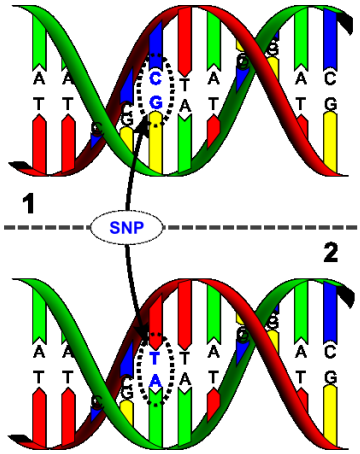

Figure 8: Single-nucleotide polymorphism. Source: `http://www.wikipedia.org`

The term *minor allele frequency* refers to the lowest allele frequency at a locus that is observed in a particular population. This is simply the lesser of the two allele frequencies for SNPs at that locus. Variations between human populations mean that a SNP allele that is common in one ethnic group may be much rarer in another.

Genetic variations in humans can affect response to pathogens, chemicals, drugs, vaccines, and even radioactive agents, thus affecting disease progression, cure and prevention. SNPs are also thought to be a key factor in personalized medicine. An important challenge in genome-wide disease association studies is to understand the interconnections between a network of the genes and their products. Hence, there is a growing need for algorithms and models for reducing biological and statistical redundancy from hundreds of thousands of SNPs. Dealing with many dependent association tests is another important statistical and computational issue. The difficulty in a SNP association study is increased by the nature of complex diseases in which the contribution of single gene is small to moderate. Instead complex diseases are thought to result from gene-gene and gene-environment interactions.

In this chapter, I analyze the Single-Nucleotide Polymorphism data provided by the Amos lab at the *University of Texas M. D. Anderson Cancer Center* (Taylor et al. 2007) using the methodology outlined in the previous chapter. The data includes biallelic measurements taken on 1260 Rheumatoid Arthritis patients and 908 normal subjects. About 550,000 measurements were taken on each subject at loci on 23 chromosome pairs. Rheumatoid arthritis (RA) is a chronic, systemic inflammatory disorder that attacks the joints. The disease causes an inflammation of the synovial membrane that leads to the destruction of the articular cartilage and stiffness in the joints.

The aim of my analysis of RA SNP data is to build a predictive model for the disease based on subject genotypes. My approach is two-staged: first I use a graphical model to find the SNPs that are likely to be the best predictors of the disease, and then I use those to build a prediction model. This approach performs well in cross-validation studies.

5.1 Alternative approaches

In the context of genome-wide disease association studies, various approaches have been proposed to develop disease-genotype models. These include Linkage Disequilibrium (LD)

based SNP selection and supervised SNP selection. Zhang and Jin (2003) proposed a two-stage approach to this problem: first, they identified haplotype blocks and then found the tagSNPs that best distinguish the haplotypes within a haplotype block. In related approaches, Anderson and Novembre (2003) and Mannila et al. (2003) found haplotype block boundaries using minimum description length methods.

These methods, however, tend to be less useful if there is significant diversity in the genotypes of the sampled population. Owing to recombination events over several generations, i.e., if the sampled population is intermixed, there probably exist no haplotype blocks of significant meaning.

Neale and Sham (2004) used a sliding window approach. They propose computing a test statistic by combining p-values from multiple independent tests. Let p_i be the p-value of association between SNP i and the disease, and m be the number of SNPs in the sliding window. Then their statistic is computed as $-2\sum_{i=1}^{m}\log(p_i)$, and has a nominal chi-square distribution with $2m$ degrees of freedom. Their method incorporates the ordering of SNPs on the chromosome. Furthermore, results across adjacent windows are merged to detect chromosome regions with significant associations. Alternative test based filter approaches have also been proposed in Hoh and Ott (2000), Levin et al. (2005), Sun et al. (2006), Song and Elston (2006) and Cheng et al. (2005), amongst many others. One major limitation of these test-based filter approaches is that they may produce many highly correlated SNPs or genes with redundant information, that are of little value for prediction.

Approaches based on logistic regression have been proposed by, for example, Durrant et al. (2004). However, logistic regression based methods are applicable only when analyzing a small subset of SNPs. To see why, suppose that the disease (variable 1) has strong marginal associations with both SNPs 2 and 3. Then a logistic regression can pick model $[1\ 2]\,[1\ 3]$ or $[1\ 2\ 3]$. Thus, the logistic model will select both SNPs 2 and 3 as predictors of the disease. However, the true disease model could be $[1\ 2]\,[2\ 3]$. This is especially true if

SNPs 2 and 3 are close to each other on the chromosome. Since in the true disease model the disease is associated with SNP 3 only through SNP 2, the former is not required to predict the disease and in fact, including it will hurt prediction. This necessitates the use of models that explicitly model conditional independence relationships.

At this stage, I note that the model class that I consider is fairly rich and should find applications to many other types of datasets. It is also worth pointing out the important differences of my model when compared to Verzilli, Stallard, and Whittaker (2006) and Zhang and Liu (2007). Verzilli, Stallard, and Whittaker (2006) use a decomposable graphical model to search for multi-locus patterns of association around a causative site. Their model space is restricted to form cliques of neighboring SNPs and their model incorporates a prior that uses the location information of each SNP on the genome. Such a model largely ignores epistatic information, i.e., the interaction of SNPs on different chromosomes. Also, although it is not explicitly mentioned in their paper, they do not allow cliques to share edges.

Zhang and Liu (2007) use an epistasis model in their method. They allow SNPs to be associated with disease either marginally or through SNP interactions. However, all SNPs that affect the phenotype through interactions are restricted to interact with each other. To see why this is problematic, consider a true disease model that corresponds to Figure 9a. The methodology of Zhang and Liu (2007) searches amongst models of the type shown in Figures 9b, c, d and e, while only model Figures 9e covers all the interactions in the true model. However, the model in Figure 9e contains many more interactions, and involves many more degrees of freedom than it needs to, and so its predictive performance must necessarily suffer. Also, since model in Figure 9e uses many more degrees of freedom than the true disease model in Figure 9a, it's harder to detect such a model from data sets of moderate size. The algorithm used by Verzilli, Stallard, and Whittaker (2006) forms bigger cliques by "merging" smaller cliques, and so is also prone to such problems.

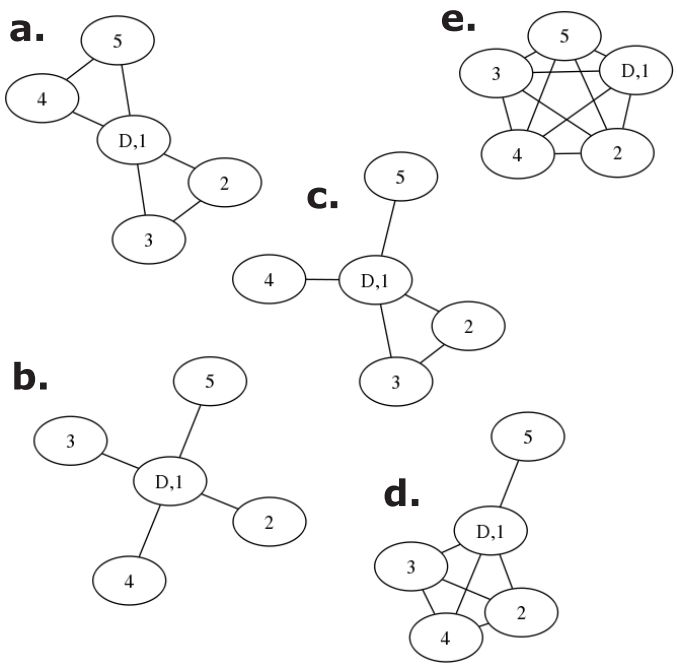

Figure 9: Models in Zhang and Liu (2007)

Zhang and Liu (2007) do not impose a restriction on the size of the cliques, and Verzilli, Stallard, and Whittaker (2006) have allowed cliques that can include up to 8 vertices. In my opinion, my restriction of the clique size to 4, along with edge sharing is sufficient for most practical applications. I now describe the application of my method to SNP data.

5.2 Outline of SNP analysis

5.2.1 Pre-selection of SNPs

My graphical model method can easily handle $\simeq$ 100 nodes (see Subsection 4.7.1). Nonetheless, as the number of nodes increases, the model space can become enormous. I, therefore need a pre-selection method to choose the SNPs that will likely be the best predictors of the disease node amongst the $\simeq$ 550,000 SNPs. I propose pre-selecting SNPs based on the ideas of a graphical model. A SNP whose strong association with the disease node is

explained by another SNP is will not be a good predictor of the disease. The pre-selection method is applied to individual chromosomes and is explained in the following steps.

1. Rank each SNP on the chromosome in the order of its association with the disease node. I measure the strength of the association based on the Bayes' factor based on the iMoM prior. To illustrate, suppose variable 1 represents the disease node, and S denotes a particular SNP node. Then the strength of the association of the two nodes is obtained as the Bayes' factor of the test $H_1 : [1][S]$ versus $H_2 : [1\ S]$. The SNPs with strongest association with the disease are assigned the lowest rank. Let all the SNPs on the chromosome be denoted in that rank order as $S_1, S_2, \ldots, S_n$. Let I denote the set $I = \{i : 1 \leq i \leq n\}$. Let $i^* = 1$. Then iterate between the following steps.

2. For all SNP indices j, such that $j \in I, j > i^*$, test the models $H_1 : [1\ S_{i^*}][S_{i^*}\ S_j]$, and $H_2 : [1\ S_j][S_1\ S_j]$ versus the saturated model involving the three nodes, $H_3 : [1\ S_{i^*}\ S_j]$.

   (a) If the model H_1 fits well to the data and model H_2 does not, redefine set $I = I \bigcap \{j\}^c$.

   (b) If the model H_2 fits well to the data and model H_1 does not, redefine set $I = I \bigcap \{i^*\}^c$. Skip to step (3).

   (c) If both models H_1 and H_2 fit well to the data, redefine set $I = I \bigcap \{j\}^c$.

   I gauge the fit of the models H_1 and H_2 by computing their Bayes' factors when tested against the saturated model involving a three-way interaction of the disease node and the two concerned SNPs. I say H_1 fits well to the data if $BF_I(3|1) < 0.01$, and H_2 fits well to the data if $BF_I(3|2) < 0.01$.

3. Let i_1^* be the SNP index such that $i_1^* = min(I \cap \{i : i > i^*\})$. Reset i^* to i_1^*.

Here, A^c represents the complement of set A. Then, I only use the SNP indices in set I for the next step of the analysis. Rephrasing step (2) above, if the model H_1 fits well to the data and model H_2 does not, we do not include S_j in the next step of the analysis. Similarly, if the model H_2 fits well to the data and model H_1 does not, we do not include S_{i^*} in the next step of the analysis.

As mentioned above, I apply my pre-selection method to one chromosome at a time. This is based on the belief that two SNPs are more likely to provide similar information about the disease if they are on the same chromosome than when they are on different chromosomes. Hence, if SNP S_1 on chromosome 6 seems to be explaining away the disease association of SNP S_2, and the two SNPs are on different chromosomes, I retain both for next step of the analysis.

5.2.2 Graphical model analysis

As a second step of the analysis, I apply the method described in Section 4.8 to the set of nodes consisting of the disease node and the SNPs that were pre-selected in Subsection 5.2.1. All the cliques containing the disease node are recorded. I then build a predictive model for the disease from these cliques.

5.2.3 Prediction model

To build a predictive model, I first rank the cliques containing the disease in order of their predictive ability. I gauge the predictive ability of the cliques containing the disease node by the strength of associations of the disease node and the other nodes in that clique. For example, let us say I obtain a clique of size 2 and a clique of size 3 containing the disease node: $[1\ S_1]$ and $[1\ S_2\ S_3]$. I gauge the predictive ability of the clique of size 2 from the Bayes' factor of the test of $H_2 : [1\ S_1]$ versus $H_1 : [1][S_1]$, and the predictive ability of the clique of size 3 from the Bayes' factor of the test of $H_2 : [1\ S_2\ S_3]$ versus $H_1 : [1][S_2\ S_3]$.

5.3 Results of cross-validation study

For my analysis, I converted the missing values of SNPs to another category of the categorical variable corresponding to that locus. I set aside 131 patients and 89 normal subjects as the *validation* sample. The *testing* sample hence consisted of 1129 patients and 819 normal subjects.

Table 17: Predictive cliques for the disease node in the SNP data. I used different parameter settings of the local and MoM priors. The disease node is labeled as "1." For each prior and parameter setting, the second column contains the number of times that clique was formed in 100 runs, and the fourth column contains the logarithm of the metric of predictive ability described previously.

Local ($c = 1$)			Local ($c = 2$)			Local ($c = 3$)		
1,28,29	48	250.63	1,28,29	51	245.85	1,27	36	243.12
1,27,28	78	248.18	1,27,28	59	244.10	1,28,29	44	243.04
1,58,112	19	202.50	1,27	21	243.71	1,27,28	48	241.69
1,4,58	21	192.21	1,4,58	26	187.08	1,4,58	37	184.06
1,7,58	20	181.91	1,58,106	23	184.10	1,58,106	33	181.08
1,13,91	33	171.92	1,14,58	23	182.75	1,14,58	28	179.73
1,64,91	31	145.35	1,13,91	50	166.78	1,13,91	51	163.76
1,88,92	28	137.46	1,64,91	45	140.20	1,91,109	24	144.57
1,45,88	26	133.03	1,88,92	44	132.31	1,64,91	34	137.18
1,4,106	19	132.98	1,45,88	20	127.88	1,88,92	61	129.29
MoM ($\tau_1 = 0.174$)			MoM ($\tau_1 = 0.348$)			MoM ($\tau_1 = 0.696$)		
1,27	57	243.76	1,27	87	242.15	1,27	77	240.48
1,28,29	32	239.81	1,4,58	73	174.23	1,58	98	171.33
1,27,28	30	239.45	1,91,109	77	134.57	1,91,109	79	128.69
1,4,58	67	180.10	1,64,91	50	127.14	1,88	94	121.04
1,58,106	53	177.10	1,88	95	122.74	1,13	99	112.84
1,91,109	41	140.44	1,13	97	114.54	1,112	98	109.41
1,64,91	55	133.01	1,112	96	111.11	1,14	64	94.95
1,88,92	64	125.09	1,14	83	96.65	1,98	95	93.20
1,13	99	116.21	1,98	90	94.90	1,42	95	92.06
1,112	90	112.78	1,42	94	93.76	1,7	76	91.99

The pre-selection scheme described above yielded 127 SNPs. Including the disease node, I therefore entered 128 variables into the graphical model analysis, and obtained the most important cliques using the algorithm described in Section 4.8. I then selected the best

150 cliques in terms of the number of times they were formed in 100 runs of the algorithm described in Section 4.8. Amongst those 150 cliques, I ranked the cliques that contained the disease node according to the metric of predictive ability as defined above.

I then used the 10 cliques with best predictive ability to obtain conditional probabilities for the phenotype status given combinations of the genotypes corresponding to those cliques. Table 17 shows the cliques with best predictive ability obtained for different prior and parameter settings. Table 18 contains the names of the SNPs that appeared in the cliques in Table 17.

Table 18: Names of some important SNPs

SNP Index	SNP Name
4	rs9442372
7	rs788160
13	rs9812051
14	rs6437394
27	rs2395175
28	rs660895
29	rs9275224
42	rs6458368
45	rs213212
58	rs3132332
64	rs10501396
88	rs7194895
91	rs2567494
92	rs4267379
98	rs4534931
106	rs2238663
109	rs503084
112	rs6063641

Each observation in the *validation* sample was then classified for different threshold values of these conditional probabilities. The results of this out of sample cross-validation are provided in Table 19. If there was a lower order interaction and higher order interaction containing the lower order interaction in the list of the 10 cliques with the highest predictive ability, I used only to higher order interaction to compute the classification probabilities.

Table 19: Cross-validation results for SNP data. The left-most column shows the threshold probability for out of sample classification. For each prior and parameter setting, the left column shows the percentage of subjects that could be classified with these probabilities, and the right column shows the misclassification error rate for those individuals.

Threshold %	Local ($c = 1$)		Local ($c = 2$)		Local ($c = 3$)	
99	33.18	2.74	30.91	2.94	32.73	2.78
95	54.09	4.20	50.45	5.41	49.09	4.63
90	65.00	8.39	60.45	7.52	59.54	6.87
80	78.64	13.29	77.27	12.35	73.18	10.56
70	89.09	18.37	84.55	15.05	86.36	14.74
60	93.64	20.87	91.36	16.92	95.00	18.66
50	100	21.82	100	19.09	100	21.82
Threshold %	MoM ($\tau_1 = 0.174$)		MoM($\tau_1 = 0.348$)		MoM($\tau_1 = 0.696$)	
99	34.54	3.94	39.09	3.49	40.00	2.27
95	52.73	6.89	60.00	10.61	60.00	8.33
90	65.91	9.65	68.64	10.60	70.45	12.26
80	76.82	13.02	81.36	16.20	79.55	14.29
70	84.54	16.13	87.73	17.62	89.09	19.39
60	92.73	19.61	95.45	19.05	95.00	21.05
50	100	21.36	100	21.36	100	22.73

CHAPTER VI

SUMMARY AND DISCUSSION

This chapter summarizes the work presented in this dissertation. I have extended the work of Johnson (2005) and used Bayes' factors in model selection. Bayes' factors computed from the LRS have several advantages. The theoretical framework for defining Bayes' factors is well established. These Bayes' factors require setting just one prior parameter and standard criteria for setting this value are available. Furthermore, a model selection methodology based on Bayes' factors can be extended directly to non-linear models.

Expressions for these test-based Bayes' factors are simple. Bayes' factors can be extended to model selection algorithms which involve simultaneous comparisons of a large number of models. The clustering method and graphical model method based on Bayes' factors computed from LRS performs well in both appropriate simulated and real examples.

The clustering method presented in this dissertation can be easily extended to thousands of genes. This method is appropriate for gene expression data in which one expects each gene to participate in interactions with only a few other genes. The graphical model method is more appropriate for SNP data where one expects a more complex association structure between the SNPs and the disease. For such data, clustering-type methods tend to propose very big models which are not practically feasible or supported by most model selection criteria.

Although it is currently not computationally possible to infer a network of one-half million SNPs, my method can be used to infer the local network structure around individual nodes (SNP or disease). Moreover, if interest lies in predicting disease status, my pre-selection method combined with my graphical model method, can be used to obtain a good set of predictors. The results of a cross-validation study confirm this assertion.

REFERENCES

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary data and polychotomous data response data. *Journal of the American Statistical Association* **88**, 669–679.

Anderson, E. and Novembre, J. (2003). Finding haplotype block boundaries by using the minimum-description-length principle. *American Journal of Human Genetics* **73**, 336–354.

Bertucci, F., Houlgatte, R., Benziane, A., Granjeaud, S., Adelaide, J., Tagett, R., Loriod, B., Jacquemier, J., Viens, P., Jordan, B., Birnbaum, D., and Nguyen, C. (2000). Gene expression profiling of primary breast carcinomas using arrays of candidate genes. *Human Molecular Genetics* **9**, 2981–2991.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B* **36**, 192–236.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B* **25**, 220–233.

Birch, M. W. (1964). The detection of partial association. *Journal of the Royal Statistical Society, Series B* **26**, 313–324.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Carvalho, C. M., Massam, H., and West, M. (2007). Simulation of hyper-inverse wishart distributions in graphical models. *Biometrika* **94**, 647–659.

Cerny, V. (1985). A thermodynamical approach to the travelling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications* **45**, 41–51.

Chen, W., Ghosh, D., Raghunathan, T., and Kardia, S. (2004) *A Bayesian Method for Finding Interactions in Genomic Studies*, Technical Report, The University of Michigan Department of Biostatistics Working Paper Series. Available at: **http://www.bepress.com/umichbiostat/paper48**.

Cheng, R., Ma, J., Elston, R. C., and Li, M. D. (2005). Fine mapping functional sites or regions from case-control data using haplotypes of multiple linked SNPs. *Annals of Human Genetics* **69**, 102–112.

Chien, W., Kumagai, T., Miller, C. W., Desmond, J. C., Frank, J. M., Said, J. W., and Koeffler, H. P. (2004). Cyr61 suppresses growth of human endometrial cancer cells. *Journal of Biological Chemistry* **79**, 53087–53096.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression (2nd Ed.)*. New York: Springer-Verlag.

Cohen, J. (1992). A power primer. *Psychological Bulletin* **112**, 155–159.

Cooper, C. S. (2001). Applications of microarray technology in breast cancer research. *Breast Cancer Research* **3**, 158–175.

Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics* **8**, 522–539.

Davidson, R. R. and Lever, W. E. (1970). The limiting distribution of the likelihood ratio statistic under a class of local alternative. *Sankhya, Series A* **32**, 209–224.

Dobra, A., Han, C., Jones, B., Nevins, J., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis* **90**, 196–212.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B* **57**, 45–97.

Durrant, C., Zondervan, K. T., Cardon, L. R., Hunt, S., Deloukas, P., and Morris, A. P. (2004). Linkage Disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *American Journal of Human Genetics* **75**, 35–43.

Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Gong, J., Ammanamanchi, S., Ko, T. C., and Brattain, M. G. (2003). Transforming growth factor beta 1 increases the stability of p21/WAF1/CIP1 protein and inhibits CDK2 kinase activity in human colon carcinoma FET cells. *Cancer Research* **63**, 3340–3346.

Goodman, L. A. (1970). Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables. *Journal of the American Statistical Association* **66**, 339–344.

Goodman, L. A. (1971). The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association* **65**, 226–256.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. Chicago: University of Chicago Press.

Harkin, D. P., Bean, J. M., Miklos, D., Song, Y. H., Truong, V. B., Englert, C., Christians, F. C., Ellisen, L. W., Maheswaran, S., Oliner, J. D., and Haber, D. A. (1999). Induction of GADD45 and JNK/SAPK-dependent apoptosis following inducible expression of BRCA1. *Cell* **28**, 575–586.

Hodges, J. S. (1987). Uncertainty, policy analysis and statistics. *Statistical Science* **2**, 259–291.

Hoh, J. and Ott, J. (2000). Scan statistics to scan markers for susceptibility genes. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 9615–9617.

Hu, J. and Johnson, V. E. (2009). Bayesian model selection using test statistics. *Journal of the Royal Statistical Society, Series B* **71**, 143–158.

Hu, J., Joshi, A., and Johnson, V. E. (2009). Log-linear models for gene association. *Journal of the American Statistical Association* **104**, 597–607.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)* **4**, 249–264.

Jefferys, W. H. and Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist* **80**, 64–72. Preprint available at: **`http://www.stat.duke.edu/~berger/papers/ockham.html`**.

Johnson, V. E. (2005). Bayes' factors based on test statistics. *Journal of the Royal Statistical Society, Series B* **67**, 689–701.

Johnson, V. E. and Rossell, D. (2008). Properties of Bayes' factors based on test statistics. *Scandinavian Journal of Statistics* **35**, 354–368.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society, Series B*, in press. Available at: **`http://www.bepress.com/mdandersonbiostat/paper42`**.

Jones, B., Carvalho, M., C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science* **20**, 388–400.

Jones, B. and West, M. (2005). Covariance decomposition in undirected gaussian graphical models. *Biometrika* **92**, 779–786.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.

Lacroix, M. and Leclercq, G. (2004). About GATA3, HNF3A, and XBP1: Three genes co-expressed with the Oestrogen receptor-alpha gene (ESR1) in breast cancer. *Molecular and Cellular Endocrinology* **219**, 1–7.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.

Leamer, E. E. (1978). *Specification Searches*. New York: Wiley.

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

Levin, A. M., Ghosh, D., Cho, K. R., and Kardia, S. L. (2005). A model-based scan statistics for identifying extreme chromosomal regions of gene expression in human tumors. *Bioinformatics* **21**, 2867–2874.

Lombaerts, M., van Wezel, T., Philippo, K., Dierssen, J. W., Zimmerman, R. M., Oosting, J., van Eijk, R., Eilers, P. H., van de Water, B., Cornelisse, C. J., and Cleton-Jansen, A. M. (2006). E-cadherin transcriptional downregulation by promoter methylation but not mutation is related to epithelial-tomesenchymal transition in breast cancer cell lines. *British Journal of Cancer* **94**, 661–671.

Luo, X., Ding, L., Xu, J., and Chegini, N. (2005). Gene expression profiling of leiomyoma and myometrial smooth muscle cells in response to transforming growth factor-beta. *Endocrinology* **146**, 1097–1118.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89**, 1535–1546.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63**, 215–232.

Mannila, H., Koivisto, M., Perola, M., Varilo, T., Hennah, W., Ekelund, J., Lukk, M., Peltonen, L., and Ukkonen, E. (2003). Minimum description length block finder, a method to identify haplotype blocks and to compare the strength of block boundaries. *American Journal of Human Genetics* **73**, 86–94.

Moulton, B. R. (1991). A Bayesian approach to regression selection and estimation with application to a price index for radio services. *Journal of Econometrics* **49**, 169–193.

Naderi, A. and Hughes-Davies, L. (2008). A functionally significant cross-talk between Androgen receptor and ErbB2 pathways in Estrogen receptor negative breast cancer. *Neoplasia* **10**, 542–548.

Neale, B. and Sham, P. (2004). The future of association studies: Gene-based analysis and replication. *American Journal of Human Genetics* **75**, 353–362.

Panet-Raymond, V., Gottlieb, B., Beitel, L. K., Pinsky, L., and Trifiro, M. A. (2000). Interactions between androgen and estrogen receptors and the effects on their transactivational properties. *Molecular and Cellular Endocrinology* **25**, 139–150.

Potamias, G., Koumakis, L., and Moustakis, V. (2004). Gene selection via discretized gene-expression profiles and greedy feature-elimination. In G. A. Vouros and T. Panayiotopoulos (eds.), *Methods and Applications of Artificial Intelligence 3025*, Lecture Notes in Computer Science. Berlin: Springer.

Quenneville, L. A., Trotter, M. J., Maeda, T., and Tron, V. A. (2002). P53-dependent regulation of heat shock protein 72. *British Journal of Dermatology* **146**, 786–791.

Raftery, A. E. (1996). Approximate Bayes' factors and accounting for model uncertainty in generalized linear models. *Biometrika* **83**, 251–266.

Ragione, F. D., Cucciolla, V., Criniti, V., Indaco, S., Borriello, A., and Zappia, V. (2003). p21Cip1 gene expression is modulated by Egr1: A novel regulatory mechanism involved in the resveratrol antiproliferative effect. *Journal of Biological Chemistry* **278**, 23360–23368.

Resnick, S. I. (1999). *A Probability Path*. Boston: Birkhäuser.

Sanga, S., Broom, B. M., Cristini, V., and Edgerton, M. E. (2009). Gene expression meta-analysis supports existence of molecular apocrine breast cancer with a role for androgen receptor and implies interactions with ErbB family. *BMC Medical Genomics* **2:**59, doi:10.1186/1755–8794–2–59. Available at:

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2753593/pdf/1755--8794--2--59.pdf.

Schafer, J. and Strimmer, K. (2005). An empirical Bayes' approach to inferring large-scale gene association networks. *Bioinformatics (Oxford, England)* **21**, 754–764.

Song, K. and Elston, R. C. (2006). A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. *Statistics in Medicine* **25**, 105–126.

Spurdle, A. B., Chang, J., Byrnes, G. B., Chen, X. Dite, G. S., McCredie, M. R. E., Giles, G. G., Southey, M. C., Chenevix-Trench, G., and Hopper, J. L. (2007). A systematic approach to analysing genegene interactions: Polymorphisms at the microsomal Epoxide Hydrolase EPHX and Glutathione S-transferase GSTM1, GSTT1, and GSTP1 loci and breast cancer risk. *Cancer Epidemiology, Biomarkers & Prevention* **16**, 769–774.

Sun, Y., Levin, A., Boerwinkle, E., Robertson, H., and Kardia, S. L. R. (2006). A scan statistic for identifying chromosomal patterns of SNP association. *Genetic Epidemiology* **30**, 627–635.

Tavare, S. and Ewens, W. J. (1997). The Ewens' sampling formula. In N. L. Johnson and N. Balakrishanan (eds.), *Multivariate Discrete Distributions*. New York: Wiley.

Taylor, K. E., Chen, W., Amos, C. I., and Criswell, L. A. (2007). Genome-wide single-nucleotide polymorphism linkage analyses of quantitative rheumatoid arthritis phenotypes in Caucasian NARAC families. In *Genetic Analysis Workshop 15*, BMC Proceedings. 11-15 November 2006, St. Pete Beach, FL. Available at: **http://www.biomedcentral.com/1753-6561/1/S1/S105**.

Verzilli, C. J., Stallard, N., and Whittaker, J. C. (2006). Bayesian graphical models for genomewide association studies. *American Journal of Human Genetics* **79**, 100–112.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A. J., Marks, J. R., and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 11462–11467.

Weston, B. S., Wahab, N. A., and M., M. R. (2003). CTGF mediates TGF-beta-induced fibronectin matrix deposition by upregulating active alpha5beta1 integrin in human mesangial cells. *Journal of the American Society of Nephrology* **14**, 601–610.

Wu, X., Yong, Y., and Kalpathi, R. S. (2003). Interactive analysis of gene interactions using graphical gaussian model. In *3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics*, pp. 63–69. Troy, NY: Rensselaer Polytechnic Institute.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner (eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pp. 233–243. North-Holland/Elsevier.

Zhang, K. and Jin, L. (2003). HaploBlockFinder: Haplotype block analysis. *Bioinformatics* **19**, 1300–1301.

Zhang, Y. and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics* **39**, 1167–1173.

APPENDIX A

DERIVATION OF BAYES' FACTORS

Let Z_n denote the LRS. Since it is hard to integrate a non-central χ^2 density, I employ a d-dimensional column vector of random variables with a multivariate normal distribution. Let $\boldsymbol{Y}$ denote a random vector: $Z_n \xrightarrow{\mathcal{D}} Z = \boldsymbol{Y}'\boldsymbol{Y}$. Since the interest lies in the asymptotic distribution of Z_n, it is sufficient to obtain a result for the distribution of Z. I obtain the marginal (with respect to the parameters) density of Z, denoted as f_Z in the following way.

1. I first integrate the density of $\boldsymbol{Y}$ with respect to the alternative prior densities in equation (2.16) and equation (2.18).
2. Then transforming $\boldsymbol{Y}$ to $\boldsymbol{W} = \boldsymbol{Y}^2 = \{Y_1^2, \ldots, Y_d^2\}$, I obtain the density of $\boldsymbol{W}$.
3. Finally, I use the convolution formula to obtain a density for Z, denoted as h_Z.

Then the density h_Z can be shown to be the same as the density f_Z by application of Fubinis' theorem (Resnick 1999) twice: first to interchange the order of integration and the convolution and then to interchange the order of the integration and the transformation. With the introduction of the random vector $\boldsymbol{Y}$, the above test of hypotheses in equation (2.12) vs. equation (2.13) could be seen as asymptotically equivalent to the following test

$$Y \sim N_d\left[\bar{\boldsymbol{C}}_{11}^{\frac{1}{2}}\boldsymbol{\delta}^{\boldsymbol{Y}}, \boldsymbol{I}_d\right] \tag{A.1}$$

$$H_1' : \boldsymbol{\delta}^{\boldsymbol{Y}} = 0 \tag{A.2}$$

against the alternative

$$H_2' : \boldsymbol{\delta}^{\boldsymbol{Y}} = \boldsymbol{\delta} \tag{A.3}$$

Here, $\bar{\boldsymbol{C}}_{11}^{\frac{1}{2}}$ denote a symmetric matrix so that $\bar{\boldsymbol{C}}_{11}^{\frac{1}{2}}\bar{\boldsymbol{C}}_{11}^{\frac{1}{2}} = \bar{\boldsymbol{C}}_{11}$. This matrix is guaranteed to exist because the matrix $\bar{\boldsymbol{C}}_{11}$ is symmetric and positive-definite. The introduction of the random vector $\boldsymbol{Y}$ enables me to use the results from the linear model case (Johnson and Rossell 2010) and extend them directly to calculate the Bayes' factor based on likelihood ratio statistic in the following way. I first apply a non-local alternative prior to the vector $\boldsymbol{\delta}$ to calculate the marginal distribution of $\boldsymbol{Y}$ under the alternative. Then I compute the marginal distribution of Z under the null and the alternative. Knowing the marginal null and alternative distributions, the Bayes' factor could hence be computed. Let's for illustration, consider the case in which the default MoM prior is applied to $\boldsymbol{\delta}$. We will later show that the result is readily obtained for default iMoM priors as well.

Let $\bar{\boldsymbol{C}}_{11}^{-\frac{1}{2}} = \left(\bar{\boldsymbol{C}}_{11}^{\frac{1}{2}}\right)^{-1}$. To use the results from the derivations for the linear model case (Johnson and Rossell 2010), we substitute $\bar{\boldsymbol{C}}_{11}^{\frac{1}{2}}$ for $\boldsymbol{X}$ and $\boldsymbol{\delta}$ for $\boldsymbol{\theta}_1$. The following subsitutions are obtained as consequence: $\bar{\boldsymbol{C}}_{11}^{-1}$ for $\boldsymbol{\Sigma}$, $\hat{\boldsymbol{\delta}} = \bar{\boldsymbol{C}}_{11}^{-\frac{1}{2}}\boldsymbol{Y}$ for $\hat{\boldsymbol{\theta}}_1$, and $\boldsymbol{Y}'\boldsymbol{Y}$ for $\hat{\boldsymbol{\theta}}_1'\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\theta}}_1$. Therefore, the likelihood of the pseudo-data under the alternative is obtained as follows:

$$\begin{aligned} f\left(\boldsymbol{Y}|\boldsymbol{\delta}\right) &= \frac{1}{(2\pi)^{\frac{d}{2}}}\exp\left[-\frac{1}{2}(\boldsymbol{Y}-\bar{\boldsymbol{C}}_{11}^{\frac{1}{2}}\boldsymbol{\delta})'(\boldsymbol{Y}-\bar{\boldsymbol{C}}_{11}^{\frac{1}{2}}\boldsymbol{\delta})\right] \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}}\exp\left[-\frac{1}{2}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})'\bar{\boldsymbol{C}}_{11}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})\right] \\ &= \frac{f(\hat{\boldsymbol{\delta}}|\boldsymbol{\delta})}{\left|\bar{\boldsymbol{C}}_{11}\right|^{\frac{1}{2}}} \end{aligned} \tag{A.4}$$

Given the prior in (2.16), the marginal density of $\boldsymbol{Y}$ under the alternative is obtained as:

$$
\begin{aligned}
f_2(\boldsymbol{Y}) &= \int f(\boldsymbol{Y}|\boldsymbol{\delta})\,\pi_M(\boldsymbol{\delta})\,d\boldsymbol{\delta} \\
&= \frac{\int f\left(\hat{\boldsymbol{\delta}}|\boldsymbol{\delta}\right)\pi_M(\boldsymbol{\delta})\,d\boldsymbol{\delta}}{\left|\bar{\boldsymbol{C}}_{11}\right|^{\frac{1}{2}}} \qquad \text{(A.5)} \\
&= \frac{\mu_k}{\prod_{i=0}^{k-1}(d+2i)}\frac{1}{(1+n\tau)^{k+\frac{d}{2}}}\frac{1}{(2\pi)^{\frac{d}{2}}}\exp\{-\frac{\boldsymbol{Y}'\boldsymbol{Y}}{2(n\tau+1)}\} \qquad \text{(A.6)}
\end{aligned}
$$

As the numerator of the left hand side of (A.5) above has been derived as part of computations for the linear model result (Johnson and Rossell 2010), the right hand side of (A.6) can be directly obtained from it. Also, μ_k is the k^{th} raw moment of the $\chi^2_d(\lambda)$ distribution where $\lambda = \frac{n\tau}{(1+n\tau)}\boldsymbol{Y}'\boldsymbol{Y}$.

To get the Bayes' factor, we need to determine the distributions of $Z = \boldsymbol{Y}'\boldsymbol{Y}$ under the null and alternative hypothesis. Noting that both the null and alternative distributions depend on $\boldsymbol{Y}$ through terms involving $\boldsymbol{Y}'\boldsymbol{Y}$, the Bayes' factor could be derived in a straightforward fashion. Formally speaking, let $\boldsymbol{Y} \sim g(\boldsymbol{Y}'\boldsymbol{Y})$. Then transforming $\boldsymbol{Y}$ to $\boldsymbol{W} = \boldsymbol{Y}^2 = \{Y_1^2, \ldots, Y_d^2\}$, we obtain the density of $\boldsymbol{W}$ as:

$$
f_{\boldsymbol{W}}(\boldsymbol{W}) = \begin{cases} \left[\prod_i \frac{1}{\sqrt{W_i}}\right] g(\boldsymbol{W}'\underline{\mathbf{1}}), & W_i \geq 0 \quad \forall i \\ 0, & \text{otherwise} \end{cases} \qquad \text{(A.7)}
$$

Here, $\underline{\mathbf{1}}$ is a column vector of ones. Let set A_z be defined as $A_z \equiv \{f_{\boldsymbol{W}}(\boldsymbol{W}) > 0\} \bigcap \{\boldsymbol{W}'\underline{\mathbf{1}} = z\}$. Now, transforming from $\boldsymbol{W}$ to $\boldsymbol{W}^* \equiv \{\boldsymbol{W}'\underline{\mathbf{1}}, W_2, \ldots, W_d\}$, the determinant of the Jacobian of the monotonic transformation is 1. The distribution of $Z = \boldsymbol{W}'\underline{\mathbf{1}}$ could thus be obtained in a straightforward fashion as below:

$$\begin{aligned} f_Z(z) &= \int f_{\boldsymbol{W}^*}(z, W_2, \ldots, W_d) \quad \mathrm{d}W_2 \ldots \mathrm{d}W_d \\ &= \int_{A_z} f_{\boldsymbol{W}}(\boldsymbol{W})\mathrm{d}\boldsymbol{W} = g(z) \int_{A_z} \left[\prod_i \frac{1}{\sqrt{W_i}} \right] \mathrm{d}\boldsymbol{W} \end{aligned} \tag{A.8}$$

The result in (A.8) leads to the following interesting observations:

1. Under both the null and the alternative, the distribution of Z is of the form of $f_Z(z) = g(z) \cdot h(z)$. The latter multiplier, $h(z)$ is the same irrespective of the form of the function $g(\cdot)$.

2. The multiplier term can be determined by observing that the density $N_d\left(\boldsymbol{V}; \boldsymbol{0}; \boldsymbol{I}_d\right)$ depends only on terms involving $\boldsymbol{V}'\boldsymbol{V}$. So, the function $h(z)$ can be obtained readily by comparing a χ^2_d density to $N_d\left(\boldsymbol{V}; \boldsymbol{0}; \boldsymbol{I}_d\right)$ with z substituted for $\boldsymbol{V}'\boldsymbol{V}$. Specifically,

$$h(z) = \frac{\frac{1}{\Gamma\left(\frac{d}{2}\right)2^{d/2}} z^{d/2-1} \exp\left(-\frac{z}{2}\right)}{\frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{z}{2}\right)} = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} z^{d/2-1} \tag{A.9}$$

3. In the test of interest, both the null and alternative marginal distributions of $\boldsymbol{Y}$ depend only on $\boldsymbol{Y}'\boldsymbol{Y}$. Because the multiplier $h(\cdot)$ appears in both the null and alternative distributions of Z, the Bayes' factor could be obtained directly by replacing the term $\boldsymbol{Y}'\boldsymbol{Y}$ by Z_n in the densities under the null and alternative distributions and dividing them. Therefore, the Bayes' factor in favor of the alternative against the null is obtained as:

$$BF(2|1) = \frac{\mu_k^*}{\prod_{i=0}^{k-1}(d+2i)} \frac{1}{(1+n\tau)^{k+\frac{d}{2}}} \exp\{\frac{1}{2}\frac{n\tau}{(n\tau+1)}Z_n\} \tag{A.10}$$

where μ_k^* is the k^{th} raw moment of the $\chi^2_d(\lambda^*)$ distribution where $\lambda^* = \frac{n\tau}{(n\tau+1)}Z_n$.

4. Note that the result in (A.10) would be obtained if we substitute for $\hat{\boldsymbol{\theta}}_1'\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\theta}}_1$ by Z_n in the Bayes' factor result for the linear model case (Johnson and Rossell 2010). We can do this because the MoM prior in (2.16) produces a marginal density for $\boldsymbol{Y}$ that depends only on $\boldsymbol{Y}'\boldsymbol{Y}$. Suppose instead of using the MoM prior in (2.16), the following iMoM prior were applied to $\boldsymbol{\delta}$:

$$\pi_I(\boldsymbol{\delta}) = c_I \left[\frac{\boldsymbol{\delta}'\bar{\boldsymbol{C}}_{11}\boldsymbol{\delta}}{n\tau}\right]^{-\frac{\nu+d}{2}} \exp\left[-\left(\frac{\boldsymbol{\delta}'\bar{\boldsymbol{C}}_{11}\boldsymbol{\delta}}{n\tau}\right)^{-k}\right] \tag{A.11}$$

where

$$c_I = \left|\frac{\bar{\boldsymbol{C}}_{11}}{n\tau}\right|^{1/2} \frac{k}{\Gamma(\nu/2k)} \frac{\Gamma(d/2)}{\pi^{d/2}} \tag{A.12}$$

Again comparing to the linear model case (Johnson and Rossell 2010), the marginal density of $\boldsymbol{Y}$ under the alternative depends only on $\boldsymbol{Y}'\boldsymbol{Y}$. Hence, the Bayes' factor under this iMoM alternative prior could be written down directly as:

$$BF(2|1) = \left(\frac{2}{n\tau}\right)^{d/2} \frac{k\Gamma(d/2)}{\Gamma\left(\frac{\nu}{2k}\right)} \frac{E_z\left[\left(\frac{n\tau}{z}\right)^{(\nu+d)/2} \exp\{-(n\tau/z)^k\}\right]}{\exp\{-\frac{1}{2}Z_n\}} \tag{A.13}$$

where $z \sim \chi^2_d(Z_n)$.

5. With the MoM prior and $k = 1$, the marginal density of Z under the alternative is:

$$\begin{aligned} f_2(Z) &= \frac{1}{d}\left(d + \frac{n\tau}{1+n\tau}Z\right) \frac{1}{(1+n\tau)^{1+\frac{d}{2}}} \frac{1}{(2\pi)^{\frac{d}{2}}} \\ &\qquad \exp\{-\frac{Z}{2(n\tau+1)}\} \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} z^{d/2-1} \qquad \text{(A.14)} \\ &= \frac{\exp\{-\frac{z}{2(1+n\tau)}\}}{(1+n\tau)} \left\{ \begin{array}{c} \left[\frac{1}{\Gamma\left(\frac{d}{2}\right)} \frac{z^{\frac{d}{2}-1}}{\{2(1+n\tau)\}^{\frac{d}{2}}}\right] + \\ n\tau\left[\frac{1}{\Gamma\left(1+\frac{d}{2}\right)} \frac{z^{\frac{d}{2}}}{\{2(1+n\tau)\}^{\frac{d}{2}+1}}\right] \end{array} \right\} \qquad \text{(A.15)} \end{aligned}$$

This is a mixture of scaled χ^2_d and χ^2_{d+2} distributions. Comparing this to the result with local priors, which yields a scaled χ^2_d under the alternative, it could be seen that the accelerant for the Bayes' factor under the null, as seen in Figure 2, comes from the latter part of the mixture.

VITA

Adarsh Joshi was born in Bhilai in the state of Chhattisgarh, India. He graduated from the Indian Institute of Technology Bombay (IIT-B), India in August 2003 with a Bachelor of Technology (B.Tech.) in Mechanical Engineering. In August 2006, he received a Master of Science degree in Statistics from Texas A&M University. In June of 2007, he began pursuing his doctoral research at the University of Texas M. D. Anderson cancer center under the joint program between that institution and the Texas A&M University. He received his Ph.D. in Statistics from Texas A&M University in May of 2010 under the direction of Dr. Valen E. Johnson (primary affiliation: the University of Texas M. D. Anderson cancer center) and Dr. David B. Dahl. His research interests include Bayesian methods, computational statistics and Bioinformatics.

Dr. Joshi may be reached at:

Department of Statistics
Texas A&M University
3143 TAMU
College Station
TX 77843-3143
c/o David B. Dahl or Valen E. Johnson
email: adarshjoshi.stat@gmail.com

The typist of this dissertation is Adarsh Joshi