Abstract

FENG, SHENG. Statistical Studies of Genomics Data (Under the direction of Dr. Zhao-Bang Zeng, Dr. Bruce Weir, Dr. Russ Wolfinger, Dr. Helen Zhang and Dr. Leonard Stefanski)

In recent years, studies on Genetics and Genomics have become one of the most active fields in science. The Genetic and Genomics data have several significant and unique characteristics that bring great challenges for data analysis. Three statistical studies have been presented in this dissertation. In chapter 1, an empirical Bayesian approach has been developed in a linear mixed model for Microarry data analysis. In chapter 2, a multiple order Markov chain model is applied to summarize the local correlation patterns among multiple genetic markers in linkage disequilibrium mapping. In chapter 3, a shrinkage method is being developed to integrate Biological prior knowledge presented in moment statistics. This new method may be useful in some genetic network studies.

Statistical Studies of Genomics Data

by

Sheng Feng

A Dissertation

submitted to the advisory committee on graduate studies of

North Carolina State University

in partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

DEPARTMENT OF STATISTICS

Raleigh, NC

December, 2004

APPROVED BY:

Zhao-Bang Zeng Chair of Advisory Committee	Bruce S. Weir Co-Chair of Advisory Committee
Russ D. Wolfinger	Hao Helen Zhang
Leonard A. Stefanski	

To Lan and my parents

Biography

Sheng Feng was born in Chengdu, Sichuan, China on July 15, 1973. He attended the University of Science and Technology of China in 1991, where he earned a Bachelor of Science degree in Biology in 1996. In 1999, he came to the University of Nebraska at Lincoln and received a Master of Science degree in statistics in May, 2001. He continued his studies towards a Ph.D degree in statistical genetics at North Carolina State University.

Acknowledgements

I would like to express my deep gratitude to my advisors, Dr. Zeng and Dr. Weir. For the past three years, they have been advising me throughout all my works, especially on the study of Linkage Disequilibrium, with valuable ideas, generous support and continuous encouragement. I learned from them not only the statistical techniques and skills, but also the essential principles of research. It is an enjoyable and unforgettable experience to work with them.

I am very much grateful to Dr. Wolfinger as well, for his guidance and providing me the opportunity to work on the project of microarray data analysis. His brilliant ideas made the work outstanding. My sincere thanks also go to Dr. Zhang and Dr. Stefanski, who have helped me to develop and to better understand the statistical shrinkage analysis.

Finally, I would like to thank my wife, Lan Lan, and my parents. Their love, support and patience have been the greatest and most wonderful gifts for me. I dedicate all my love and this dissertation to them.

Contents

List of Tables

List of Figures

Overview

Only about 150 years ago was the basis of genetics founded by Gregor Mendel, when he studied some simple inheritance traits in pea plants. Growing extremely fast, today genetics/genomics has become one of the most active fields in science.

In recent a few years, the invention and development of some novel techniques in molecular biology, such as microarray and DNA sequencing, has brought the genetics/genomics studies to a new level. Benefited from these techniques, information of thousands of features (e.g., genes, markers, chemical compounds, etc.) can be simultaneously collected in a single experiment, which allows biologists to expand the scope of their research to the whole genome. A typical research question in these studies is often to search for a small subset of the thousands of features that are responsible for the variation in phenotypic traits. Though eventually the true features need to be verified by biological experiments, statistical analysis is needed in early steps to narrow down the number of candidates.

There are two significant characteristics in the datasets that make this problem a unique challenge for statisticians. The first characteristic is the so called "large p, small n" problem, i.e., the number of features p is large, while the number of replicates n is often small. Generally, "large p" causes serious multiple comparison problems; "small n" limits the power of statistical tests. The second characteristic is that, many features are correlated following certain underlying biological mechanisms. Sometimes these biological mechanisms themselves are the focus of the study, so the modeling of these correlations are essential; in other cases when the correlations are not of interest, a good understanding of the correlation patterns usually helps us to make more accurate and precise statistical inferences about the core research problems. How to model the correlations among thousands of features and

to integrate this information in analysis remains a big problem.

A simple two-step strategy has been widely used to analyze this type of data. In the first step, a certain statistical test is applied on each feature to quantify the association between the observed features and the phenotypic variation. In the next step, some statistical procedures are used to adjust the multiple-testing problems. A critical value is set at an arbitrary level of statistical significance. The features having test scores more extreme than the critical value are declared to be significant. This strategy will be discussed later with two different names. In Chapter 1, it is called the "single gene analysis" in microarray data analysis; in Chapter 2, it is called the "single marker analysis" in linkage disequilibrium mapping studies.

The two-step strategy is logically sound and practically useful. Many features have been truly detected by this approach. However, it is also well known that this approach is not perfect. There are at least two apparent problems. First, by testing one feature at a time, the correlations among the features has been ignored. Second, when n is too small, the quality of the statistical test of each single feature is questionable, since many of the statistical tests applied are based on large sample theories.

There are three chapters in this thesis. Each chapter is independent and focuses on a different topic. However, since all the statistical works are developed to analyze genetic data, those common characteristics or problems are addressed in all chapters. As the biological backgrounds and the statistical motivation for the three studies are different, a separate introduction is included in each chapter. A brief overview is given below to summarize the statistical objectives in these three studies.

In Chapter 1, an empirical Bayes approach has been developed to improve the quality of the variance component estimates in a linear mixed model for microarray data. The huge oligo-nucleotide array dataset contains p=14010

genes for each of five Drosophila lines. For each line, the minimum number of replications are applied, i.e., two biological replicates and two technical replicates ($n = 2 \times 2$). The research question is to find genes that have different expression patterns in different lines with respect to their mating behaviors. Because of the "small n" problems, the variance components in the mixed model may not be well estimated. As a result, if the "single gene analysis" is applied, no significant genes are detected. The proposed empirical Bayes approach assumes that the variance component of each gene comes from a certain statistical distribution, which may be empirically estimated from all genes. This distribution is then treated as the prior distribution of the variance component. Thus, assuming some similarities among all genes, the information across all genes is used to stabilize the variance estimates of each gene. The performances and some properties of the posterior estimators of the variance components are investigated by simulation studies. Results show that, for the majority of genes, the empirical Bayes estimators of the variance component have smaller bias and smaller variance than the common estimators. The power of the statistical test is also improved.

In the empirical Bayes study, the possible correlations among genes are not modeled, due to the lack of biological models (and hence, the statistical hypothesis) of the correlations. But in some genetic studies, such information is available. The linkage disequilibrium (LD) mapping study described in Chapter 2 is an example. In typical LD mapping studies, besides the phenotype trait values, a large number of genetic markers are typed for each individual. One goal of LD mapping is to locate the genetic variants associated with the trait on chromosomes. It is generally believed that, on average the closer the two markers are, the higher the association is, and *vice versa*. This belief serves as a basic foundation of LD mapping. Thus, to understand and to model the correlations among multiple markers is important. A multiple order Markov chain approach is introduced to summarize the local correlations among multiple markers. Statistical tests are developed

to test some specific correlation patterns, from which an LD map is constructed. This LD map may be useful in LD mapping studies by integrating the marker correlation information. Some simulation studies show that the proposed method may also be a good tool in some genetic studies where the LD is the main focus.

In Chapter 3, a statistical shrinkage method is being developed. This method is able to integrate some common type of biological information into the data analysis. Some applications of this method include: (1) when sample size n is so small that the biological signal and noise can not be well separated, but fortunately some biological information is available which may be helpful to suppress the noise and to make a better statistical prediction; (2) gene network studies. A genetic network study may be conducted by microarray experiments. Usually, a simple network containing just a few genes is already known. The research question is to look for more genes involved in this network. It is obvious that the known network is critical for this study. But this information is available in a specific way (a network) that some commonly statistical methods (i.e., Bayesian statistics) may have problems to utilize this information. The proposed shrinkage approach is developed to take this specific information into account. From a statistical point of view, the proposed method no longer follows the basic principle of unbiasedness. But it has some other desired advantages which makes itself a useful and attractive method.

In each study, some statistical questions remain unsolved. Further development on those methodologies are discussed in the future works section in each chapter.

Chapter 1

Empirical Bayes Analysis of Variance Component Models for Microarray Data

Abstract

A gene-by-gene mixed model analysis (called "single gene analysis") is a useful statistical method for assessing significance for microarray gene differential expression (Chhabra et al, 2003, Chu et al, 2002, 2003, 2004, Jin et al, 2001, Wolfinger et al, 2001). While data of thousands of genes are collected in a typical microarray experiment, the sample size for each gene is usually relatively small, which could limit the statistical power of the analysis. In this report, we introduce an empirical Bayes approach for general variance component models and apply it to microarray data. The power can be improved by integrating prior information on variance components estimated from all genes. The approach starts with a series of single gene analyses. The estimated variance components from each gene are transformed to the "ANOVA components". This transformation makes it possible that the marginal distribution of each ANOVA component is estimated independently. For every gene, the posterior density of each ANOVA component can be easily derived based on previous work (Wolfinger and Kass, 2000). The means of

the posterior distributions are inversely transformed to compute the posterior estimates of the variance components. These posterior estimates may be thought of as the results of the naive estimates from the single gene analysis shrunk towards the prior density. In the real data example, no genes are declared to be significantly different after standard adjustments for multiple comparisons for the single gene analysis. However, the empirical Bayes approach declares hundreds of genes to be significantly differentially expressed. A simulation study is designed to investigate some statistical properties of the empirical Bayes estimators of the variance components. The test size and power of the statistical test are also discussed.

1.1 Introduction

Microarray experiments have been widely used to compare gene expression patterns under different biological backgrounds and conditions. Some recent comprehensive reviews of mircoarray experiments and data analysis include the Nature Reviews - 2004 web focus on microarrays (http:\\ www.nature.com\reviews\focus\microarrays), the Nature Genetics supplement (Nature Genetics, 2003) and Speed (2003).

In a microarray experiment, thousands of genes arrayed on the same chip receive the same treatments. The same factorial design across all chips is shared by each of all genes. Based on this specific data structure, a popular strategy has been applied for data analysis: a single statistical test is performed and a test score is obtained for each and all genes; then a threshold is set over these thousands of test scores. Genes with test scores exceeding the threshold are claimed to be significantly differentially expressed.

The two-step gene by gene mixed model analysis proposed by Wolfinger et al (Wolfinger et al, 2001) (called "single gene analysis" in this paper) is one example. This approach fits a linear mixed model for each gene and claims

significant differential expression of a gene if its associated t-statistic is larger than a preset cutoff value. Other examples, varying on different single gene analysis methods and/or multiple testing adjustment techniques, include Li and Wong (2001a), Irizarry et al (2003), Lonnstedt and Speed (2001), Kerr et al (2000), Kerr et al (2002), Tusher et al (2001), Hochberg and Westfall (2000), and Storey and Tibshirani (2003).

In real data analysis, the statistical power of the single gene analysis is often limited. First, for various reasons, not enough chips or replications are used. For each gene, the sample size is usually small. Second, the statistical tests are performed on a gene-by-gene basis, which causes multiple testing problems. A highly conservative cutoff may prevent detecting any truly differentially expressed genes at a given sample size.

In this report, we will focus on the improvement of the statistical power influenced by the small sample size problems. We will not discuss multiple testing problems, though we will apply the Bonferroni adjustment and a false discovery rate controlling procedure (Benjamini and Hochberg, 1995) to show some results.

Suppose a microarray experiment involves p genes and several random factors (for instance, see the real data described in data and methods), which can be well described by a variance component mixed model. The single gene analysis produces one estimate for each variance component for each gene, and a total of p estimates for p genes. When p is large, a histogram plot of those n estimates shows an empirical distribution, which gives a feeling on how the variance component distributes across all genes. Obviously, this global information is valuable and may help us to derive a better estimator for the variance components for each gene. This is our idea of using Bayesian analysis by treating the empirical global information as the prior.

Lonnstedt and Speed (2001) described an Empirical Bayes method to es-

timate the posterior odds ratio in a simply designed microarray experiment with one variance component, and claimed that the method could be extended to two variance components. Smyth (2004) developed the model of Lonnstedt and Speed into a practical approach. Note that this method works directly on variance components, which are often correlated with each other. So, it may be difficult to extend their work to more complicated cases (i.e., in our example with three variance components) because of the higher-order integrations.

Bayesian methods have been applied in variance component models. Two pioneers, Box and Tiao, gave an explicit theoretical discussion on this issue some decades ago (Box and Tiao, 1973). Using the typical Gaussian distributions for errors and random effects and a non-informative Jeffrey's prior for the variance components, Wolfinger and Kass (2000) developed a quick and reliable method using an independence chain algorithm to estimate the marginal posterior density of the variance components in mixed models. This work extends the classical restricted maximum likelihood (REML) analysis of mixed models and provides foundational ideas for our Empirical Bayes approach of variance component models for microarray data.

1.2 Data and Methods

Drosophila data:

The *Drosophila melanogaster* microarray study investigates the effects of male genotype on post-mating gene expression in female flies using *Drosophila* Affymetrix GeneChips (Affymetrix Inc. Santa Clara, CA). Five experimental treatments ("state") consist of: (1) virgin Canton S females, (2) Canton S females mated to Canton S males, (3) Canton S females mated to mutant line 1 males, (4) Canton S females mated to mutant line 2 males or (5)

Oregon R strain females mated to Oregon R males. For each experimental group, two independent RNA extractions were performed ("prep") and two replicates of each extraction ("chip") were arrayed for a total of four chips per experimental treatment.

The GeneChips contains probe sets representing unique genes, and each probe set consists of 20 probe pairs. Each probe pair has a perfect match (PM) oligonucleotide probe, which is designed exactly complementary to a preselected 25mer of the target gene, and a mismatch probe (MM), which is identical to PM except for one single nucleotide difference at position 13. According to Lockhart et al. (1996), the purpose of the mismatch probe is to serve as an internal control of hybridization specificity. Researchers have different opinions on what measure should be used as the response value for each probe pair in each gene. Some popular choices include the base-2 logarithm of the difference of the PM and MM measures (suitably adjusted for negative values), if we believe that MM serves as a proper internal control; the base-2 logarithm of the PM only, if we do not wish to incorporate any MM information; and the intermediate choice, log(PM) - 0.5log(MM), as suggested by Efron et al (2000). In this study, we choose the base-2 logarithm of the PM measure as the response.

The expression patterns of a total of 14,010 genes are under investigation. Let y_{gijkl} be the base-2 logarithm of the measurement from gene g (g=1,...,14010), "state" i (i=1,2,3,4,5), "prep" j (j=1,2), "chip" k (k=1,2) and "probe" (the PM) l (l=1,...,20). The treatment "state" and the factor probe are treated as fixed. Researchers are interested in comparing the gene expression patterns between each pair of treatments. The factor prep is nested in state. For every state, each of the two preparations is considered as being randomly drawn from a population of preparations. The factor chip is nested in each (state $\times$ prep) combination and is also regarded as random. So, for each gene, the data have a nested structure with three vari-

ance components, the prep(state), the chip(state× prep) and the experiment error.

There are several statistical methods available to analyze the oligonucleotide array data, such as multiplicative model analysis (Li and Wong, 2001a), mixed model analysis (Chu et al, 2002) and fixed linear model analysis (Irizarry et al, 2003). Our study (Chu et al, 2004) shows that the other two methods may not work well under certain conditions, while the simple mixed model analysis seems to be always appropriate for this type of data. This comparison is not a focus in this report. We just use the result and apply mixed models for the oligonucleotide array data.

The two-step single gene mixed model analysis.

This approach was proposed by Wolfinger et al (Wolfinger et al, 2001). The first step is normalization. By treating genes as random replicates, this step aims to remove the main effects from all undesired factors averaged over all genes. For this study, the normalized model could be written as:

$$y_{gijkl} = \mu + state_i + prep(state)_{ij} + chip(state \times prep)_{ijk} + e_{gijkl},$$

where u represents an overall mean value, $state_i$ is the main effect for fly line i, $prep(state)_{ij}$ is the main effect for the i^{th} state and j^{th} prep combination, $chip(state \times prep)_{ijk}$ is the main effect for the i^{th} state, j^{th} prep and k^{th} chip combination. Note that no genetic effects are modeled in this step. All genetic variation is then left in the error term.

In the second step, the residuals from the normalization model, denoted as r_{gijkl}, become the response variable for a gene model, which could be written as:

$$r_{gijkl} = G_g + state_{gi} + probe_{gl} + prep(state)_{gij} + chip(state \times prep)_{gijk} + \gamma_{gijkl}.$$

All effects indexed by g are assumed to serve similar roles to those from the normalization model, but at the individual gene level (with subscript

"g" for distinction). The random effects $prep(state)_{ij}$, $chip(stateprep)_{ijk}$, e_{gijkl} , $prep(state)_{gij}$, $chip(state \times prep)_{gijk}$ and γ_{gijkl} are all assumed to be normally distributed with mean zero and variance components σ^2_{ij}, σ^2_{ijk}, σ^2_e, σ^2_{gij}, σ^2_{gijk} and σ^2_{gijkl}, respectively. Those parameters are estimated by REML. Hypothesis tests are performed with appropriate standard errors, which are functions of the estimates of the variance components and sample size. For more details of this method, please refer to the original paper (Wolfinger et al, 2001).

Bayesian analysis of mixed models with non-informative prior:

Introduced by Wolfinger and Kass (2000), this approach provides a flexible alternative to REML estimation. From a Bayesian point of view, the variance components are assumed to be random variables following some distributions. Since the variance components are not independent of each other in the usual formulation of the model, it is difficult to analytically derive the joint posterior densities in closed forms when two or more variance components are involved in the model. Wolfinger and Kass suggested a transformation of the variance components to the ANOVA components, whose elements are independent of each other. The ANOVA components are defined as the expected mean squares from a traditional ANOVA table. They are linear combinations of the variance components with coefficients usually corresponding to effective sample sizes. For example, in a balanced randomized block experiment with k observations in each block, the ANOVA components are σ^2_{error} and $\sigma^2_{error} + k\sigma^2_b$, where σ^2_{error} and σ^2_b are the residual and block variance components, respectively. The minimum variance quadratic unbiased equations (MIVQUE(0)) (Wolfinger et al, 1994) are used to determine the coefficients of the linear transformation.

Box and Tiao (1973) showed that, for each ANOVA component σ^2_{AC}, the marginal posterior densities can be formulated as: $p'(\sigma^2_{AC}) \propto \pi(\sigma^2_{AC})L(\sigma^2_{AC})$, where $p'(\sigma^2_{AC})$, $\pi(\sigma^2_{AC})$ are the posterior and the prior density, respectively;

$L(\sigma^2_{AC})$ is the likelihood function parameterized in terms of σ^2_{AC}. They further showed that, when the non-informative Jeffrey's prior was used, the posterior densities of the ANOVA components can be well approximated by the inverse Gamma densities:

$$IG_Y(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-\alpha-1} exp(-\frac{\beta}{y}).$$

These inverse gamma distributions can be estimated and are taken as the base densities for an independence chain algorithm to simulate the posterior distributions of the ANOVA components. For more details about this approach please refer to the original paper (Wolfinger and Kass, 2000).

This method will be used in step two in our proposed Empirical Bayes analysis. Note that, since a non-informative prior $\pi(\sigma^2_{AC})$ is used, the posterior densities $p'(\sigma^2_{AC})$ of the variance components are also non-informative. They are equivalent to the REML estimators in terms of informative prior usage. When a desired informative prior, denoted as $pr(\sigma^2_{AC})$, is considered, the non-informative posterior density $p'(\sigma^2_{AC})$ can be updated to the "adjusted posterior density" $p(\sigma^2_{AC}) = pr(\sigma^2_{AC})p'(\sigma^2_{AC}) \propto pr(\sigma^2_{AC})\pi(\sigma^2_{AC})L(\sigma^2_{AC})$. It is this informative posterior density $p(\sigma^2_{AC})$ that serves as the posterior density in this analysis. For an explanation of $p(\sigma^2_{AC})$ and the reason we use it, instead of the regularly defined posterior density $f(\sigma^2_{AC}) \propto pr(\sigma^2_{AC})L(\sigma^2_{AC})$, see the discussion section.

An Empirical Bayes Approach:

For each gene g, the variance components (σ^2_{gij}, σ^2_{gijk} and σ^2_{gijkl}) are assumed to be random variables following some statistical distribution. Note we do not specify the distributions of variance components here. Instead, we will specify the distributions of the ANOVA components as shown below.

The single gene mixed model analysis is applied first. A total of $n = 14010$ sets of variance component estimates are obtained, as well as the 14010 sets of transformed ANOVA components. Histogram plots are used to show the

empirical distribution for each ANOVA component. The prior densities of the ANOVA components are estimated from their empirical distribution. The procedure can be described as following:

1. Obtain the non-informative posterior density for each gene: The Bayesian analysis of mixed models is performed with the non-informative Jeffrey's prior to obtain the inverted gamma posterior density of each ANOVA component for each gene. Practically, this step can be done with SAS Proc Mixed using the PRIOR statement (SAS online help). For each gene g in the example, denote the set of three inverted gamma densities as: $p'_g(\sigma^2_{AC}) = (IG_{g1}(\alpha'_{g1}, \beta'_{g1}), IG_{g2}(\alpha'_{g2}, \beta'_{g2}), IG_{g3}(\alpha'_{g3}, \beta'_{g3}))$.

2. Estimate the informative prior: In the example, the prior distributions for the three ANOVA components are assumed to be inverted gamma with density $pr = (IG_{r1}(\alpha^r_1, \beta^r_1), IG_{r2}(\alpha^r_2, \beta^r_2), IG_{r3}(\alpha^r_3, \beta^r_3))$. To estimate the parameters α^r_m and $\beta^r_m, (m = 1, 2, 3)$, we plot the histograms of the estimated α'_{gm} and β'_{gm} for $g = 1, ..., 14010$ from step 1, and estimate the modes of their empirical distributions. The estimated modes are taken as the estimates for the parameters α^r_m and β^r_m. This is an empirical version of formal hierarchal Bayesian modeling on the parameters α'_{gm} and β'_{gm}. Instead of specifying distributions, we employ only an empirical estimate (the mode) of α^r_m and β^r_m. Obviously, there are other ways to estimate α^r_m and β^r_m. Given abundant (14010) data points, those methods are expected to produce similar results.

3. Update the non-informative posterior density p'_g to the informative posterior density p_g: For gene g, the final posterior density is denoted by p_g, which can be derived as:

$$p_g(\sigma^2_{AC}) = pr(\sigma^2_{AC})p'_g(\sigma^2_{AC}) = (IG_1(\alpha_1, \beta_1), IG_2(\alpha_2, \beta_2), IG_3(\alpha_3, \beta_3)),$$

where $\alpha_m = \alpha'_m + \alpha^r_m + 1$ and $\beta_m = \beta'_m + \beta^r_m$ for $m = 1, 2, 3$. The result follows directly the property of multiplying two inverted gamma densities.

4. Inverse transformation: The means of the posterior density for each

ANOVA component σ^2_{AC} and each gene are estimated. The estimates are inversely transformed to their corresponding values in terms of variance components σ^2. These values are the posterior estimates of the variance components.

5. Mixed model analysis with posterior estimates of variance components held constant: The single gene mixed model analysis is performed again. The test statistic is constructed by treating the posterior estimates of the three variance components as true parameters. The standard normal distribution is used as an approximation for the distribution of the test statistics under the null hypothesis. Gene significance is claimed if the test statistics is larger than a preset cutoff value, adjusted for multiple testing (see results).

1.3 Results

Estimation of the variance components:

Figure 1.1 shows the variance component estimated from the single gene analysis and from the Empirical Bayes analysis. The global distributions of the 14010 variance component estimates from the single gene analysis (the red curves in the upper three graphs) are flatter and widely spread; while the posterior distributions (the green curves) are sharper and more concentrated. The bottom three graphs show how the single gene estimates shrink towards the prior distribution to produce the posterior estimates (only one point, the mode, of the prior distribution is shown). The posterior estimates can be viewed as an optimal compromise between the single gene estimates and the prior.

Increased power to detect the significance:

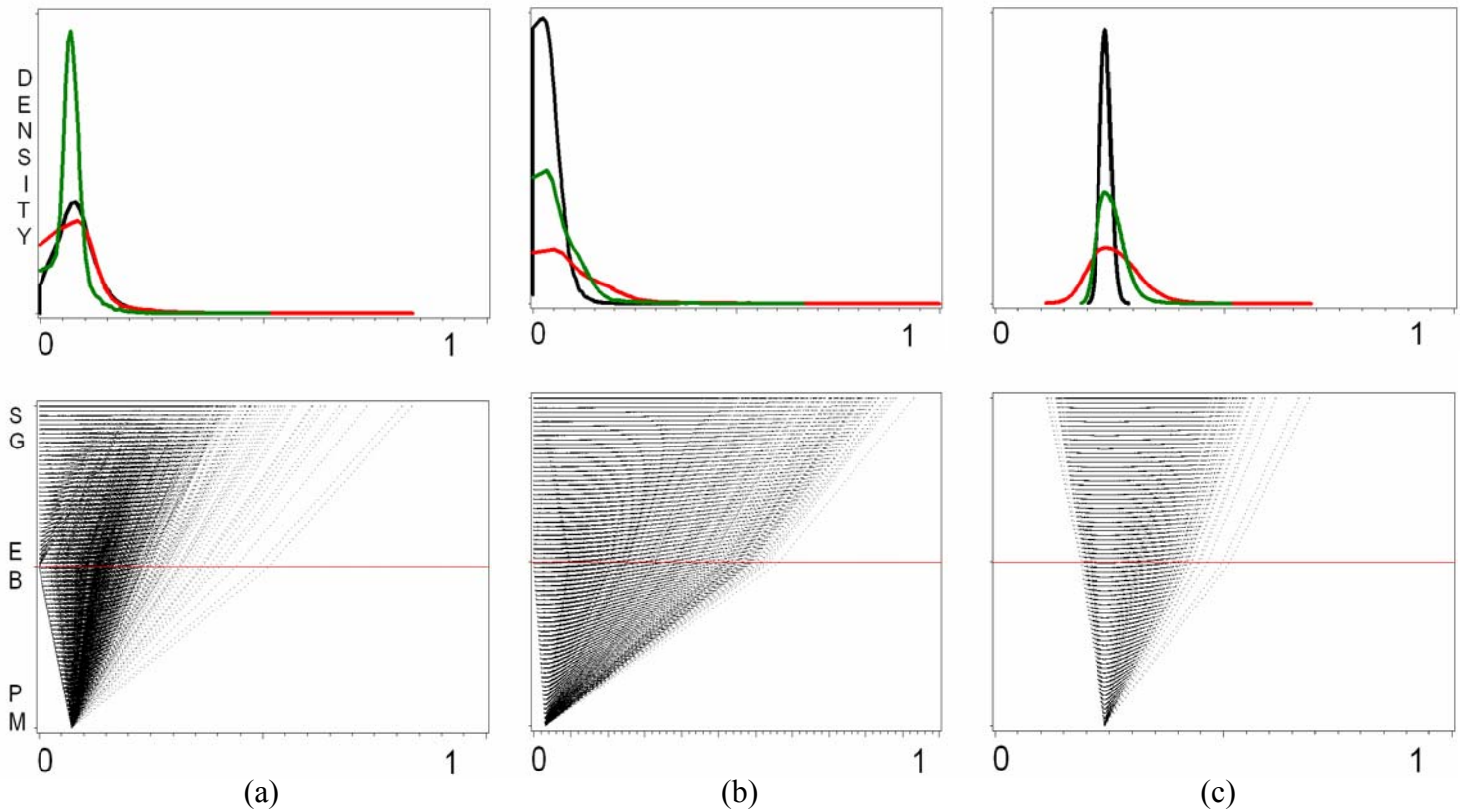


Figure 1.1 Variance component estimates from different methods. Upper: For the empirical Bayesian method, the prior (red), likelihood (black) and posterior (green) distributions of the 3 variance components: (a) σ_{gij}, (b) σ_{gijk} and (c) σ_{gijkl}. Bottom: The variance component estimated from the single gene analysis (SG), the empirical Bayesian analysis (EB), and just a point estimate: the prior mode (PM). The x axis is the squared root of variance component estimates.

	Number of significant genes			
	Bonferroni		False Discovery Rate	
Contrast	S.G.A.	E. B.	S.G.A.	E. B.
Male1_vs_CS	0	5	0	38
Male1_vs_OR-R	0	24	0	91
Male1_vs_Male2	0	9	0	28
Male1_vs_Virgin	0	7	0	48
CS_vs_OR-R	0	33	0	276
CS_vs_Male2	0	22	0	209
Cs_vs_Virgin	0	37	0	183
OR_R_vs_Male2	0	50	0	184
OR-R_vs_Virgin	0	48	0	190
Male2_vs_Virgin	0	9	0	68

Table 1.1. Number of significant genes detected by the single gene analysis (S.G.A.) and the Empirical Bayes analysis (E.B.). The overall experimental error rate is set to be 0.05.

Researchers are interested in all 10 possible pair-wise comparisons between the 5 state levels. Table 1.1 gives the number of genes showing significant expression for each of the 10 contrasts. Multiple testing is adjusted by both Bonferroni (left columns) and false discovery rate (right columns). In both cases, the Empirical Bayes approach is able to declare hundreds of significant genes where the single gene analyses declare none.

1.4 Simulation Study

A Monte Carlo simulation study is performed to investigate (1) some statistical properties of the empirical Bayes estimators of the variance components, as compared to the single gene analysis estimators, and (2) the type I error and power of the statistical test.

Simulation Design:

Data containing 5500 genes are generated. The experiment design structure is exactly the same as that of the Drosophila data. For each gene, the response variable is:

$$y_{gijkl} = G_g + state_{gi} + probe_{gl} + prep(state)_{gij} + chip(state \times prep)_{gijk} + \gamma_{gijkl}.$$

All parameters, except the *state* means $state_{gi}$, take values from the parameter estimates (by the single gene analysis) of the first 5500 genes from the Drosophila data. Among the 5500 genes, 5000 are designed to be non-significant genes and 500 are significant genes. For the non-significant genes, all 5 *state* means are set to be 0; for each of the significant gene g, the mean of *state* 1 is 0, the mean of *state* 5 is c_g, where c_g is set to be large such that, given the known variance components and the normal assumptions, the statistical power to detect the difference between *state* 1 and *state* 5 is close to

1. The means of *state* 2, 3, 4 of gene g then take value 0.25 c_g, 0.5 c_g and 0.75 c_g.

For example, to create data for gene g, the random effects $prep(state)_{gij}$, $chip(state \times prep)_{gijk}$ and γ_{gijkl} are generated independently from each of the three normal distributions with mean 0 and variances σ^2_{gij}, σ^2_{gijk} and σ^2_{gijkl}, respectively. Those three variance components take values as the parameter estimates from the single gene analysis of the real dataset. For the fixed effects, the 20 probe measures take values as the parameter estimates from the real data as well. But the treatment means ($state_{gi}$) have to be adjusted, since in the real dataset the treatment differences are too small and non-significant. Given the three variance components and the design structure, we first calculate the standard deviation of the contrast of two treatment means. The mean of *state* 1 is set to be 0 and the mean of *state*5 is set to be z folds of the standard deviation (for an initial setting, $z = 2$). After the initial dataset of 500 genes are generated, treatment mean difference between *state* 1 and *state* 5 is tested for each gene with the mixed model. The statistical power is estimated by the proportion of the number of the significant tests over all 500 tests. If this proportion is less than 0.9, then the treatment difference is set to be larger (i.e., z is adjusted to take a larger value, e.g., $z = 2.5$) and the data are generated again. This process is repeated until the estimated power is greater than 0.9. Now the values of $state_{g5}$ are set to be c_g, and the means of $state2, 3, 4$ of gene g take value 0.25 c_g, 0.5 c_g and 0.75 c_g.

A total of 50 Monte Carlo simulation replicates are generated.

Analyses and results:

(1) Comparison of the two variance components estimators

Both the single gene analysis approach and the empirical Bayes approach are applied for each simulated data set. To compare the two variance compo-

nents estimators, the bias, the variance and the mean square error (MSE) of each estimator is estimated for every single gene. Since there are 5500 genes, we think it is appropriate to report the results at the global level. The better estimator should be the one having smaller mean bias (MB), mean variance (MV) and mean MSE (MMSE), where the mean is taken over all 5500 genes. Specifically, those statistics are defined as:

$$MB = \frac{1}{G} \sum_{g=1}^{G=5000} |\overline{\hat{\sigma}^2_g} - \sigma^2_g|,$$

$$MV = \frac{1}{G} \sum_{g=1}^{G=5000} \sum_{s=1}^{S=50} (\hat{\sigma}^2_{gs} - \overline{\hat{\sigma}^2_g})^2,$$

$$MMSE = \frac{1}{GS} \sum_{g=1}^{G=5000} \sum_{s=1}^{S=50} (\hat{\sigma}^2_{gs} - \sigma^2_g)^2,$$

where $\hat{\sigma}^2_{gs}$ is the variance component estimate for gene g, in sample s, $\overline{\hat{\sigma}^2_g}$ is the mean of $\hat{\sigma}^2_{gs}$ in the 50 samples. The results are reported in Table 1.2.

The results show that, for all three variance components, the empirical Bayes estimators have smaller mean bias, smaller mean variance and hence smaller mean MSE than the single gene analysis (REML) estimators. At the first glimpse, the results seem unusual, since for any data set of given size, often bias and variance can not be controlled at the same time in estimation. An explanation will be given in the discussion section.

Figure 1.2 shows the comparison between the two estimators of the variance component σ^2_{gijkl} (experiment error) over 5500 genes. It is obviously that the empirical Bayes estimates (blue) are more concentrated around the true values (the black horizontal line) than the REML estimates (red).

(2) The type I error and power

For testing treatment differences in each single gene, a t-test with 5 degrees of freedom is used for the single gene analysis. In the step 5 of the

	MMSE ($x10^{-4}$)		MB ($x10^{-3}$)		MV ($x10^{-4}$)	
	SGA	EB	SGA	EB	SGA	EB
VC1	2.39	1.08	1.8	0.7	2.38	1.07
VC2	0.79	0.44	1.0	0.1	0.78	0.44
VC3	0.42	0.18	0.1	0.0	0.42	0.18

Table 1.2. Comparisons between the 2 variance components estimators

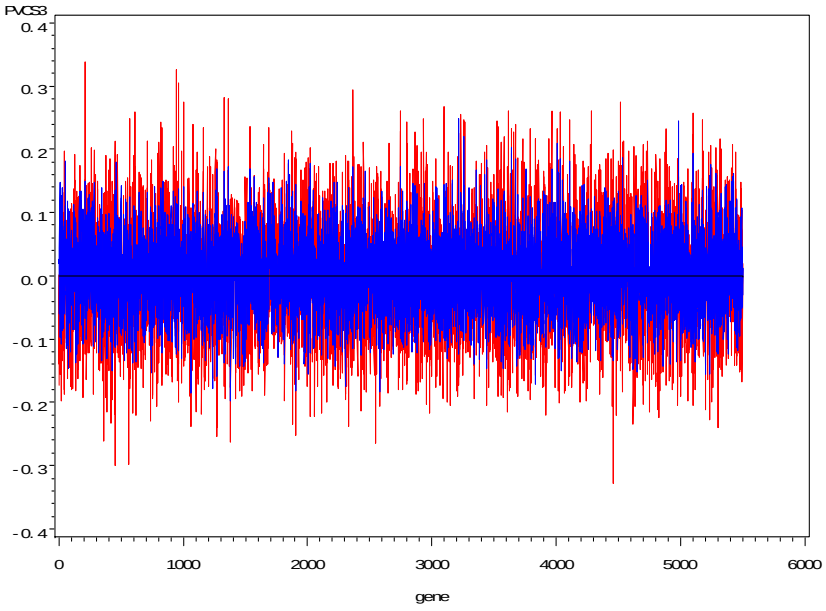


Figure 1.2. Comparison between the REML estimator (red) and the EB estimator (blue) of the experiment error over 5500 genes. The x-axis contains the gene numbers; the y-axis is the estimated experiment error, standardized by the true value: $(\bar{\hat{\sigma}}_g^2 - \sigma_g^2)/\sigma_g^2$. The length of the short vertical red or blue lines on each gene represents the values of the estimates. The horizontal line at 0 is the true parameter (self –standardized to be 0).

empirical Bayes algorithm, the test statistic is constructed by replacing the single gene analysis (REML) estimates of the variance components by the posterior empirical Bayes estimates. A natural question arises: what is the distribution of the new test statistic under the null hypothesis?

Unfortunately, the exact answer is unknown analytically. One possible solution is to numerically approximate the true distribution by applying certain re-sampling techniques. However, the computation could be extremely intensive. In stead, we approximate the distribution of the test statistic by a standard normal distribution. This trick is usually applied when there is reasonable belief or evidence that the "true values" of the variance components are replaced in the test statistic. Since we expect (and observed in the simulation) the empirical Bayes estimates are much "closer" to the true values of the parameters than the single gene analysis estimates, this normal approximation is used in the algorithm.

It is thus important to know how good, especially from the test size point of view, this normal approximation is. For this purpose, we compare three approaches, the single gene analysis with t-test of 5 degree of freedom, the empirical Bayes analysis with normal test approximation and the single gene analysis with normal test given the true variance components, by investigating the test size and the power. Based on the design of the simulation, there are 5000 non-significant genes and 500 significant genes, so the test size and power can be estimated at the same time. The results are reported in Table 1.3.

The results indicate that, for all three approaches, the test size is around 0.05 (for the EB approach, the number is a little bit larger than 0.05, but the inflation seems to be negligible). This suggests that, in this simulation, when the test size is considered, the standard normal distribution is an appropriate approximation for the distribution of the empirical Bayes test statistic under the null hypothesis. Figure 1.3 shows the empirical distribution of

	Test Size			Power		
Contrast	SGA	EB	True	SGA	EB	True
T1 vs. T2 (.25c)	.049	.056	.050	.124	.146	.220
T1 vs. T3 (.50c)	.050	.053	.053	.348	.480	.510
T1 vs. T4 (.75c)	.047	.051	.050	.662	.810	.832
T1 vs. T5 (c)	.048	.048	.043	.844	.904	.914
T2 vs. T3 (.25c)	.050	.054	.053	.134	.166	.224
T2 vs. T4 (.50c)	.051	.057	.057	.392	.518	.564
T2 vs. T5 (.75c)	.049	.050	.044	.642	.724	.764
T3 vs. T4 (.25c)	.048	.053	.049	.126	.184	.228
T3 vs. T5 (.50c)	.051	.054	.045	.346	.412	.448
T4 vs. T5 (.25c)	.046	.049	.049	.096	.146	.194

Table 1.3. Test size and power calculation

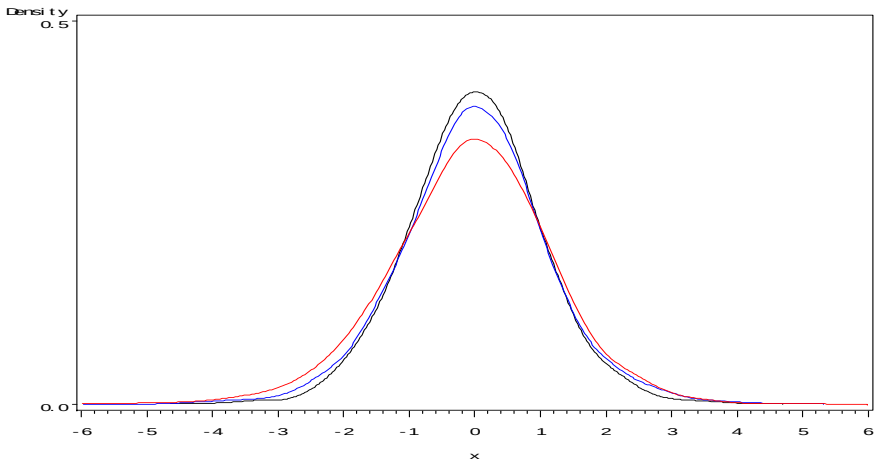


Figure 1.3. The empirical distribution of the test statistics under the null hypothesis from the 3 different approaches: the black curve, SGA with a normal test, given the true variance component value (should be a standard normal in theory); the red curve, SGA with a *t*-test (should be *t*-distribution with 5 degree of freedom in theory); the blue curve, EB with a normal approximation.

the test statistics under the null hypothesis (from the 5000 non-significant genes). It appears that, though the distribution of the empirical Bayesian test statistic (blue curve) is unknown, a standard normal distribution (black curve) may be a good approximation. On the other hand, the power of the empirical Bayes approach is noticeably higher than that of the single gene analysis with a t-test; but lower than that of the single gene analysis with a standard normal test given the true variance components. Combining results from both sides, we conclude that, comparing to the single gene analysis approach, the empirical Bayes approach does improve the power of detecting true significant genes.

1.5 Discussion

Microarray data contain thousands of genes observed under the same experimental design structure, sometimes involving several random factors, or variance components. We may treat the variance component estimates obtained from the single gene analysis as a realization from a prior distribution. In this way, the low power caused by small sample size may be improved by integrating information from all genes.

The posterior density is

$$p_g(\sigma^2_{AC}) = pr(\sigma^2_{AC})p'_g(\sigma^2_{AC}) \propto pr(\sigma^2_{AC})\pi(\sigma^2_{AC})L(\sigma^2_{AC}) \propto \pi(\sigma^2_{AC})f_g(\sigma^2_{AC}),$$

where $f_g(\sigma^2_{AC}) \propto pr(\sigma^2_{AC})L(\sigma^2_{AC})$. To understand $p_g(\sigma^2_{AC})$, we may think of it as $f_g(\sigma^2_{AC})$ being adjusted by the non-informative Jeffrey's prior $\pi(\sigma^2_{AC})$. Since $p_g(\sigma^2_{AC})$ and $f_g(\sigma^2_{AC})$ only differ in $\pi(\sigma^2_{AC})$, they are equivalent in terms of usage of the prior information $pr(\sigma^2_{AC})$. However, in comparison to $f_g(\sigma^2_{AC})$, whose closed form solution is difficult to obtain, the adjusted posterior density $p_g(\sigma^2_{AC})$ can be estimated in a much more convenient way since (1) the non-informative posterior density $p'_g(\sigma^2_{AC})$ can be directly sim-

ulated from the independence chain algorithm; and (2) both $pr(\sigma^2_{AC}$ and $p'_g(\sigma^2_{AC})$ are inverted Gamma densities, so that the parameters in $p_g(\sigma^2_{AC})$ can be easily estimated (as shown in data and methods). This is the reason that we estimate $p_g(\sigma^2_{AC})$ instead of $f_g(\sigma^2_{AC})$.

The empirical Bayes method can be thought as a statistical shrinkage method (actually, in general all Bayes estimators can be thought as shrinkage estimators). Like other shrinkage methods, such as the ridge regression (Hoerl and Kennard, 1970) and the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), this method trades off decreased variance for possibly increased bias in parameter estimation. During the shrinkage process, the variance of the variance component estimator is expected to be reduced, while large posterior bias may be introduced if the destination of the shrinkage, in our example, the prior, is biased from the true values. However, in this empirical Bayes method, the prior density of each variance component is estimated from all 14010 genes. The histogram plot of the estimates over the 14010 genes gives a pretty good indication on the "true" distribution of the variance component. The simulation results show this estimated prior density represents the true distribution well enough so that the shrinkage (posterior) estimator not only has smaller variance as expected, but also has smaller bias than the REML estimator (the REML estimator is known to be asymptotically unbiased, but for small sample, i.e., in this Drosophila data set, it could be biased). Overall, for small samples, the empirical Bayes method tends to produce variance component estimators with smaller MSE than the REML estimators.

Our proposed empirical Bayes method works on transformed variance components so that the obstacle of non-independence between variance components is avoided, which make it a powerful approach to accommodate more complicated models. The Bayesian inference is performed only on the variance components, without any further informative distribution assumptions

on other parameters (i.e., treatment means). This makes it possible that the hypothesis testing, including composite comparisons between treatments, is performed in a regular way ("regular" means in a classic linear mixed model framework) with the differences being that the posterior estimates of the variance components are held constant in the final analysis and the test statistics is reasonably assumed to be approximately normally distributed. These features largely simplify the analysis and reduce the computational burden. This kind of analysis is readily performed using mixed model software such as SAS Proc Mixed.

The real data example in this paper is from Affymetrix oligonucleotide array data, and represents a case where treatment effects are somewhat subtle. The proposed approach borrows strength across all of the genes in a classical fashion and greatly increases the number of significant genes. The approach can be used in any case where a general mixed model is appropriate, including those applied two-color cDNA array data (i.e., Jin et al, 2001).

1.6 Future Work

There are still some interesting questions in the procedure which deserve future studies. Basically, we assume that the variance components are random variables following certain known distributions. Then those distributions, or some summary statistics from those distributions are estimated empirically. One question is, in this specific Drosophila dataset, the histogram of the estimated variance components is smooth and has a single mode. Our assumption seems to be perfectly valid. However, it is possible that a prior density is hard to estimate by certain known densities, if the empirical curve is not that "good", i.e., not smooth, or with multiple modes. Under those situations, the proposed method may not work well. However, the idea of

borrowing information from thousands of genes to improve the parameter estimator of a single gene is still applicable. Though the whole prior distribution of the variance component is difficult to estimate, some information, such as empirical mean and variance of the distribution can still be obtained. The point shrinkage approach introduced in chapter 3 may be applied to integrate that prior moment information to generate a posterior estimator for the variance components. Another question is, after the posterior density of the variance component is estimated, only one point from the density (mean or mode) is used as the posterior estimate. It is not known how much information has been lost by this procedure. The choice of choosing mean or mode may cause a large difference in result, especially when the posterior density is highly skewed. This problem may be solved by drawing a sample from the posterior density directly so that the distribution of the test statistics can be estimated directly. The sampling is not difficult since the ANOVA components are independent of each other. The work is in progress.

Chapter 2

Estimating and Testing Linkage Disequilibrium Patterns by Multiple Order Markov Chains and the Linkage Disequilibrium Map

2.1 Introduction

Many common diseases, such as cystic fibrosis, diabetes, cancer, stroke, schizophrenia, heart disease, asthma, etc, are usually caused by the combined effects of genetic variants and environmental factors. The efforts of defining and hunting for those factors have never ceased. However, until today, except for some diseases with single Mendelian genetic factors (e.g., cystic fibrosis), only very a few successes have been declared for the much more common complex diseases (e.g., all other diseases listed above).

By modeling the recombination events in meiosis, pedigree-based linkage mapping is a powerful tool for locating Mendelian disease genes, mainly because the single genetic signal is simple and strong, and a good understanding

of the underlying biological mechanism helps to increase the resolution of the mapping results. See Kerem et al, (1989) and Riordan et al, (1989) for examples of the linkage mapping on cystic fibrosis; and see Weir (1996) for a brief review on some statistical issues related to linkage mapping. In complex disease studies, however, linkage mapping often produces weak and inconsistent results. In contrast to Mendelian diseases, a large number of genetic variants are likely involved in the complex diseases, with small contribution from each variant. Limited by sample/pedigree size, the resolution of the linkage mapping is relatively low, often in the range of 1 centiMorgans, or roughly 1 million bases in human. This is too wide for further molecular studies.

Studies of linkage disequilibrium (LD) mapping aim to estimate the locations of genetic variants on chromosomes at a much finer scale and predict their genetic effects. LD mapping is based on a measure LD, a statistical concept quantifying the degree of non-random association between two or multiple alleles at different loci in a target population. The LD between two loci has been extensively studied. Many mathematical formulations are suggested as the two-locus LD measures. For example, two commonly used two-locus LD measures include the statistical correlation γ (Hill and Robertson, 1968), and D or its normalized value $D^{'}$ (Lewontin, 1964)(γ and D will be defined in the next section). It is important to emphasize that, whatever form it takes, the LD in the current population is created by all evolutionary events throughout history, e.g., selection, recombination, mutation, drift, migration, etc.. This makes an unambiguous genetic interpretation of LD difficult and a decent biological modeling of LD, especially for multiple-locus LD, almost impossible. The study of LD mapping is popular since we optimistically hope that, by considering all recombination events in history, this approach has the potential to increase the mapping resolution, despite all other sources of variance in evolution.

In a typical LD mapping study in a natural population, a random sample of uncorrelated individuals is drawn. For each individual in the sample, two types of data are collected, the phenotype data and the genotype data. The phenotype data, which may also be called "traits", quantify the physical appearance and characteristics of an individual. The data can be continuous, such as body weight or blood pressure (quantitative trait); or categorical, such as disease status (categorical trait). The genotype data contain information of the genetic constitution of an individual through genetic markers, the DNA sequence variations. The locations of the genetic markers and the specific allele at each marker of each individual are recorded. In recent years, the single nucleotide polymorphism (SNP) markers have been widely used in LD mapping. It is estimated that there could be as many as 10 million SNP markers in human populations (roughly 1 SNP every 300 bases), which constitute more than 90% of the total variation in human genome (HapMap website). These high-density SNP markers are desirable for a genome-wide scan for disease alleles. In this study, the proposed method is developed for SNP markers.

Each SNP marker usually has two alleles and hence is usually modeled as a *Bernoulli* random variable. The statistical correlation γ (a LD measure) between two loci has the same magnitude, but maybe opposite signs, no matter which two alleles are considered. Thus for SNP markers, the two-locus γ may be simply understood as a measure quantifying the correlation between these two *Bernoulli* variables (this may not be true if a marker has multiple alleles).

There are two types of correlation in the dataset, the correlation among multiple markers and the correlation between the markers and the trait. Generally, the latter one is the main focus of the LD mapping study. A better understanding of the correlation among markers is surely helpful to make more precise and accurate statistical statements on the correlation

between the markers and the trait.

A commonly used strategy in LD mapping is to evaluate the association between *each* single marker and the trait by some statistical test. The philosophy underlying the test is that the markers that are significantly associated with the trait may be close to the disease allele, or themselves may be the disease alleles (the so-called Quantitative Trait Nucleotide, or QTN). By applying this single marker analysis, each marker is tested independently. The correlations among markers are not considered. The statistical power of this approach could be limited.

Several methods for multiple-marker LD mapping have been proposed (Terwilliger, 1995; Devlin, et al, 1996; Xiong, et al, 1997). In those methods, multiple markers in a specific genome region are included in a likelihood model with all genetic parameters estimated simultaneously. Compared to the single marker analysis, the statistical power is enhanced. However, strictly speaking, these methods can be viewed as a simple "combination of multiple single marker analysis"(McPeek and Strahs, 1999). The information of the LD background is still not modeled and used in the analysis, as commented by Jorde (Jorde, 2000).

The two-locus LD is a key measure for the single marker analysis. However, to fully characterize a p biallelic marker system, $(2^p - 1)$ correlation measures are needed. This includes C_p^1 single marker allelic frequencies; C_p^2 two-locus LD; C_p^3 three-locus LD; ... ; and C_p^p p-locus LD. The number of two-locus LD measures only contribute to a very small percentage of the total number of correlation measures. As McPeek and Strahs pointed out "...to consider only pairwise information on LD among loci when extensive multi-locus haplotypes are available is a tremendous waste of valuable information"(McPeek and Strahs, 1999). Many mathematical formulations have been suggested for multiple-locus (> 2) LD during the past century (Geiringer, 1944; Bennett, 1954; Slatkin, 1972; Gorelick and Laubichler,

2004). But unlike the well studied two-locus LD, very few LD mapping studies and population genetics studies are based on multiple-locus LD measures (see Hastings, 1984 for an example).

One apparent obstacle is how to feasibly model the multiple-locus LD and to use this information in the LD mapping studies. The number of multiple-locus LD measures could be extremely large, even when p is moderate. In practice, not all LD can be estimated, since in general not all 2^p different types of gamete can be observed in a finite sample. Also, the sample distributions, or even the variance, of those LD estimators are difficult to obtain. So it is hard to test the significance of these LD measures, especially for higher order (e.g., when order > 3) LD (Weir, 1996). Further, even if the LD background can be well modeled, it is not clear how these LD measures can be efficiently used in LD mapping to improve statistical power.

Alternative methods have been developed. Instead of using each single marker, the whole hyplotype is taken as a single genetic variant (e.g., McPeek and Strahs, 1999; Lin, 2003; Zhao, et al, 2004). A haplotype is a combination of alleles at a group of markers which are located closely together on the same chromosome and which tend to be inherited together. Basically, all LD measures are just functions of haplotype frequencies. With all LD information embodied in hyplotypes, this approach avoids the need to model the LD background while still being able to use the information. However, the haplotype model may have as many as $(2^p - 1)$ parameters, i.e., the frequency of each possible haplotype for p biallelic markers. In other words, with the same number of parameters, the haplotype model is equivalent in dimension to the "full" single marker model with all higher order LD considered. In real data analysis, this approach is feasible since the number of observed haplotypes in a sample is usually much smaller than all possible haplotypes, thus a large number of parameters do not appear in the model. Practically, limited by computational capabilities and sample size, the haplotype approach may be

used only in one or several candidate gene regions with very a few markers (personal communication with Dr. D.Y.Lin).

Does this complicated haplotype approach have higher statistical power than a series of simple single marker analysis? There is no straight answer to this question. Some studies suggest that the results largely depend on the local LD patterns. It seems that the haplotype model performs better in a local chromosome region where markers are in higher order LD, but not as well if markers are in lower order LD (Akey, et al, 2001; Long and Langley, 1999; Kaplan and Morris, 2001; Zhang, K et al, 2002).

In summary, the single marker analysis ignores the correlation among markers. The approach is simple but it may not be sufficient. The haplotype approach takes all LD background into account, but it is not simple and its application is limited. Obviously, a logical improvement is to find a balance between the two extreme models based on the local LD patterns. Sharing similar idea, Zhang, X., et al (2003) developed the Bayesian Adaptive Regression Splines (BARS) method aiming to "bridge the gap between single locus and haplotype-based tests". In this method, a statistic (e.g., LD) from single marker analysis is obtained first and then a single estimation of the disease locus is made by adjusting all marker information with a nonparametric regression.

We will take a different approach, which is more straightforward. Stimulated by the genetic map in linkage analysis, an LD map is constructed as the first step to summarize the local multiple-locus LD patterns. Then a likelihood based mapping method will be developed based on this local LD map. This chapter will mainly focus on the first step. The second step will be briefly discussed in the section of future studies.

At the first step, the statistical challenges are:

(1) to find an easy, efficient and reliable statistical tool to summarize the

local multiple-locus LD patterns among, possibly, millions of SNP markers and;

(2) to construct an LD map, where the information of local LD pattern is represented in a way such that this information can be used in LD mapping.

For challenge (1), a direct approach is to estimate and test all the multiple-locus LD. However, as discussed previously, there are many technical difficulties. In this chapter, we introduce a simple approach involving the multiple-order Markov chain models. This approach summarizes the local higher order LD patterns along the chromosome in terms of the order of Markov chains. The relations between the Markov chain parameters and the LD measures are explored. A better fit of a Markov chain model of a certain order in the local chromosome region indicates the existence of the same and lower order of LD patterns in this region. Consequently, a local LD map can be constructed based on the multiple-order Markov chain model fitting results.

The idea of LD map comes from the concept of the linkage map in designed cross populations or pedigrees. It is known that a specific type of correlation pattern among multiple markers can be well modeled by some genetic maps in a linkage analysis in experimental cross populations. Through a designed cross from inbred lines, the recombination events can be observed. A linkage map can thus be generated by modeling the interference effects of double recombination with different assumptions. The Haldane genetic map assumes the interference effects are absent, i.e., the recombination event in one interval is independent of those in any other intervals. Between two loci 1 and 2, the map distance x_{12} is defined as a logarithm function of the recombination rate c_{12}. Specifically,

$$x_{12} = -0.5\ log(1 - 2c_{12}).$$

x takes value from 0 to 1, with unit "Morgan", which is defined as the distance along which one recombination event is expected to occur between locus 1

and 2, per gamete, per generation. Depending on the specific cross design, the quantity $(1-2c_{12})$ (the "linkage parameter") is related to the statistical correlation γ. For example, in the backcross design, $\gamma_{12} = (1-2c_{12})$. In this case $x_{12} = -0.5log(\gamma_{12})$. Further, assuming no interference, for loci 1, 2, 3 in this order, $(1-2c_{13}) = (1-2c_{12})(1-2c_{23})$, or $\gamma_{13} = \gamma_{12}\gamma_{23}$. This is a specific type of correlation pattern (multiplicative correlation) among multiple markers, which directly implies $x_{13} = x_{12} + x_{23}$.

Note that, just from a statistical point of view, if three binary random variables M_1, M_2, M_3, are truly generated from a Markov chain model of order 1 ($MC1$), then it can be easily shown that:

$$\gamma_{13} = \gamma_{12}\gamma_{23}. \qquad (Eq.1)$$

So $Eq.1$ is an $MC1$ property. Consequently, if all SNP markers in a chromosome region can be appropriately modeled by $MC1$, then an additive map can be constructed with map distance being the logarithm of the statistical correlation γ. The Haldane genetic map is such an additive map with a Markov chain interpretation, since its assumption of no interference satisfies the Markovian property.

When an LD map is to be constructed from a set of SNP data, certainly, it is inappropriate to simply assume that the markers are well modeled by $MC1$ model. Instead, the $MC1$ property $Eq.1$ can be examined and tested statistically. For some sub-regions of the chromosome where the $Eq.1$ holds, the map distance is additive. For other parts of the chromosome, the level of the complexity of the local LD pattern is recorded.

2.2 Methods and Results

2.2.1 Model settings and vocabularies

The method has been developed under some simplified situations. Consider a dataset containing C independent haplotypes from a natural population with random mating. Hardy-Weinberger equilibrium is assumed. The hyplotype data can be obtained either directly from biological experiments (e.g., Deluca et al, 2003) or estimated from diplotype data by some statistical methods (Weir and Cockerham, 1979; Hawley and Kidd, 1995; Long, et al, 1995).

Suppose chromosome c has m_c biallelic SNP markers, with the two alleles (states) denoted as 1 and 0. The SNP marker i is modeled by a *Bernoulli* random variable M_i for $i = 1, ..., m_c$, which has observation $m_{i,c}(m_{i,c} = 1, 0)$ on chromosome c $(c = 1, ..., C)$. Let $P_i = P(M_i = 1)$ be the probability of allele 1 for marker M_i in the population. $H_{(i,k)}$ is the haplotype containing k consecutive markers starting from the i^{th} marker, e.g., $M_i, M_{i+1}, ..., M_{i+k-1}$. $h_{c(i,k)}$, a $k \times 1$ vector of the observed $H_{(i,k)}$ on the chromosome c, i.e., $h_{c(i,k)} = [m_{i,c}, m_{i+1,c}, ..., m_{i+k-1,c}]$. $P(H_{(i,k)} = h_{c(i,k)})$ is the population frequency for the specific haplotype $h_{c(i,k)}$. The conditional probability that a marker M_i takes value $m_{i,c}$ given its previous r markers $M_{i-r}, ..., M_{i-1}$ taking values $h_{c(i-r,r)}$, is denoted as $P(M_i = m_{i,c}|H_{(i-r,r)} = h_{c(i-r,r)})$, or $P^{(i,H_{(i-r,r)})}_{(m_{(i,c)},h_{c(i-r,r)})}$. If we assume the haplotype counts follow a multinomial distribution (Weir, 1996), the maximum likelihood estimate of the haplotype frequency $P(H_{(i,k)})$ is: $\hat{P}(H_{(i,k)}) = (\text{ counts of } H_{(i,k)})/n$; the conditional probability $P(M_i = m_{i,c}|H_{(i-r,r)} = h_{c(i-r,r)})$ can be estimated as $\hat{P}(H_{(i-r,r+m_{i,c})})/\hat{P}(H_{(i-r,r)})$, where $H_{(i-r,r+m_{i,c})}$ represents the haplotype $[M_{i-r}, ..., M_{i-1}, M_i = m_{i,c}]$.

The two-locus LD measure D_{12} between locus 1 and 2 is defined as:

$$D_{12} = P_{12} - P_1 P_2,$$

where $P_{12} = P(H_{(1,2)})$. For three loci, the definition of three-locus LD by Bennett (1954) is adopted:

$$D_{123} = P_{123} - P_1 D_{23} - P_2 D_{13} - P_3 D_{12} - P_1 P_2 P_3,$$

where $P_{123} = P(H_{(1,3)})$.

D' is a normalized version of D, which is defined as:

$$D'_{12} = \begin{cases} D_{12}/max(-P_1P_2, -(1-P_1)(1-P_2)) & \text{if } D_{12} < 0, \\ D_{12}/min(P_1(1-P_2), (1-P_1)P_2) & \text{if } D_{12} > 0. \end{cases}$$

Another two-locus LD measure γ_{12} is defined as:

$$\gamma_{12} = D_{12}/\sqrt{P_1(1-P_1)P_2(1-P_2)}.$$

It can be shown that, D_{12}, γ_{12} are the statistical covariance and correlation between the two *Bernoulli* variables M_1 and M_2, respectively.

A third two-locus LD measure ρ is defined as (Zhang, W. et al, 2002(a)):

$$\rho_{12} = |D_{12}|/min[P_1P_2, P_1(1-P_2), (1-P_1)P_2, (1-P_1)(1-P_2)].$$

In random samples, ρ is numerically equal to D'. An LD map constructed based on this measure will be briefly mentioned in the Discussion section.

2.2.2 Moving Multi-order Markov chain

Multiple-order Markov chain models have been used to measure how far the correlations between adjacent DNA bases extend along a chromosome (Weir, 1996). In this study, a multiple-order Markov chain model is applied to measure the complexity of the higher order LD patterns.

A Moving window is defined along the chromosome. The window size is defined as the number of markers in the window. The window size could be flexible. If the overall complexity of LD pattern in the whole region is of

Table 2.1: The Number of Parameters in Multiple-Order Markov Chain Models for windows of length ω

ω	Number of parameters in MC models						
	MC0	MC1	MC2	MC3	MC4	...	MCr
1	1	-	-	-	-	.	-
2	2	3	-	-	-	.	-
3	3	5	7	-	-	.	-
4	4	7	11	15	-	.	-
5	5	9	15	23	31	.	-
.	.	.	.	.	.	.	.
ω	ω	$2\omega-1$	$4\omega-5$	$8\omega-17$	$16\omega-49$	...	$(2^r-1)+2^r(\omega-r)$

interest, then the whole region can be viewed as a window with the maximum size m_c; if more detailed local information is of interest, the window size should be set small. For a window with specific size ω ($\omega \leq m_c$), the highest order of Markov chain that can be applied is $(\omega-1)$; while the lowest order is always 0, implying independence among markers. For simplicity, the Markov chain of order r is denoted as MCr.

The Markov chain models applied in this study are non-stationary. All transition probabilities are locus-specific. For a chromosome region containing ω markers $(M_1, M_2, ..., M_\omega)$, the MCr model has (2^r-1) initial parameters and $(2^r(\omega-r))$ conditional parameters. The numbers of parameters are summarized in Table 2.1.

The likelihood of observing $h_{c(1,\omega)} = (m_{1,c}, m_{2,c}, ...m_{\omega,c})$ is a function of the Markov chain order r:

$$L_c(MCr) = P(H_{(1,r-1)} = h_{c(1,r-1)}) \times P(M_r|H_{(1,r-1)} = h_{c(1,r-1)})\times$$

$$P(M_{r+1}|H_{(2,r-1)} = h_{c(2,r-1)})\times ...\times P(M_\omega|H_{(\omega-r+1,r-1)} = h_{c(\omega-r+1,r-1)}).$$

For C independent chromosomes, the total likelihood is

$$L(MCr) = \prod_{c=1}^{N} L_c(MCr).$$

There are two types of parameters in the likelihood, the $2^r - 1$ initial parameters $P(H_{(1,r-1)} = h_{c(1,r-1)})$ and the $2^r(\omega - r)$ conditional probabilities $P(M_r|H_{(1,r-1)} = h_{c(1,r-1)})$, ... , $P(M_\omega|H_{(\omega-r+1,r-1)} = h_{c(\omega-r+1,r-1)})$. As shown in last section, the maximum likelihood estimates of both types of parameters can be obtained by assuming that the haplotype counts follow a multinomial distribution. In the sample, the likelihood $L(MCr)$ is calculated by replacing all parameters with their maximum likelihood estimates. Note that only the observed haplotypes contribute to the likelihood.

Within each moving window of size ω, Markov chain models with different orders, from 0 to r, are applied to fit the data and compared by the Bayesian information criterion (BIC). Some studies suggest that the BIC is a valid criterion of estimating the orders of Markov chain models (Katz, 1981; Cisizar and Shields, 1999; Finesso, 1992). The BIC is defined as:

$$BIC(r) = \text{Constant} - 2\log(L(MCr)) + d_r \log(s_r),$$

where $d_r = (2^r - 1) + 2^r(\omega - r)$ is the number of free parameters in the MCr model; s_r is the number of observations, i.e., $n(\omega - r)$, the number of subsequences of length $(r+1)$. The order of the MC model with the smallest BIC value is selected and recorded as a statistic associated with this window. As the window moves along the chromosome, a series of such statistics are collected. Those statistics provide a summary of the Markov chain model fitting results in each window along the chromosome.

A *Drosophila* dataset, reported by DeLuca et al, 2003, is analyzed by the proposed multiple-order Markov chain models. In the dataset, 36 biallelic markers (31 SNP markers, 5 insertion/deletion markers) are genotyped within the 5.5kb Dopa decarboxylase (D*dc*) gene region. The haplotype data are available by DNA sequencing on 173 *Drosophila* lines derived from the

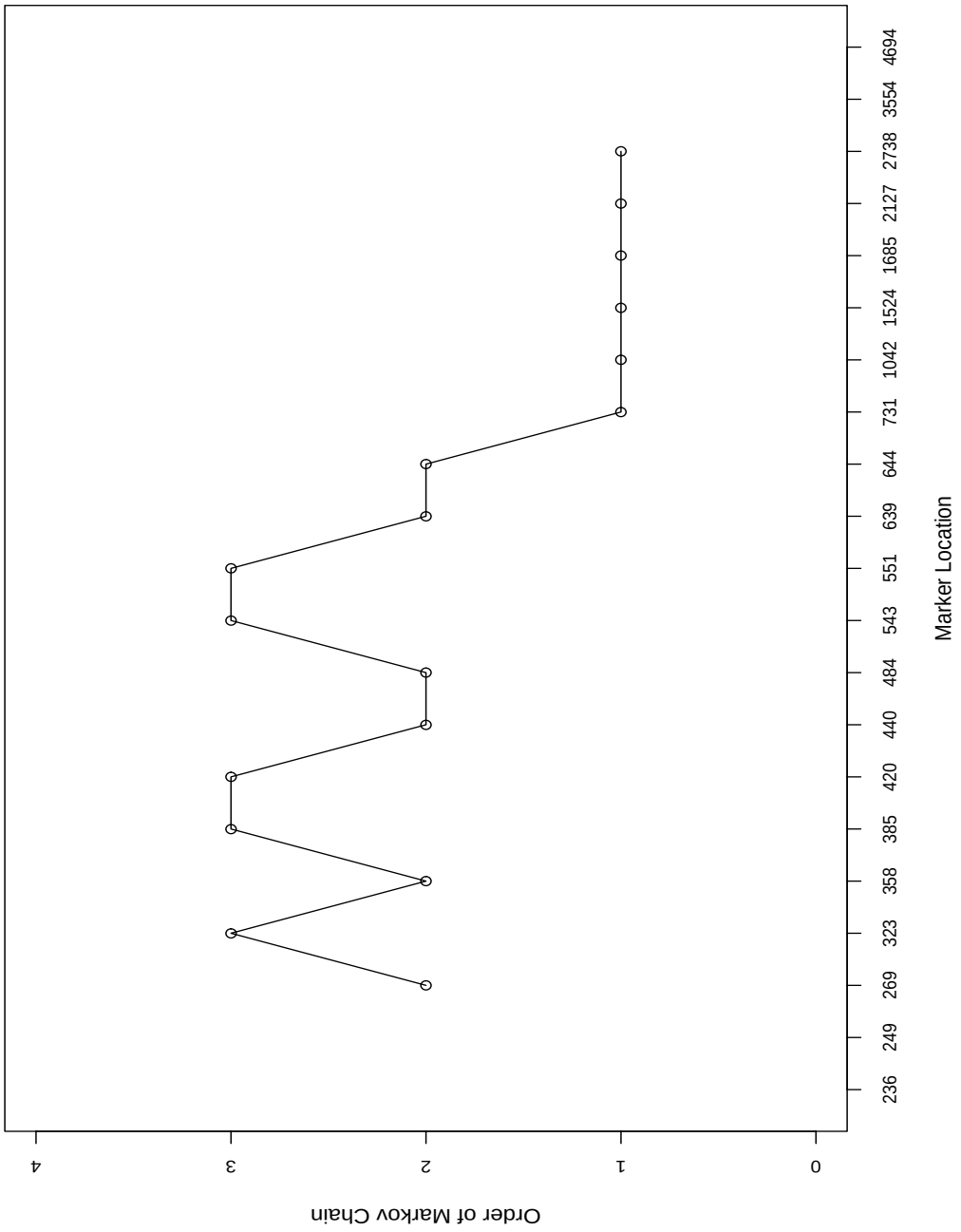


Figure 2.1: The multi-order Markov chain modeling of the 5.5kb D*dc* gene region with 21 binary markers. The promoter region is from position 236 to 731; the remainder gene region is from position 731 to 4694.

Raleigh population. As in many other statistical methods, the markers with minor allele frequency less than 0.05 (15 markers) are not used in the analysis. The locations of the remaining 21 markers are shown in Figure 2.1. Among the 21 markers, 14 are in the promoter region (236-731bp) and 7 are in the remainder of the gene region (1042-4694bp). Markov chain models with order from 0 to 4 are applied to fit the data. Since the detailed local LD information is desired in this study, the window size is set to be 5 (the smallest possible for the $MC4$ model). So, in this example, m=21, ω=5, r=4. In Figure 2.1, the first point at the (269, 2) indicates that, in the first window containing 5 markers at location (236, 249, 269, 323, 358), centered at location 269, the BIC criterion chooses the $MC2$ model as the best model for this window. As the window moves 1 marker right along the chromosome, to (249, 269, 323, 358, 385), the $MC3$ model fits data best and gives the point (323, 3). And so on ..., until the window reaches the last 5 markers (1685, 2127, 2738, 3554, 4694), where the $MC1$ model fits data best.

Summarizing the LD patterns (see next section for relations between the Markov chain orders and the higher order LD) in this way provides more local information than the popular two dimensional pairwise LD plot (Figure 3.a., DeLuca et al, 2003). In the pairwise LD plot, many significant two-locus LD are detected, but their locations seem to be randomly distributed in the two dimensional plot. In Figure 2.1, it is obvious that the LD patterns change dramatically from the promoter region (before locus 731) to the remainder gene region (after locus 731). This result may suggest that, in the D*dc* gene, fewer recombination events have occurred in the promoter region than in the coding region (see the example 1 in Application).

In applying this Markov chain approach, the computation is simple since the likelihood is easy to calculate. Also, we do not need to estimate (or to approximate) the sample variance of the higher order LD, which is hard to obtain. In the following analysis, the maximum order of Markov chain is set

to be 4. Theoretically there is no such restriction (we may set $\omega = m = 21$, and fit MC20 model to the data). But in real data analysis, limited by the number of observed haplotypes in a finite sample, a large number of parameters will either be 0, or not be estimable if the order of Markov chain model is too high. To our experience, we have not observed any examples in which $MC5$ model or higher (i.e., six-locus LD) is selected by BIC, either in real data or simulated data. So in practice, we suggest to restrict the maximum order of Markov chain to be 4 to avoid the non-estimable problems and to save computing time.

2.2.3 Explanation of the results of the multi-order Markov chain modeling in terms of LD measures

The interpretation of the results of the Markov chain modeling is not trivial. For example, if the result shows that the $MC2$ model is a better model than the $MC1$ model in a certain window, what do we learn from it? Since the $MC2$ model characterizes the dependence among three markers, intuitively the result may suggest that in average the LD extends to three markers in this window. But the explicit patterns on how the three markers are correlated are still not clear.

In order to clarify the specific correlation patters embodied in each order of MC model, the connection between the parameters of the MC models and the multiple-locus LD measures need to be established. Since both are functions of hyplotype frequencies, the connection can be built by re-parameterizations.

When two MC models of different orders are compared, the higher order MC model with more parameters may be considered as the "full" model and the lower order MC model as the "reduced" model. In general, the difference in the number of parameters may indicate the number of constraints on the

full model parameters. For example, consider the simplest case where the $MC0$ model and the $MC1$ models are compared within a window of size 2 containing markers M_1, M_2. Since the $MC1$ model has 3 parameters and the $MC0$ model has 2, only one constraint is expected. For MC models, how the lower order MC model is nested in the higher order MC model is not clear, since the parameters in each order of MC model are different. In order to determine this constraint, the 3 parameters, P_1, P_2 and γ_{12} are expressed in terms of the 2 $MC0$ model parameters P_1 and P_2. It is obvious that, the constraint is $\gamma_{12} = 0$. This is the interpretation of comparing the $MC0$ model with the $MC1$ model in this two-marker-window.

Now consider a more complicated example. In a window of size 3 with markers M_1, M_2 and M_3, what are we examining when we compare the reduced $MC1$ model with the full $MC2$ model?

For a $MC1$ model, there are 5 parameters (P_1, $P(M_2 = 1|M_1 = 1) = P^{(2,1)}_{(1,1)}$, $P(M_2 = 0|M_1 = 0) = P^{(2,1)}_{(0,0)}$, $P(M_3 = 1|M_2 = 1) = P^{(3,2)}_{(1,1)}$ and $P(M_3 = 0|M_2 = 0) = P^{(3,2)}_{(0,0)}$. To fully model 3-biallelic markers, 7 parameters are necessary: the 3 parameters of single marker alleles, P_1, P_2, P_3; the 3 pair-wise LD parameters, D_{12}, D_{23} and D_{13}; and the 1 parameter of 3-locus LD, D_{123}. Note that instead of γ, D is used here since the reparameterization in D is easier.

The 7 parameters can be expressed with the 5 $MC1$ model parameters:

$P_1 = P_1$;

$P_2 = P_1 P^{(2,1)}_{(1,1)} + (1 - P_1)(1 - P^{(2,1)}_{(0,0)})$;

$P_3 = P_2 P^{(3,2)}_{(1,1)} + [P_1(1 - P^{(2,1)}_{(1,1)}) + (1 - P_1)P^{(2,1)}_{(0,0)}](1 - P^{(2,1)}_{(0,0)})$;

$D_{12} = P_1(P^{(2,1)}_{(1,1)} - P_2)$;

$D_{23} = P_2(P^{(3,2)}_{(1,1)} - P_3)$;

$$D_{13} = P_1[P^{(2,1)}_{(1,1)})P^{(3,2)}_{(1,1)} + (1 - P^{(2,1)}_{(0,0)})(1 - P^{(3,2)}_{(0,0)})] - P_1P_3;$$

$$D_{123} = P_1P^{(2,1)}_{(1,1)}P^{(3,2)}_{(1,1)} - P_1D_{23} - P_2D_{13} - P_3D_{12} - P_1P_2P_3.$$

Two constraints on those 7 parameters are expected. Recall that the $Eq.1$ is a property of $MC1$ model, so $\gamma_{13} = \gamma_{12}\gamma_{23}$, or equivalently $D_{13} = D_{12}D_{23}/P_2(1 - P_2)$ should be one constraint. Intuitively, another constraint should be a function containing the three-locus LD, since the $MC2$ model does contain parameters involving all 3 markers while the $MC1$ model does not. Some derivations determine the second constraint: (C2) $D_{123} = (1 - 2P_2)D_{13}$. Both constraints can be easily verified given the reparameterization formula above.

So, (C1) and (C2) are the two constraints in terms of the LD measures for the comparison of the reduced $MC1$ model versus the full $MC2$ model in this 3-marker window.

Some important comments:

(I) The way we search for constraints seems to be somewhat unusual in statistics. Often, for a series of nested models, the difference in the parameterizations between the reduced model and the full model is obvious, which indicates the contents and the interpretations of the comparison. For the multi-order Markov chain models, however, the nested structure is ambiguous. So, to explain the comparison results, the Markov chain parameters are re-parameterized by the LD measures. Thus the comparison between MC models can be interpreted by the constraints on the LD measures. In the $MC1$ *vs.* $MC2$ model example, it is clear that by comparing $MC1$ model with $MC2$ model, we are simultaneously examining if the three-locus LD (D^*_{123}, see comment (II)) is significantly different from 0, and if the multiplicable property ($Eq.1$) holds.

(II) One definition of D_{123} is (Bennett, 1954):

$$D_{123} = P_{123} - P_1 D_{23} - P_2 D_{13} - P_3 D_{12} - P_1 P_2 P_3.$$

It is basically the statistical covariance of the 3 markers M_1, M_2, M_3 (Wang, 2001). For the 3 markers, if the $MC1$ model is preferred over the $MC2$ model, it makes intuitive sense that the three-locus LD is absent. However, we observe that $D_{123} = (1-2P_2)D_{13} \neq 0$. If we let: $D^*_{123} = D_{123} - (1-2P_2)D_{13}$, then the constraint C2 is simplified as: $D^*_{123} = 0$. The quantity D^*_{123} can be thought as a slightly modified version of D_{123}, i.e., $D^*_{123} = P_{123} - P_1 D_{23} - (1-P_2)D_{13} - P_3 D_{12} - P_1 P_2 P_3$. One interesting observation is that, D^*_{123} is sensitive to the marker order but D_{123} is not, i.e., $D_{123} = D_{132} = D_{213}$, but generally $D^*_{123} \neq D^*_{132} \neq D^*_{213}$. Note that this is not a problem for two-locus LD. However, since in almost all datasets, the order of the markers is known, D^*_{123} may have advantage by utilizing more information than D_{123}. Nevertheless, expressed in D^*_{123}, the constraint C2 is now a simple null hypothesis, which allows us to further test it by a likelihood test (explained later).

(III) The window size ω also determines the constraints. The window size effect has been briefly discussed in the last section. Now let us take a deeper look. All results derived above assume a small window, i.e., to compare $MC1$ model vs. $MC2$ model, ω is set to be 3, so that the $MC2$ model is the full model. The question is: if the window size ω is set large, what are the changes in the constraints? As a simple example, consider $\omega = 4$, the $MC0$ model and the $MC1$ model are compared. From Table 2.1, in this window the $MC1$ model has three more parameters than the $MC0$ model. It is easy to check that the three constraints are: $\gamma_{12} = \gamma_{23} = \gamma_{34} = 0$. Similarly, if the $MC1$ model and the $MC2$ model are compared in this window, the 4 constraints are: $\gamma_{13} = \gamma_{12}\gamma_{23}$, $\gamma_{24} = \gamma_{23}\gamma_{34}$, $D^*_{123} = 0$, $D^*_{234} = 0$. From this result, we see the explicit forms of the constraints have not been changed, but the number of constraints increases, corresponding to the addition of more markers in the window.

(IV) For comparisons between higher order Markov chain models, a systematic search for the constraints has not been developed yet. There are two difficulties. First, the difference in the number of parameters increases rapidly as the order of Markov chain increases; second, the definition of higher order D^* is unknown. For example, to compare the $MC3$ model (23 parameters) with the $MC4$ model (31 parameters) given $\omega = 5$, there will be 8 constraints (Table 2.1). It is reasonable to guess that the adjusted five-locus LD $D^*_{12345} = 0$ (of course $D_{12345} \neq 0$) is one of the 8 constraints, but the mathematical formulation of D^*_{12345} is not available yet. However, if the BIC prefers the reduced $MC3$ model than the full $MC4$ model, then it may be reasonable to conclude that D^*_{12345} is not significant. In general, it makes sense that the order of the Markov chain (say, r) represents the order of the LD, since both model the dependence among $(r + 1)$ markers.

(V) In all discussion so far, the model comparison is based on the BIC criterion. To test individual constraints of interest, statistical tests need to be developed.

2.2.4 Test $Eq.1$ and construction of the LD map

As discussed previously, the multiplicative correlation pattern $Eq.1$ is a property of the $MC1$ model. It is a focus of our study. We want to find the sub-regions of the chromosome where $Eq.1$ holds.

A simple strategy is to compare the $MC1$ model with the $MC2$ model, since $Eq.1$ is a constraint involved in this comparison. However, $Eq.1$ is entangled with another constraint of the three-locus LD. These two constraints are not independent. To test $Eq.1$, we develop a two-step test following a logic order. In the first step, we test the three-locus LD ($H_0 : D^*_{123} = 0$). If the test result is not significant, we further test the $Eq.1$ ($H_0 : \gamma_{13} = \gamma_{12}\gamma_{23}$).

Specifically, the likelihood of both $MC1$ and $MC2$ models, which are orig-

inally expressed in Markov chain parameters, need to be re-parameterized with the 7 parameters $(P_1, P_2, P_3, D_{12}, D_{23}, D_{13}, D^*_{123})$. Again, the haplotype frequencies are used as a bridge for the re-parameterizations. Technical details are available in Appendix 2.1. The two likelihoods are denoted as $L(MC1)$ and $L(MC2)$. A third likelihood for an "intermediate" model, the $MC2$ model with the constraint (C2) $D^*_{123} = 0$, can be calculated similarly as $L(MC2)$ but fixing $D^*_{123} = 0$. This likelihood is denoted as $L(MC2+C2)$.

In the first step, the null hypothesis $H_0 : D^*_{123} = 0$ (constraint C2) is tested against the alternative $H_a : D^*_{123} \neq 0$ by a likelihood ratio test. A large difference between the full model likelihood $L(MC2)$ and the intermediate model likelihood $L(MC2 + C2)$ rejects the null hypothesis and claims the three-locus LD is significant. If this is the case, we stop here since to further test $Eq.1$ seems to be meaningless. If the null hypothesis is not rejected, we then continue to the second step.

In the second step, $Eq.1$ is tested by a likelihood test as well. Given $D^*_{123} = 0$, the null hypothesis is $H_0 : \gamma_{13} = \gamma_{12}\gamma_{23}$, which is tested against the alternative hypothesis $H_a : \gamma_{13} \neq \gamma_{12}\gamma_{23}$. A large difference between $L(MC2 + C2)$ and $L(MC1)$ rejects the null hypothesis.

The test statistic is $X^2 = -2\log(\Delta)$ in both tests, where Δ is the ratio of the two likelihoods, one under the H_0 and the other under the H_a. Asymptotically, X^2 follows Chi-square distribution with degree of freedom 1.

These two tests are performed within this 3-locus window. As the window moves along the chromosome, a series of p-values are obtained for each test from each window (a total of $(m - 2)$ windows). The false discovery rate (FDR) procedure may be used to adjust the multiple test problem (Westfall & Young, 1996; Storey et al, 2003). Alternatively, we may simply report the individual p-value for each window, and set some critical lines to measure

Table 2.2: Testing the Two Constraints in $MC1$ Model for the D*dc* data

Window	$logL(MC1)$	$logL(MC2)$	$logL(MC2+C2)$	P1(C2)	P2(C1|C2)
1	−200.39	−200.07	−200.14	0.796	0.619
2	−296.06	−272.11	−282.87	0.001	0.000
3	−245.50	−243.87	−245.29	0.234	0.650
4	−267.92	−262.68	−261.38	1.000	–
5	−262.65	−248.90	−260.03	0.001	–
6	−236.33	−229.18	−238.96	0.002	–
7	−200.68	−158.58	−239.70	0.000	–
8	−238.86	−237.90	−238.93	0.310	0.000
9	−281.07	−247.74	−271.86	0.000	–
10	−280.40	−264.07	−268.90	0.028	0.001
11	−301.41	−274.75	−291.95	0.000	–
12	−286.01	−285.04	−287.23	0.139	0.000
13	−297.38	−293.12	−297.29	0.042	0.771
14	−317.48	−315.05	−317.63	0.108	0.000
15	−289.32	−288.79	−288.79	0.962	0.468
16	−260.99	−259.73	−259.73	0.959	0.260
17	−304.95	−303.80	−303.90	0.750	0.307
18	−293.66	−292.35	−292.59	0.621	0.301
19	−307.79	−299.89	−306.40	0.011	0.238

the degree of significance level.

The *Drosophila* dataset (DeLuca et al, 2003) is used again to illustrate this test scheme. The results are given in Table 2.2 (refer to Figure 2.1 for marker location).

In Table 2.2, the likelihood for the $MC1$, $MC2$, $MC2+C2$ models are given in column 2, 3 and 4. The p-values of the Chi-square tests of C2 and (C1|C2) are listed in column 5 and 6. The degrees of freedom for each test are 1. Note that, if the test of C2 is significant, then the p-value in the test of (C1|C2) is meaningless and is set to be 0 (meaning $Eq.1$ does not hold). Table 2.2 provides similar results as those from Graph 2.1. In the D*dc* gene, the Markov chain property $Eq.1$ holds in the gene coding region, but not in

the promoter region.

We expect to see $L(MC2) \geqslant L(MC2+C2) \geqslant L(MC1)$. But in practice, while $L(MC2)$ is always greater than $L(MC1)$, $L(MC2+C2)$ is not always within the interval $(L(MC1), L(MC2))$. This happens because, (1)sometime $L(MC1)$ and $L(MC2)$ are too close to each other, $L(MC2+C2)$ may slightly cross the boundary of the interval due to rounding errors; (2) in some windows, D^*_{123} is very significant. But, to calculate $L(MC2+C2)$, D^*_{123} is forced to be 0. In this case some related probability parameters in the likelihood are affected and take values out of the parameter space [0,1]. Adjustments have to be made (i.e., if a probability < 0, then it is set to be 0), which makes the likelihood estimates unstable. Fortunately, this problem happens only when $D^*_{123} \neq 0$, so it is not serious. Because when $D^*_{123} \neq 0$ we do not test $Eq.1$, our main focus, anyway.

Figure 2.2 shows the LD map for the 5.5 kb D*dc* gene region with 21 binary markers. Some basic information, such as the physical location and the major allele frequency of each marker, is recorded in the LD map. The significance level of the two-locus LD is expressed by different colors. The map distance is measured by the LD distance (LDD), which is defined as the negative logarithm of the LD (expressed in γ) between a pair of adjacent markers. Whether the map distance are additive is indicated by the green lines. In Table 2.2, the Chi-square test shows that $Eq.1$ holds in window 1 and 3, but not in window 2. So a dashed green line is used to cover the three windows indicating there are some, but not strong, evidences suggesting the map distance is additive in this region. $Eq.1$ holds consistently from window 15 to window 19, so a solid green line is used to cover this region.

The three-locus LD, four-locus LD information may also be included in the LD map, if the window size is set to be large. But note that, for those higher order LD, the significance level is not available since the Chi-square test has not been developed.

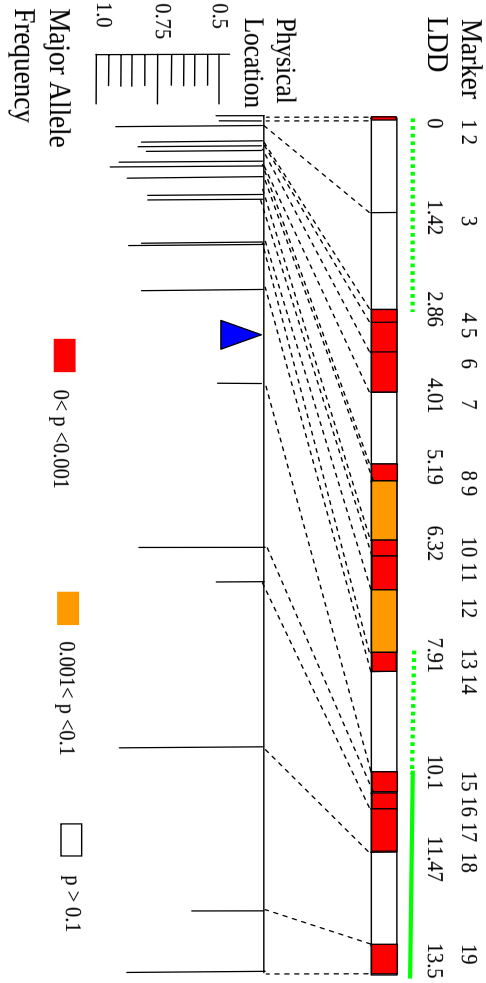


Figure 2.2: The LD map of the 5.5kb D*dc* gene region with 21 binary markers.

2.2.5 Some applications

So far we have illustrated how to use the multi-order Markov chain model to summarize the higher order LD, how to test the Markov chain property $Eq.1$ and to construct the LD map. Now, with some examples, we will show the proposed method is a useful tool and may be applied in many studies where LD is the main focus.

Example 1. Studies on some evolutionary factors.

Population genetics plays a primary role in the LD mapping studies, since LD is created during the long evolutionary history by many dynamic biological processes. It is interesting to investigate how these biological processes influence the correlation patterns among multiple markers. There have been numerous studies on this topic, but most of them focus on the two-locus LD, with only a few on higher order LD (e.g., Hastings, 1983). Since we show that the higher order LD can be reasonably well summarized by the multi-order Markov chain models, we are now able to extend the investigation to higher order LD. We performed a simple coalescence simulation to generate data under the Fisher-Wright model and under different scenarios (mutation rates and recombination rates). The coalescence simulation is done by using the MS program provided by Hudson (2003). Once the data are generated, the multi-order Markov chain models are applied to fit each data set. The order of the Markov chain model selected for each scenario represents the complexity of the correlation patterns under this specific scenario.

Again, the goal of this simulation is not to study the evolutionary history by examining multiple evolutionary factors. Instead, we want to show that, the proposed Markov chain model is a reasonably good statistical tool for these kinds of studies.

Simulation design:

Table 2.3: The Values of the Recombination Rate and Mutation Rate

Parameter Design (4Ne.ρ.n, 4Ne.c.n)		Mutation Rate ρ (/adjacent sites/generation)		
		$\rho : 5 \times 10^{-10}$	$\rho : 2 \times 10^{-9}$	$\rho : 10^{-8}$
Recombination Rate c (/site/generation)	1×10^{-11}	$(10, 0.2)$	$(40, 0.2)$	$(200, 0.2)$
	4×10^{-11}	$(10, 0.8)$	$(40, 0.8)$	$(200, 0.8)$
	1×10^{-10}	$(10, 2)$	$(40, 2)$	$(200, 2)$
	1×10^{-9}	$(10, 20)$	$(40, 20)$	$(200, 20)$
	1×10^{-8}	$(10, 200)$	$(40, 200)$	$(200, 200)$
	3×10^{-8}	$(10, 600)$	$(40, 600)$	$(200, 600)$
	1×10^{-7}	$(10, 2000)$	$(40, 2000)$	$(200, 2000)$

1. DNA segments: total number of nucleotides n = 500 kb. Based on different mutation rate, hundreds of SNP will be generated for the DNA segments.

2. Ne: the effective population size is set to be 10000;

3. 21 scenarios: The range of the parameter values (mutation and recombination) are set to be wide. The values are shown in Table 3:

4. For each of the 21 scenarios, 10 SNP data sets are created.

There are 500 independent chromosomes in each data set. 5 Markov Chain Models (order = 0, 1, 2, 3, 4) are fitted to each data set. The window size is set to be the whole 500 kb region. The likelihood and BIC of each model are calculated in each dataset.

5. The simulation result:

In Figure 2.3, the influence of the recombination rate on LD patterns is shown. The mutation rate is fixed at 5×10^{-10} mutation events/site/generation, while the recombination rate varies from 10^{-11} (low) to 10^{-7} (high) recombination events/adjacent base pair/per generation.

The results show that the recombination does influence the multiple correlation patterns. Besides the common knowledge that the two-locus LD

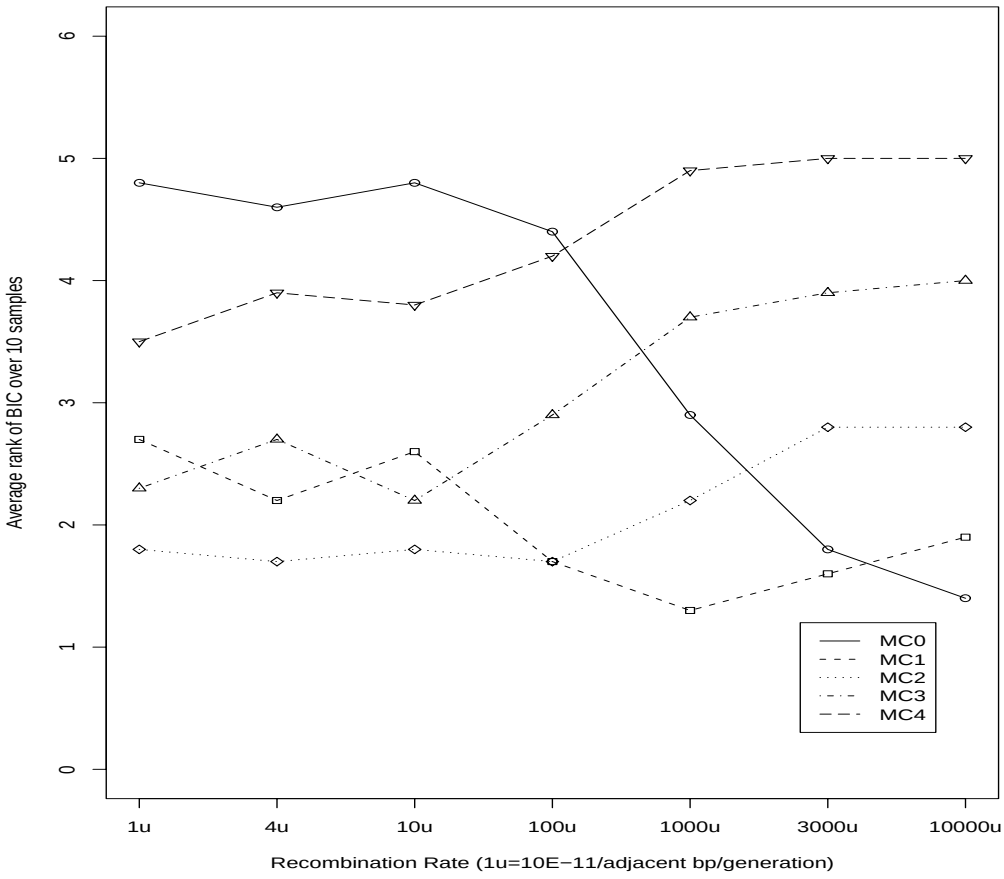


Figure 2.3: The influence of the recombination rate on LD patterns. The X-axis: the range of recombination rates. The Y-axis: the rank of BIC, averaged over 10 replication samples. Specifically, for each of the 10 dataset in each scenario, the 5 MC models are ranked by BIC criterion. The MC model with the smallest BIC is ranked 1; the model with the largest BIC is ranked 5. For the scenario, the ranking of each MC model is averaged over the 10 replication samples.

decreases over physical distance because of recombination, this simulation further implies the complexity of the LD patterns decreases as the recombination rates increases. In Figure 2.3, to better view the results, the X-axis can be divided into 3 sections. In section 1, recombination rates are low, ranging from 1u to 10u (1u = 10^{-11} recombination event/adjacent sites/generation). The $MC1$, $MC2$, $MC3$ models are approximately equally likely to be selected by BIC, with $MC2$ model slightly preferred than the other two models. At these low recombination rates, the $MC0$ model does not fit data well, while the $MC4$ model over-fits the data. As the recombination rate gets higher, the performance of the lower order Markov models, e.g., the $MC0$ and the $MC1$ models, are improved. As an example, in the second section, from 100u to 3000u (10^{-9} to 3×10^{-8}/adjacent sites/generation), the $MC1$ model fit the data best. In the third section, where the recombination rate is larger than 3000u (3×10^{-8}/adjacent sites/generation), the $MC0$ model is the best model selected, which indicates that too many recombination events break up almost any correlations between markers. At 10000u (10^{-7}/adjacent sites/generation), we observe that the 5 MC models ranked by their orders, i.e., the $MC0$ model is the best, the $MC1$ model is the second best, ..., the $MC4$ model is the least preferred model by BIC. At this high recombination rate, the likelihood for all 5 MC models are almost same. BIC selects the best model with fewest parameters.

Based on this coalescent simulation, the changes of the LD patterns over recombination rates are similar at different mutation rates. Thus results at the other two mutations rates are not shown.

This simulation suggests that, by summarizing the higher order LD patterns, the multi-order Markov chain models may reflect the changes in some evolutionary factors, e.g., the recombination rate. This feature makes this method a useful tool in some population genetics studies. For example, if the researchers want to compare LD patterns among multiple populations, it

may be helpful to apply the multi-order Markov chain method and compare the higher order LD in different populations.

Another interesting observation is that, in this simulation, overall the lower order MC models seem to fit the data quite well. In the full range of recombination rate, the $MC0, MC1, MC2$ models compete for the best model. It is especially interesting that the $MC1$ model performs well as the recombination rate is between 10^{-11} and 3×10^{-8} recombination event/adjacent sites/generation, which is so wide that this interval covers many estimated recombination rates reported from various human genetic studies. This observation indicates that the LD map may be useful in practice.

Example 2. Haplotype blocks

In recent years, the studies on haplotype block structures have attracted much attention. Basically the haplotype blocks can be viewed as some specific types of multiple correlation patterns over multiple markers. Since the proposed Markov chain models summarize higher order LD patterns, it is interesting to see whether the haplotype block structure can be captured by this multiple-order Markov chain model.

We investigate the dataset reported by Daly, et al (2001). In the study on Crohn disease, 129 trios from a European-derived population are genotyped. The whole DNA region under study is 500kb, with 103 common SNP markers (minor allele frequency $> 5\%$). Daly et al show that 93 SNP markers are clustered into 11 haplotype blocks. Within each block, the observed number of different haplotypes is strikingly less than the number of all possible hyplotypes, e.g., in the longest block 7 containing 31 markers, more than 90% of all chromosomes only have 4 haplotypes, comparing to the total 2^{31} possible haplotypes. We choose this block 7 to fit the multi-order Markov chain model. A total of 258 31bp DNA sequences (haplotype) data are generated by directly using their results, i.e., the percentages of the 4 haplotypes were

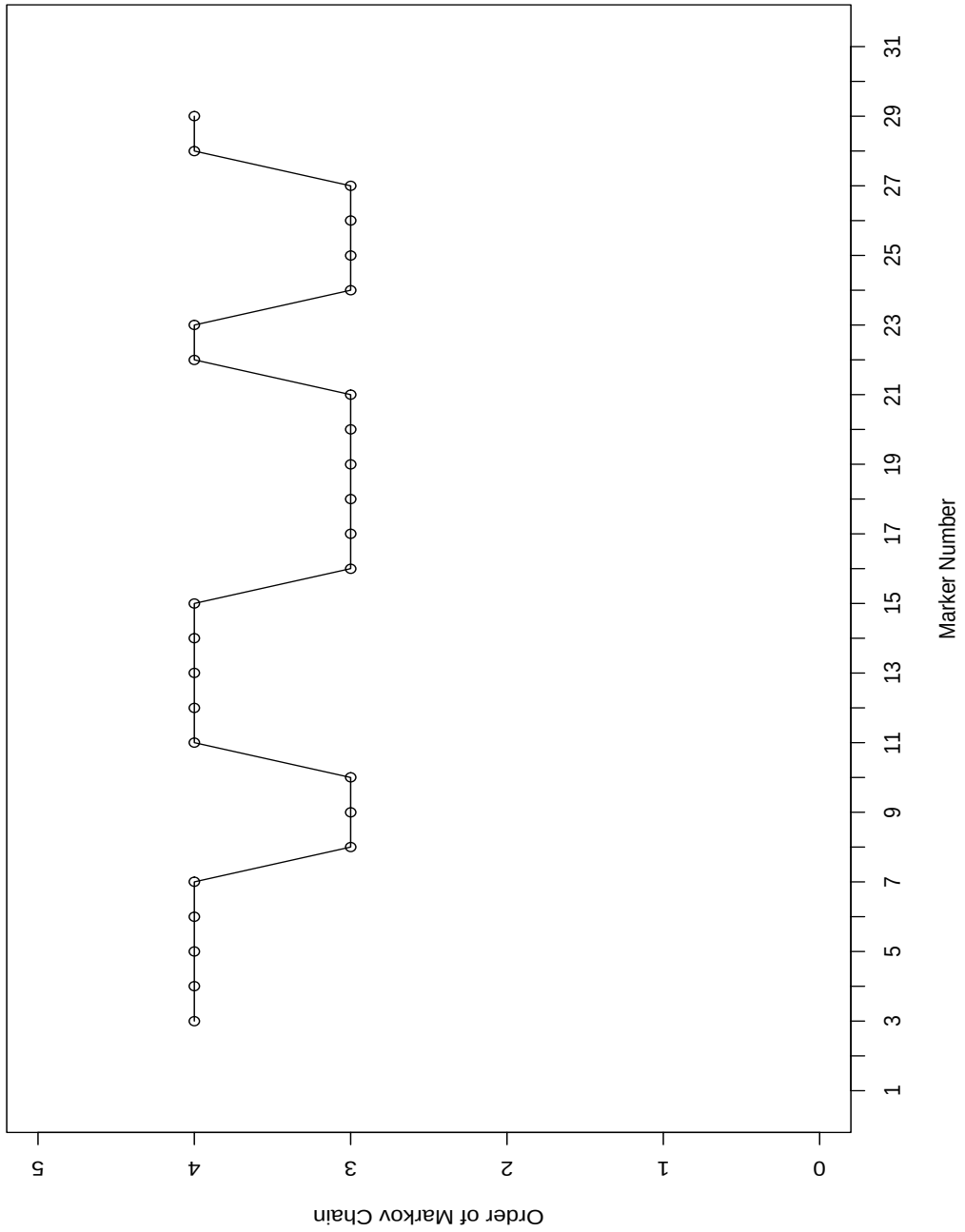


Figure 2.4: The multi-order Markov chain model fitting on the haplotype block 7 (Daly, et al, 2001). 31 SNP markers are located in the block.

reported as 40%, 14%, 21% and 12%, respectively.

Markov chain models with orders up to 4 are applied to fit the data and the window size is set to be 5. Figure 2.4 shows the model fitting results. It is obvious that higher order MC models (e.g., the $MC3$ and $MC4$ models) are preferred in this haplotype block. Since the data of the markers between adjacent haplotype blocks are not available, we do not know for sure if the LD pattern within this block is significantly different than that outside the block. But the difference is expected. So, the proposed method may be used to detect haplotype blocks structures.

2.3 Discussion

Li and Stephens (2003) wrote that:"...most current methods for interpreting and analyzing patterns of LD suffer from at least one of the following limitations:

1. They are based on computing some measure of LD defined only for pairs of sites, rather than considering all sites simultaneously.

2. They assume a 'block-like' structure for patterns of LD, which may not be appropriate at all loci.

3. They do not directly relate patterns of LD to biological mechanisms of interest, such as the underlying recombination rate. ..."

We can not agree more on those comments. If we take those three comments as three criteria to evaluate the proposed method, obviously, our multi-order Markov chain method with the LD map do not have the first two problems, but have the third one.

We believe that the third problem is very important. To relate the patterns of LD to biological mechanisms helps us to understand the insight of the model with underlying biological process. However, as we already discussed briefly before, the LD is created confoundedly by many evolution factors. It is almost impossible, and maybe inappropriate, to model LD with certain simplified biological assumptions in a natural population of unrelated individuals without a clear understanding of the history of the population for the genome region. This is the reason we do not model LD with biological processes in this study, but only characterize the higher order LD patterns. Thus the results, represented by the order of Markov chain and the LD map along the chromosome, are just statistical descriptions of correlations among markers, similar to the LD measures themselves. If the experiments are carefully designed with direct measures on certain evolution factors, e.g., the sperm typing experiment introduced in Li and Stephens (2003) directly measures the recombination rate and site, the biological modeling of LD is thus feasible and appears to be appropriate.

The proposed LD map, denoted here as the LDγ-map for comparison purposes, is different from the LD map introduced by Morton's group, denoted as the LDρ-map, which is based on a specific LD measure ρ and the Malecot equation (Maniatis et al, 2002; Zhang, W. et al, 2002 (a) (b); Lonjou et al, 2003). Some major differences are:

(1) The LDγ-map is constructed on the commonly used LD measure γ; the LDρ-map is constructed on a specific LD measure ρ. Since *Eq.*1 is only based on γ, not ρ, so the method to construct the LDγ-map is not applicable for the LDρ-map. Similarly, γ should not be used to construct the LDρ-map due to the bad model fitting result (Morton, et al, 2001).

(2) The LDρ-map only uses two-locus LD information; the LDγ-map also characterizes higher order LD. Thus, the LDγ-map is more informative.

2.4 Future Work

The next step is to use the LD map in the LD mapping studies. A likelihood based mapping method has been developed by our group (Wang, et al, 2001). In this method, the EM algorithm is applied and a closed form solution of the maximum likelihood estimates is derived. An immediate extension is to integrate the LD background information in the likelihood framework. It is expected that, in the chromosome region where the $MC1$ property $Eq.1$ holds, the unknown genetic variant can be located on the LD map. Some well developed statistical methods in linkage analysis, such as interval mapping, composite interval mapping and multiple interval mapping may be adopted in LD mapping studies.

However, in a region where higher order LD exists, the problem is still unsolved. It appears that the haplotype analysis may be an appropriate approach in those regions. So, it is tempting to build a model which combines the single marker analysis and haplotype analysis, depending on the local LD patterns. Obviously, more theoretical development is needed.

Chapter 3

Knowledge Based Shrinkage Estimation: Integrating Prior Mean and Covariance Information in a Least Square Framework

3.1 Introduction

In biological data analysis, the mechanism by which we believe the data are generated may be represented by a statistical model. Ideally, the parameters in a parametric statistical model would have clear biological interpretations and quantify the key characteristics of the underlying biological process. In practice, sometimes researchers may have certain knowledge about some of these characteristics. The knowledge is often of biological importance and is independent to the experiments from which the data are collected. It is thus preferred that the "prior" knowledge can be integrated into data analysis to obtain more accurate and precise results.

The Bayesian statistics are popular tools for this purpose. In a typical Bayesian analysis, the independent knowledge is modeled by a known statis-

tical distribution, i.e., the prior distribution. Based on the Bayesian rules, the posterior distribution of the parameters can be derived by combining both the prior information and the data information.

A critical step in Bayesian analysis is to choose a proper prior distribution. Depending on the nature of the prior knowledge, not always are we able to find a proper known distribution to summarize the information. An example is, the prior information is given in the form of some moment statistics (mean and/or covariance) of the parameters. This situation is not rare in research. For instance, in a genetic network study containing hundreds of genes, usually the correlation between a few specific genes are known. This information, independent of data and in the form of the second moment statistics, is the basis for future investigations. More generally, prior knowledge often comes from prior scientific studies. In those studies many results have been reported, hence only available, in summary moment statistics. In statistical consulting, biologists seem to be more confident to talk about the mean and/or covariance information of the parameters, rather than to specify a certain joint prior distribution. The underlying philosophy is, the limited understanding of the complicated biological system is only capable to provide limited amount of prior information, in simple and conservative forms. The common Bayesian practice for this problem is to assign a certain symmetric distribution, e.g., the normal distribution, with mean and variance fixed at the prior values (personal communication with Dr.Ghosh). Theoretically, a statistical distribution contains all (of infinity) moment information. By specifying a distribution, this Bayesian approach fixes not only the first two moments, but also all higher order moments. This may somehow exaggerate the amount of prior information, though in practice the extra information (the higher order moments) may not necessarily ruin results.

In this chapter, we will introduce a simple point shrinkage approach (PSA), which has the capability of integrating prior information provided

in forms of mean and covariance. Similar to other shrinkage methods, such as ridge regression (Hoerl and Kennard, 1970) and the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), PSA is developed in a least square framework. It trades off decreased variance for possibly increased bias, and hence improves prediction accuracy under some conditions. The PSA differs from other shrinkage methods in that: aiming for variable selection, other methods shrink the least square estimate towards the origin 0; aiming for integrating prior single point information, the PSA shrinks least square estimate towards the prior single point with its vicinity in parameter space.

The basic idea for the PSA is simple: suppose in a p dimensional parameter space, the prior information suggests that the parameter is around a certain point β_0. We hope to find a point $\hat{\beta}_{PSA}$ within a neighborhood of β_0 with radius r, such that the prediction error is minimized. The estimate $\hat{\beta}_{PSA}$ depends on both β_0 and the data. The estimating process may be understood as seeking a balance point between the prior β_0 and the least square estimate.

3.2 Method

3.2.1 Review of The Least Square, The Ridge regression and The LASSO estimators:

Consider a linear regression setting, where there are N independent observations $Y = (Y_1, Y_2, ..., Y_N)'$ of the response and p variables $X = (X_1, X_2, ..., X_p)$, where each X_j is a $(N \times 1)$ vector, so X is $(N \times p)$. Y has mean $X\beta$ and covariance matrix $\sigma^2 I_N$. β is a $(p \times 1)$ vector of unknown constants and σ^2 is an unknown positive constant.

Developed by Gauss, the least square (LS) approach minimizes the objective function

$$g(\beta) = \sum_{i=1}^{N}(Y_i - \sum_{j=1}^{p}\beta_j X_{ij})^2,$$

with solution: $\hat{\beta}_{LS} = (X'X)^{-1}X'Y$. This is an unbiased estimator with covariance matrix $V(X) = (X'X)^{-1}\sigma^2$.

The Ridge regression and the LASSO are developed by adding some constrains on $g(\beta)$.

The ridge regression minimizes

$$g(\beta) = \sum_{i=1}^{N}(Y_i - \sum_{j=1}^{p}\beta_j X_{ij})^2, \quad subject\ to\ \sum_{j=1}^{p}\beta_j^2 \leq r;$$

or equivalently (Murray et al. 1981), minimizes

$$q(\beta) = \sum_{i=1}^{N}(Y_i - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2,$$

where λ can be expressed as a function of $r(i.e., \lambda \approx 1/r)$.

The ridge regression estimator is $\hat{\beta}_R = (X'X + \lambda I_p)^{-1}X'Y$, with bias $bias(\hat{\beta}_R(\lambda)) = \lambda(X'X+\lambda I_p)^{-1}\beta$ and covariance matrix $V(\hat{\beta}_R(\lambda)) = X'X(X'X + \lambda I_p)^{-2}\sigma^2$. If the input in X matrix is ortho-normal, where $X'X = I_p$, then $\hat{\beta}_R = (\frac{1}{1+\lambda})\hat{\beta}_{LS}$, indicating the least square estimate is shrunk towards 0 as λ increases (if X is not ortho-normal, the shrinkage will have a complex form, see Hoerl and Kennard, 1970). The tuning parameter λ (or r) controls the depth of the shrinkage. Note that, if we re-write the constrain as: $\sum_{j=1}^{p}(\beta_j - 0)^2 \leq r$, then $\hat{\beta}_R$ may be interpreted as a point in the neighborhood of origin 0, with radius $\sqrt{r}$, at which $g(\beta)$ is minimized.

Similarly, the LASSO minimizes

$$g(\beta) = \sum_{i=1}^{N}(Y_i - \sum_{j=1}^{p}\beta_j X_{ij})^2, \quad subject\ to\ \sum_{j=1}^{p}|\beta_j| \leq r;$$

or equivalently, minimizes

$$q(\beta) = \sum_{i=1}^{N}(Y_i - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|.$$

The LASSO parameter estimator is $\hat{\beta}_{LASSO} = sign(\beta_{LS})'(|\hat{\beta}_{LS}| - \gamma)^+$, where "sign" is a $(p \times 1)$ vector of signs (+ or −) of the least square estimate $\hat{\beta}_{LS}$; γ is a $(p \times 1)$ vector which depends on λ.

The difference between the ridge regression and the LASSO is the shape of the neighborhood of the shrinkage target point 0. For example, in a two-dimensional parameter space, the ridge regression defines a circle around the origin 0; while the LASSO defines a square. The shape of the neighborhood determines the patterns of the shrinkage. For more details, see Frank and Friedman, 1993, and Tibshirani, 1996.

The least square estimator $\hat{\beta}_{LS}$ is known as the "Best Linear Unbiased Estimator" (BLUE). It has the smallest variance among all unbiased estimators. However, when prediction accuracy (usually measured by prediction error (PE) or mean square error (MSE)) is the major consideration, $\hat{\beta}_{LS}$ may not be the best. It can be shown that, though both $\hat{\beta}_R$ and $\hat{\beta}_{LASSO}$ are biased towards the origin 0, they have smaller variance and possibly, smaller prediction error than $\hat{\beta}_{LS}$ (Theobald 1974, Tibshirani 1996).

3.2.2 The point shrinkage approach (PSA)

3.2.2.1 The Model

The PSA minimizes

$$g(\beta) = \sum_{i=1}^{N}(Y_i - \sum_{j=1}^{p}\beta_j X_{ij})^2,$$

within a neighborhood centered at a special point of β_0 in a p-dimensional space. The shape of the neighborhood can be either a circle, i.e., $\sum_{j=1}^{p}(\beta_j -$

$\beta_{0j})^2 \leq r$, or a square, i.e., $\sum_{j=1}^{p} |\beta_j - \beta_{0j}| \leq r$, where β_0 is a known $(p \times 1)$ vector with element $\beta_{0j}, j = 1, 2, ..., p$. The shape is important for determining the pattern of shrinkage, but it is not a focus in this paper. A circle neighborhood is chosen to develop the PSA method.

Under this situation, to minimize $g(\beta)$ within the neighborhood of β_0 is equivalently to minimize

$$q(\beta) = \sum_{i=1}^{N}(Y_i - \sum_{j=1}^{p} \beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p}(\beta_j - \beta_{0j})^2, \lambda > 0.$$

The estimator $\hat{\beta}_{PSA}$ can be obtained by differentiating $q(\beta)$ with respect to β and setting the differentiation equation to be 0 and solve:

$$\hat{\beta}_{PSA} = (X'X + \lambda I_p)^{-1}(X'Y + \lambda\beta_0).$$

$\hat{\beta}_{PSA}$ is biased with $bias(\hat{\beta}_{PSA}(\lambda)) = -\lambda(X'X+\lambda I_p)^{-1}(\beta-\beta_0)$ and covariance matrix $V(\hat{\beta}_{PSA}(\lambda)) = X'X(X'X + \lambda I_p)^{-2}\sigma^2$, which is exactly the same as the covariance matrix of the ridge regression estimator. Similar to the ridge regression and the LASSO, the PSA estimator has smaller variance than the least square estimator. Further, if the condition $\lambda < 2\sigma^2/(\beta - \beta_0)'(\beta - \beta_0)$ is satisfied, $\hat{\beta}_{PSA}$ also has smaller prediction error than the least square estimator. The proof is in Appendix 3.1.

When the shrinkage destination is just a point β_0, the PSA is equivalent to the Ridge regression on a transformed scale.

Consider the objective function:

$$q(\beta) = \sum_{i=1}^{N}(Y_i - \sum_{j=1}^{p} \beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p}(\beta_j - \beta_{0j})^2, \lambda > 0.$$

Let $\beta^* = \beta - \beta_0$, then in terms of β^*, the new objective function is:

$$\begin{aligned} h(\beta^*) &= \sum_{i=1}^{N}(Y_i - \sum_{j=1}^{p}((\beta^* + \beta_0)_j)X_{ij})^2 + \lambda\sum_{j=1}^{p}(\beta_j^*)^2 \\ &= \sum_{i=1}^{N}(Y_i^* - \sum_{j=1}^{p}(\beta_j^*)X_{ij})^2 + \lambda\sum_{j=1}^{p}(\beta_j^*)^2, \end{aligned}$$

where $Y_i^* = Y_i - \sum_{j=1}^{p}\beta_0 X_{ij}$.

Note that in terms of Y_i^*, X_{ij} and β^*, $h(\beta^*)$ is just the usual Ridge regression objective function with solution:

$$\hat{\beta}^* = (X'X + \lambda I_p)^{-1}X'Y^*.$$

But $\beta^* = \beta - \beta_0$, so that in terms of $\hat{\beta}$,

$$\beta = (X'X + \lambda I_p)^{-1}X'Y^* + \beta_0 = \hat{\beta}_{PSA}.$$

Thus, many properties and conclusions that are valid in the Ridge regression are also valid in the PSA. The PSA can be thought as an extension to the Ridge regression in that:

(1) the Ridge regression is a special case of PSA when $\beta_0 = 0$;

(2) the PSA can also integrate the second moment information.

Also, the controlling mechanism of the shrinkage method is clearly illustrated in this report. This helps us to understand all shrinkage approaches.

3.2.2.2 Parameter estimation through cross-validation

$\hat{\beta}_{PSA}$ is a function of the tuning parameter λ. An estimate $(\hat{\beta}_{PSA}, \hat{\lambda})$ can be obtained by optimizing the criterion of prediction accuracy, measured by the prediction error, $PE = (Y - \hat{y})^2$; or similarly, the mean square error $MSE = E(\hat{y} - x\beta)^2 = PE - \sigma^2$.

In simulation studies, where β, β_0 and σ^2 are known, an exact real number solution $\hat{\lambda}$ may be obtained by minimizing $MSE(\hat{\beta}_{PSA})$. For real data

analysis, this theoretical solution is not available since β and σ^2 are unknown. The PSA applies the 10 fold cross-validation procedure (Efron and Tibshirani, 1993) to estimate $(\hat{\beta}_{PSA}, \hat{\lambda})$. The PE is calculated over a series of values of λ from a grid search in $(0, \infty)$. The value of λ yielding the lowest estimated PE is taken as the estimate $\hat{\lambda}$, and $\hat{\beta}_{PSA}$ can be estimated given this $\hat{\lambda}$. For technical details and an example of the 10 fold cross-validation, see the simulation study in next section.

3.2.2.3 Some properties of the PSA estimator

(1) The bias

The prior bias is defined as $\tau = \beta - \beta_0$. The posterior bias for $\hat{\beta}_{PSA}$, denoted as $bias(\hat{\beta}_{PSA})$, is a linear function of the tuning parameter λ and the prior bias τ:

$$bias(\hat{\beta}_{PSA}(\lambda, \tau)) = -\lambda(X'X + \lambda I_p)^{-1}\tau.$$

For a fixed τ, if $\lambda \to \infty$, $bias(\hat{\beta}_{PSA}(\lambda, \tau)) \to \tau$; if $\lambda \to 0$, $bias(\hat{\beta}_{PSA}(\lambda, \tau)) \to 0$. When $\lambda = 0$ (back to the least square case), $bias(\hat{\beta}_{PSA}(\lambda, \tau)) = bias(\hat{\beta}_{LS}) = 0$. τ also controls the posterior bias: when τ is small, the posterior bias could not be large; if τ is 0, so is the posterior bias.

This suggests a potential advantage of the PSA procedure: unlike the ridge regression estimator whose posterior bias can not be avoided, the posterior bias of the PSA estimator may be reduced if the prior information is good (τ is small). In such cases, the reduction of variance and MSE of $\hat{\beta}_{PSA}$ is a "net gain".

(2) The tuning parameter λ controls shrinkage

Let $\alpha_{p\times p} = (X'X + \lambda I_p)^{-1}X'X$. Then, $\hat{\beta}_{PSA}$ can be written as:

$$\hat{\beta}_{PSA} = (X'X + \lambda I_p)^{-1}(X'Y + \lambda\beta_0) = \alpha\hat{\beta}_{LS} + (I_p - \alpha)\beta_0.$$

In the simple ortho-normal case, where $\alpha_{p\times p} = (1+\lambda)^{-1} I_p$,

$$\hat{\beta}_{PSA} = \frac{1}{1+\lambda}\hat{\beta}_{LS} + \frac{\lambda}{1+\lambda}\beta_0.$$

So the PSA estimator is basically a weighted average between the prior information (the point β_0) and the least square estimator $\hat{\beta}_{LS}$. Both of the two weights $1/(1+\lambda)$ and $\lambda/(1+\lambda)$ range from 0 to 1, depending on the value of λ. If λ is large, then β_0 gets more weight, $\hat{\beta}_{PSA}$ is shrunk to β_0 and we say the shrinkage is deep; if λ is small, then $\hat{\beta}_{LS}$ gets more weight, $\hat{\beta}_{PSA}$ stays at $\hat{\beta}_{LS}$ and we say the shrinkage is slight.

The parameter λ may also be explained as the proportions of information contributed by prior and data, respectively. A large λ implies the prior is dominating, while a small λ implies the prior is less influential.

(3) Theoretical value of $\hat{\lambda}$

λ is the critical parameter controlling the shrinkage. $\hat{\beta}_{PSA}$ can not be computed without knowing λ. As stated before, in real data analysis, λ is estimated by cross validation. However, to investigate the mechanism of this shrinkage procedure, it is necessary to derive the theoretical solution of λ that minimizes $MSE(\hat{\beta}_{PSA})$.

By definition, we have:

$$\begin{aligned}
MSE(\hat{\beta}_{PSA}) &= E[(\hat{\beta}_{PSA}-\beta)'(\hat{\beta}_{PSA}-\beta)] \\
&= E[(\alpha\hat{\beta}_{LS}-\beta+(I_p-\alpha)\beta_0)'(\alpha\hat{\beta}_{LS}-\beta+(I_p-\alpha)\beta_0)] \\
&= E[(\alpha\hat{\beta}_{LS}+\alpha\beta-\alpha\beta-\beta+(I_p-\alpha)\beta_0)'(\alpha\hat{\beta}_{LS}+\alpha\beta-\alpha\beta \\
&\quad -\beta+(I_p-\alpha)\beta_0)] \\
&= E[(\hat{\beta}_{LS}-\beta)'\alpha'\alpha(\hat{\beta}_{LS}-\beta)]+[\alpha\beta-\beta+(I_p-\alpha)\beta_0)]'[\alpha\beta \\
&\quad -\beta+(I_p-\alpha)\beta_0)] \\
&= \sigma^2 tr[(X'X)^{-1}\alpha'\alpha]+[(\alpha-I_p)(\beta-\beta_0)]'[(\alpha-I_p)(\beta-\beta_0)] \\
&= \sigma^2[tr((X'X)^{-1}+\lambda I_p)^{-1}-\lambda((X'X)^{-1}+\lambda I_p)^{-2}] \\
&\quad +(\beta-\beta_0)'(\alpha-I_p)'(\alpha-I_p)(\beta-\beta_0) \\
&= \sigma^2\sum_{i=1}^{p}\frac{\delta_i}{(\delta_i+\lambda)^2}+(\beta-\beta_0)'((X'X)^{-1}+\lambda I_p)^{-2}(\beta-\beta_0),
\end{aligned}$$

where δ_i is the i^{th} eigenvalue of $X'X$.

In the special case of ortho-normal design, where $X'X = I_p$ and $\delta_i = 1$ for all i,

$$MSE(\hat{\beta}_{PSA}) = \frac{p\sigma^2}{(1+\lambda)^2}+\frac{\lambda^2}{(1+\lambda)^2}(\beta-\beta_0)'(\beta-\beta_0).$$

$\hat{\lambda}$ is solved by taking the derivative of $MSE(\hat{\beta}_{PSA})$ with respect to λ and setting the derivative to be 0:

$$\hat{\lambda} = \sigma^2/[(\beta-\beta_0)'(\beta-\beta_0)/p] \qquad Eq(3.1)$$

The second derivative $(\partial^2 MSE(\hat{\beta}_{PSA})/\partial\lambda^2) > 0$, so $\hat{\lambda}$ minimizes $MSE(\hat{\beta}_{PSA})$.

(4) The proof of $MSE(\hat{\beta}_{PSA}(\hat{\lambda})) < MSE(\hat{\beta}_{LS})$

In the ortho-normal design, $MSE(\hat{\beta}_{LS}) = MSE(\hat{\beta}_{PSA}(\lambda=0)) = p\sigma^2$, we have:

$$MSE(\hat{\beta}_{PSA}(\hat{\lambda}))-MSE(\hat{\beta}_{LS}) = -1/(p\sigma^2+bias^2(\beta_0)) < 0,$$

where $bias^2(\beta_0) = (\beta - \beta_0)'(\beta - \beta_0) = \tau'\tau$. So in this simple case, it is easy to show $MSE(\hat{\beta}_{PSA}(\hat{\lambda})) < MSE(\hat{\beta}_{LS})$, regardless how large the prior bias is.

Actually, it can be shown that the inequality is generally true. A simple proof is given below.

Theorem of Existence:

If $0 < \lambda < c = 2\sigma^2/(\beta-\beta_0)'(\beta-\beta_0)$, then $MSE(\hat{\beta}_{PSA}(\lambda)) < MSE(\hat{\beta}_{LS})$.

The two-step proof is given in Appendix 3.1.

This theorem provides a sufficient, but not a necessary, condition. Based on this theorem, the following corollary is true:

Corollary: If the function $MSE(\hat{\beta}_{PSA}(\lambda))$ can be minimized with respect to $\lambda \in (0, \infty)$ at the point of $\hat{\lambda}$, then: $MSE(\hat{\beta}_{PSA}(\hat{\lambda})) < MSE(\hat{\beta}_{LS})$.

Proof: Let $\lambda_1 \in (0, c)$, where c is a constant and $c = 2\sigma^2/(\beta-\beta_0)'(\beta-\beta_0)$. Since $\hat{\lambda}$ is the global minimizer of $MSE(\hat{\beta}_{PSA}(\lambda))$ in $(0, \infty)$, so $MSE(\hat{\beta}_{PSA}(\hat{\lambda}))$ $\leq MSE(\hat{\beta}_{PSA}(\lambda_1))$. From the Theorem of existence, $MSE(\hat{\beta}_{PSA}(\lambda_1)) <$ $MSE(\hat{\beta}_{LS})$. Hence, $MSE(\hat{\beta}_{PSA}(\hat{\lambda})) \leq MSE(\hat{\beta}_{PSA}(\lambda_1)) < MSE(\hat{\beta}_{LS})$.

(5) How is λ controlled?

Another interesting question is, since the shrinkage is controlled by the tuning parameter λ, how is λ itself controlled? This is the key question to understand the mechanism of the shrinkage.

Still consider the simple ortho-normal condition. From $Eq(3.1)$,

$$\hat{\lambda} = \sigma^2/[(\beta - \beta_0)'(\beta - \beta_0)/p],$$

given a constant variance, if the prior bias $\tau = \beta - \beta_0$ is large, then $\hat{\lambda}$ is

small, resulting in a slight shrinkage. In the extremely bad case, we have:

$$(\beta - \beta_0) \to \infty \Rightarrow \hat{\lambda} \to 0 \Rightarrow \hat{\beta}_{PSA} \to \hat{\beta}_{LS}.$$

In other words, if the prior information is bad, the PSA tends to anchor $\hat{\beta}_{PSA}$ around the unbiased $\hat{\beta}_{LS}$ to avoid being influenced by the bad prior. This is a wonderful result, implying that the estimator $\hat{\beta}_{PSA}$ is self-adjusted and is "safe".

On the other hand, if the prior bias τ is small, then $\hat{\lambda}$ is inflated. This results in a deep shrinkage towards β_0. In the extremely good case, we have:

$$(\beta - \beta_0) \to 0 \Rightarrow \hat{\lambda} \to \infty \Rightarrow \alpha \to 0 \Rightarrow \hat{\beta}_{PSA} = \alpha\hat{\beta}_{LS} + (I_p - \alpha)\beta_0 \to \beta_0.$$

The estimated posterior bias is: $bias(\hat{\beta}_{PSA}(\hat{\lambda})) = -\hat{\lambda}(X'X + \hat{\lambda}I_p)^{-1}(\beta - \beta_0) \to \beta - \beta_0 \to \tau$, which is small since τ is small. The estimated variance $V(\hat{\beta}_{PSA}(\hat{\lambda})) = X'X(X'X+\hat{\lambda}I_p)^{-2}\sigma^2$ is also small since $\hat{\lambda}$ is large. The results suggests that, if the prior is good, not only $MSE(\hat{\beta}_{PSA})$ and $V(\hat{\beta}_{PSA})$ are smaller, but the posterior bias is also under control. This is the situation when the benefit of using PSA is maximized. Of course, this is a reward of good prior information.

The theoretical results can be illustrated by a simple example. Let $p = 1$ and $\sigma^2 = 10$. Set the square of the prior bias $\tau^2 = 0, 4, 10, 16, 25, 100$, and λ from 0 to 20 by 0.05. By $Eq(3.1)$, $MSE(\hat{\beta}_{PSA}(\lambda))$ is calculated on each value of λ; $MSE(\hat{\beta}_{LS}) = \sigma^2 = 10$. The square of both the prior bias and the posterior bias, the $V(\hat{\beta}_{PSA}(\lambda))$, the $MSE(\hat{\beta}_{PSA}(\lambda))$ and the $MSE(\hat{\beta}_{LS})$ are plotted over λ with different prior bias τ in Figure 3.1.

The $MSE(\hat{\beta}_{LS}) = 10$ is independent to τ and λ. As $\lambda \to \infty$, the posterior bias is monotonically increasing (except (a), which is 0) and converges to the prior bias τ; the $V(\hat{\beta}_{PSA}(\lambda))$ is monotonically decreasing and converges to 0; and the $MSE(\hat{\beta}_{PSA}(\lambda))$ converges to τ^2. $\hat{\lambda}$ is the minimizer of $MSE(\hat{\beta}_{PSA}(\lambda))$. When τ is small (left 3 plots), $\hat{\lambda}$ may not be estimated

accurately since the MSE curve is quite flat around its minimum value. If $\tau^2 < MSE(\hat{\beta}_{LS})$, then $MSE(\hat{\beta}_{PSA}(\lambda)) < MSE(\hat{\beta}_{LS})$ for all $\lambda > 0$ (left 3 plots). When τ is large (right 3 plots), $\hat{\lambda}$ may be estimated accurately since the minimum of the MSE curve is obvious, but corresponding to a very small λ value. If $\tau^2 > MSE(\hat{\beta}_{LS})$, the Theorem of Existence ensures at least for some small $\lambda, MSE(\hat{\beta}_{PSA}(\lambda)) < MSE(\hat{\beta}_{LS})$. We observe that there exists some region of λ, in which the red curve is under the straight black line (right 3 plots).

3.2.2.4 Partial shrinkage

β_0 is a $p \times 1$ vector containing prior information for all p parameters. When $p > 1$, in practice it is possible that the prior information of each β_j has different quality, i.e., for some j, β_{0j} is close to the true β_j; for other j, β_{0j} is far away from the truth. In those cases, the PSA tends to shrink the two sets of parameters in different depth. For the subset of β with good priors, the shrinkage is deep and $\hat{\beta}_{PSA} \approx \beta_0$; for the subset of β with bad priors, the shrinkage is slight and $\hat{\beta}_{PSA} \approx \hat{\beta}_{LS}$. This feature provides the basis of variable selection for the ridge regression and the LASSO, given $\beta_0 = 0$. For details, see Tibshirani, 1996.

3.2.2.5 The β_0:

As a statistical method, the PSA is developed to integrate moment prior information of parameters in data analysis within a least square framework. However, the choice of β_0 is not limited to be the prior knowledge of the location of the parameters. Theoretically, β_0 can be any points in the parameter space. In application, another meaningful choice of β_0 is the origin 0. For example, in ridge regression and LASSO, even we have no prior information about the true parameters, β_0 is set to be 0. The goal is to select a subset of variables that have no effects on the response (a subset of parameters whose true values are 0). As all parameters are shrunk towards 0, the ones with

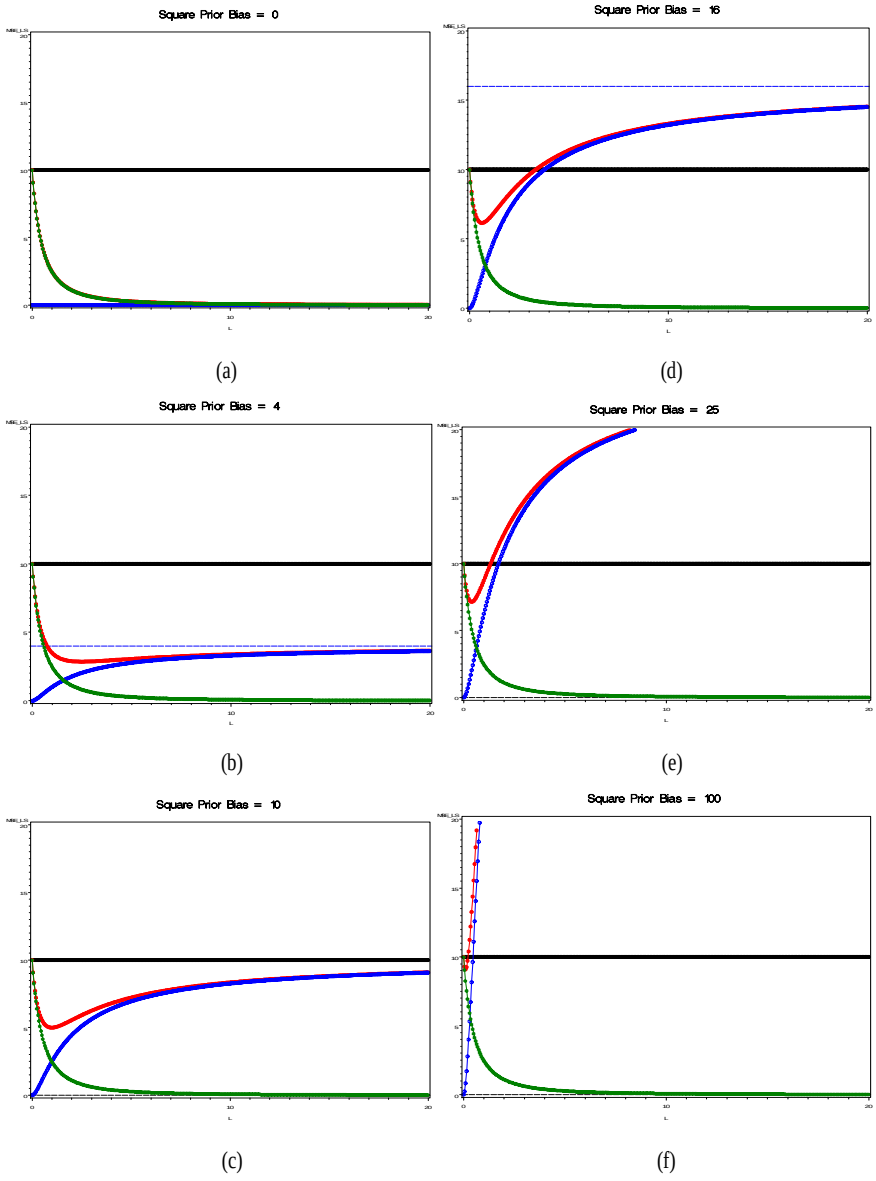


Figure 3.1. Some statistics of $\hat{\beta}_{PSA}(\lambda)$ as a function of λ and the prior bias.
The blue curve: the square of the posterior bias; the blue dot line: the square of the prior bias; the green curve: V($\hat{\beta}_{PSA}(\lambda)$); the red curve: MSE($\hat{\beta}_{PSA}(\lambda)$); the black line: MSE ($\hat{\beta}_{LS}$). (a) $\tau^2 = 0$; (b) $\tau^2 = 4$; (c) $\tau^2 = 10$; (d) $\tau^2 = 16$; (a) $\tau^2 = 25$; (a) $\tau^2 = 100$.

small effects get deeper shrinkage than the ones with large effects, since 0 is the "true value" of the parameters with small effects.

3.3 A Simulation Study

A simple simulation study has been designed to compare the PSA with the least square method. Since λ is estimated by cross validation, the performance of this procedure is tested. The theoretical results derived previously are also examined in the simulation.

3.3.1 Design:

100 data sets are generated. In each data set, 2 variables Y_1 and Y_2 are randomly drawn from a bi-variate normal distribution BVN($\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$) = BVN(10,10,30,30,0) 200 times. A normal density is chosen only because it is easy to simulate in SAS program. The properties of normal distribution are not used in the following analysis. Two responses (Y_1 and Y_2) are generated because a 2-D plot gives better vision of shrinkage than a 1-D plot. Y_1 and Y_2 are set to be independent.

The mean of Y_1 and Y_2 (μ_1, μ_2) are the parameters of interest. To estimate them from the simulated data, a linear model is applied for Y_1 and Y_2:

$$Y_i = X\mu_i + \epsilon, \qquad i = 1, 2,$$

where Y_i is a (200×1) vector, X is a (200×1) vector of 1's. ϵ is a (200×1) error vector with covariance matrix $\sigma^2 I_{200}$.

The quality of the prior information is measured by prior bias, $PB = \sqrt{2}|\mu_i - \mu_0| = \sqrt{2}|\tau|$ (because the dimension 2). Table 3.1 shows that the designed PB ranges from 0 unit (the best prior) to 1000 units (the worst

Table 3.1: The Designed Values of the Prior Bias

Scenario	$\tau(std)$	Prior μ_0
1	0.0	10.00
2	0.1	10.15
3	0.2	10.30
4	0.3	10.45
5	0.4	10.60
6	0.5	10.75
7	0.7	11.05
8	1.0	11.50
9	1.5	12.25
10	2.0	13.0
11	3.0	14.5
12	4.0	16.0
13	5.0	17.5
14	10	25.0
15	100	160
16	1000	1510

prior). The "unit" is defined as the standard deviation of the least square estimate of μ_i, i.e., 1 unit=$\sqrt{X'X^{-1}}\sigma = 1.5\sqrt{2}$.

3.3.2 The 10 Fold Cross Validation:

The cross-validation process can be briefly described as follows: the dataset is randomly equally divided into 10 subsets. 9 shares form the training set, and the remaining 1 is the validation set. (β, λ) are estimated in the training set and is evaluated in the validation set. Specifically, in the training set, λ takes values from points on the grid. In this simulation, $log(\lambda)$ takes values from -10 to 10 by step 0.2. β is estimated at each λ value in the training set. The $(\hat{\beta}(\lambda))$ is then fixed in the validation set to calculate the prediction error (PE). The λ value producing the smallest average PE is the estimated $\hat{\lambda}$. This procedure is repeated for all 10 validation sets with the same grid value of λ. Based on this $\hat{\lambda}$, $\hat{\beta}_{PSA}(\hat{\lambda})$ is estimated using the whole data set.

This $(\hat{\lambda}, \hat{\beta}_{PSA}(\hat{\lambda}))$ is the 10 fold cross-validation estimator. For more details, see Chapter 17, Efron and Tibshirani, 1993.

3.3.3 Analysis and Results:

For each dataset, given a value of the prior bias PB, the prior mean is $\mu_0 = \mu + PB$. $\hat{\lambda}$ is obtained by the cross-validation procedure and the PSA estimators for μ_1, μ_2 are $\hat{\beta}_{PSA.1} = (X'X + \hat{\lambda}I_p)^{-1}(X'Y_1 + \hat{\lambda}\mu_0)$, and $\hat{\beta}_{PSA.2} = (X'X + \hat{\lambda}I_p)^{-1}(X'Y_2 + \hat{\lambda}\mu_0)$. The least square estimators are $\hat{\beta}_{LS.1} = (X'X)^{-1}X'Y_1$, and $\hat{\beta}_{LS.1} = (X'X)^{-1}X'Y_2$. For a total of 100 data sets, the 100 $(\hat{\beta}_{PSA.1}, \hat{\beta}_{PSA.2})$ and $(\hat{\beta}_{LS.1}, \hat{\beta}_{LS.2})$ are plotted in a 2-D graph (Figure 3.2).

In figure 3.2, the red circles are the least square estimates, which remain at the same locations in all the four plots. Plot (a) represents a worst situation, where the prior information is far away from the truth. In this case, the results show that the PSA estimator $(\hat{\beta}_{PSA.1}, \hat{\beta}_{PSA.2})$ is almost the same as the least square estimator $(\hat{\beta}_{LS.1}, \hat{\beta}_{LS.2})$. λ is estimated at 10^{-4} (Figure 3.5), shrinkage is not observed (note that, in the plot, the blue dots and the red circles are at the same location, so the red circles are invisible). The quality of prior information is better in plot (b), (c) and (d), where the prior bias are set to be 10, 1 and 0 standard deviations, respectively. In plot (b), the prior mean is (25, 25), in the right-upper direction of the true parameter value (10,10). It is observed that the least square estimator $(\hat{\beta}_{LS.1}, \hat{\beta}_{LS.2})$ shrinks to the prior mean in the right-upper direction to generate the PSA estimator $(\hat{\beta}_{PSA.1}, \hat{\beta}_{PSA.2})$. In the plot, it appears that $(\hat{\beta}_{PSA.1}, \hat{\beta}_{PSA.2})$ is a little biased, while $MSE(\hat{\beta}_{PSA})$ is slightly smaller than $MSE(\hat{\beta}_{LS})$ (Figure 3.4). In plot (c), the prior mean is (11.5, 11.5), much closer to the truth than (b). Deep shrinkage is observed. Plot (d) represents the best situation, where the prior mean is the truth. In plot (d), all least square estimators are deeply shrunk to the truth (10, 10) forming a small but dense cluster, which contains more than 70 blue dots. In this case, λ is estimated at 10^5

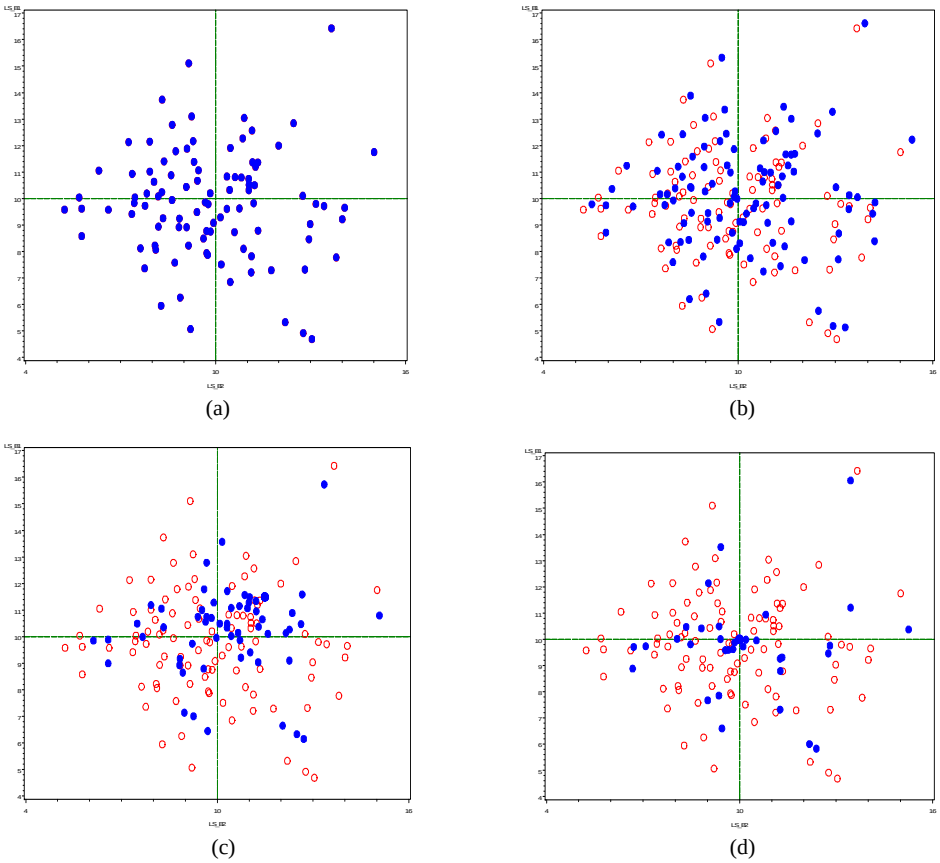
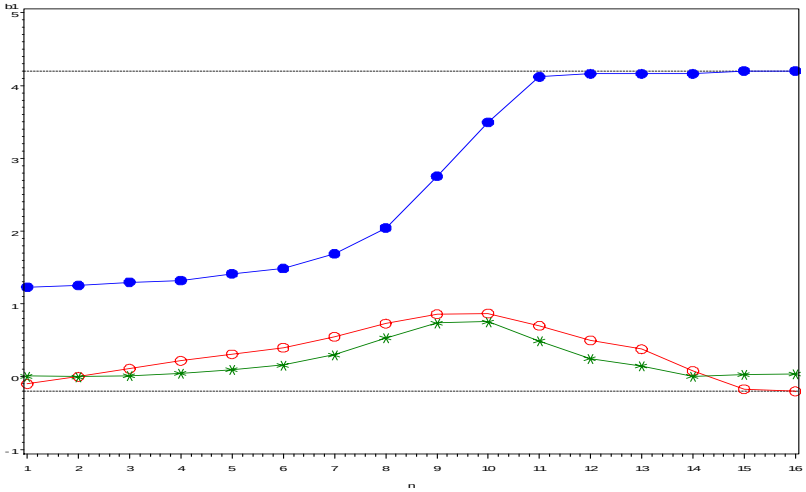


Figure 3.2. Plot of $\hat{\beta}_{LS}$ and $\hat{\beta}_{PSA}$ under different prior bias, (a) scenario 16, where PB=1000 units; (b) scenario 14, where PB=10 units; (c) scenario 8, where PB=1 unit and (d) scenario 1, where PB=0. Red circle: $\hat{\beta}_{LS}$, Blue dot: $\hat{\beta}_{PSA}$. The green dot lines indicate the true parameter values (10, 10).

(Figure 3.4) and $MSE(\hat{\beta}_{PSA})$ is much smaller than $MSE(\hat{\beta}_{LS})$ (Figure 3.5) Note that, when the prior bias is 0, we expect both the posterior bias and variance of ($\hat{\beta}_{PSA}$ are 0. In plot (d), however, we do observe some variation. This happens because in the simulation, the range of λ is limited to be less than e^{10} in the grid search. $\hat{\lambda}$ can not reach ∞, the expected value when prior bias is 0. So the depth of the shrinkage is limited. This is a typical problem for procedures involving grid search. This problem also appears in some other results (i.e., see Figure 3.4). But in application, the problem may not be serious since the prior information is unlikely to be exactly the truth.

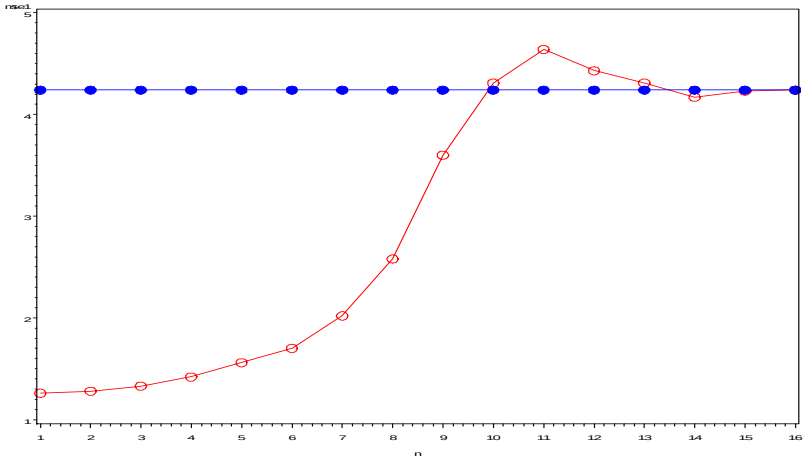
Figure 3.3 shows how the posterior bias and variance of ($\hat{\beta}_{PSA}$ change over the 16 different scenarios. The results show that, as the prior bias increases, the variance monotonically increases; but the posterior bias increases first and takes the maximum value when the prior bias is about 1 - 1.5 standard deviation away from the truth, and then decreases to 0. This happens because the posterior bias is a function of both the prior bias and λ. If the prior bias is too bad, then λ decreases sharply so that $\hat{\beta}_{PSA}$) remains at $\hat{\beta}_{LS}$), which is unbiased.

Figure 3.4 shows how MSE changes over the 16 different scenarios. The $MSE(\hat{\beta}_{PSA})$ takes its minimum value when there is no prior bias. Then it increases to its maximum value $MSE(\hat{\beta}_{LS})$ when the prior point is 1.5 standard deviations away from the truth, and stays there afterward. In some scenarios (specifically, 11-14), the results show $MSE(\hat{\beta}_{PSA}) > MSE(\hat{\beta}_{LS})$, which should not have happened in theory. Again, the grid search causes the problem. In a grid search, λ has to take the grid values, while the true λ that minimizes MSE is not likely exactly on those grid points. Thus the $\hat{\lambda}$ is an approximation of the true minimizer λ_{min}. So the $MSE(\hat{\beta}_{PSA}(\hat{\lambda}))$ is larger than the true $MSE(\hat{\beta}_{PSA}(\lambda_{min}))$. In this simulation, the grid is set from e^{-10} to e^{10} with step $e^{0.2}$. If the grid is designed to have more points



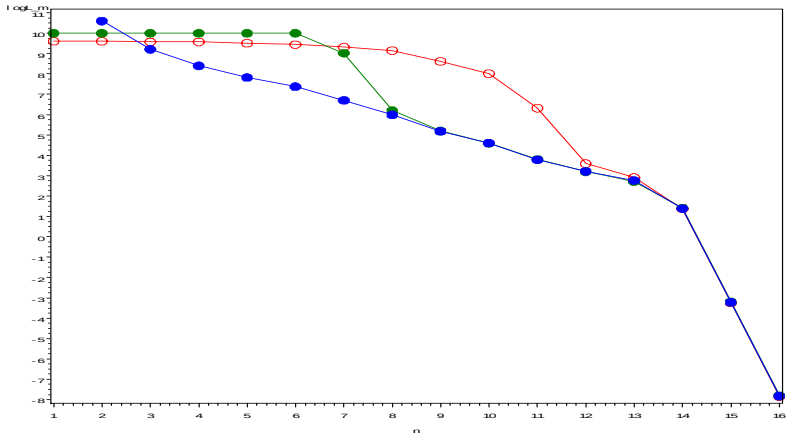
Prior Bias small $\rightarrow\rightarrow\rightarrow\rightarrow\rightarrow$ large

Figure 3.3. Posterior bias and variance of $\hat{\beta}_{PSA}$, estimated from 16 different scenarios. The black dash lines: the variance (up) and squared bias (bottom) of $\hat{\beta}_{LS}$. Red curve: the posterior bias of $\hat{\beta}_{PSA}$; green curve: the squared posterior bias of $\hat{\beta}_{PSA}$; blue curve: the variance of $\hat{\beta}_{PSA}$.

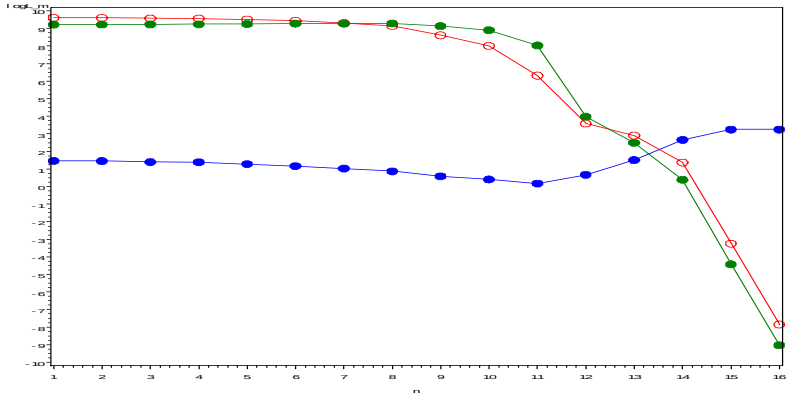


Prior Bias small $\rightarrow\rightarrow\rightarrow\rightarrow\rightarrow$ large

Figure 3.4. Estimated prediction error from 16 different scenarios: The red circle: MSE($\hat{\beta}_{PSA}$), the blue dot: MSE($\hat{\beta}_{LS}$).



(a)



(b)

good prior $\rightarrow\rightarrow\rightarrow\rightarrow\rightarrow\rightarrow\rightarrow$ bad prior

Figure 3.5. Estimating the tuning parameter λ from different scenarios. (a) The mean estimate of λ. The blue line is the λ estimated from equation (Eq (3.1)), the other two lines are the estimates from the cross-validation: the green line and the red line are the median and mean of the estimates from the 10 cross-validation replications. In (b), the red line is the mean estimate of λ (same as the red line in (a)), the green line is the standard deviation of the estimated λ , and the blue line is the Co-efficiency of variance (CV).

with higher density (i.e., from same range of e^{-10} to e^{10} but with step $e^{0.1}$), we observe that the problem is much less serious.

The results from Figure 3.3 and 3.4 suggest that, the standard deviation may be a critical value for the prior bias. If the prior bias is less than a standard deviation, benefit is expected by using the PSA; otherwise, little advantages are expected, though there is also little hazard to apply this method.

Figure 3.5 shows some statistical properties of the estimate of the tuning parameter λ. Since the range of $\hat{\lambda}$ is too wide, the logarithm of $\hat{\lambda}$ is used. In plot (a), $log(\hat{\lambda})$ decreases monotonically as the quality of the prior information becomes worse, indicating bad prior results in less shrinkage. The exact distribution of $\hat{\lambda}$ is unknown but it is obviously highly skewed to the right. The mean and the median differs significantly (4 units in a log scale). Plot (b) shows that, if the prior is good, both the mean and the variance of $\hat{\lambda}$ is large; as the prior becomes worse, the mean and the variance both get smaller. This observation agrees with the theoretical results from Figure 3.1. The range of the magnitude of $\hat{\lambda}$ is very wide, from 10^{-4}(bad prior) to 10^5 (good prior), indicating a big difference in the depth of shrinkage. By $Eq(3.1)$, when $\hat{\lambda} = 10^5$, the prior weights almost 100% in the posterior estimator $\hat{\beta}_{PSA}$, $\hat{\beta}_{PSA}$ is very close to the β_0. On the other hand, when $\hat{\lambda} = 10^{-4}$, the prior has no weight, $\hat{\beta}_{PSA}$ is close to the least square estimate $\hat{\beta}_{LS}$.

In conclusion, the PSA produces a reasonable good estimator $\hat{\beta}_{PSA}$ for parameter β. It always has a smaller (with good prior), or at least equal (with bad prior) prediction error than the least square estimate $\hat{\beta}_{LS}$. When prior information well represents the truth, most advantages are expected by applying the PSA procedure; when prior information is far away from the truth, the PSA gives a conservative estimator which is close to the unbiased least square estimator $\hat{\beta}_{LS}$.

3.4 Extensions and Future Work

3.4.1 Integrating the second moment information:

An immediate extension of the PSA is to integrate the prior second moment information (variance/covariance matrix) of the parameter vector β_0. This work may be interesting in the genetic network studies. In those studies, the correlation between genes are the main focus. Usually some correlations between a few specific genes are already known and are the basis of searching other genes in the network. The work has not completely done yet. Some early results are reported below.

Suppose the prior second moment information is stored in a $p \times p$ matrix D:

$$\begin{pmatrix} Var(\beta_{01}) & Cov_(\beta_{01}, \beta_{02}) & . & Cov_(\beta_{01}, \beta_{0p}) \\ Cov_(\beta_{02}, \beta_{01}) & Var(\beta_{02}) & . & Cov_(\beta_{02}, \beta_{0p}) \\ . & . & . & . \\ Cov_(\beta_{0p}, \beta_{01}) & Cov_(\beta_{0p}, \beta_{02}) & . & Var(\beta_{0p}) \end{pmatrix}.$$

With this prior D matrix, the objective function

$$\begin{aligned} q(\beta) &= \sum_{i=1}^{N}(Y_i - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p}(\beta_j - \beta_{0j})^2 \\ &= (Y - X\beta)'(Y - X\beta) + \lambda(\beta - \beta_0)'(\beta - \beta_0), \end{aligned}$$

is substituted by a generalized $q'(\beta)$ function

$$q'(\beta) = (Y - X\beta)'(Y - X\beta) + \lambda(\beta - \beta_0)'D^{-1}(\beta - \beta_0).$$

Now, instead of a point β_0, the least square estimate $\hat{\beta}_{LS}$ is shrunk to an eclipse centered at β_0 in a p-dimensional space.

The PSA solution for β is:

$$\frac{\partial q'(\beta)}{\partial \beta} = 0 \Rightarrow \tilde{\beta}_{PSA} = (X'X + \lambda D^{-1})^{-1}(X'Y + \lambda D^{-1}\beta_0).$$

$\tilde{\beta}_{PSA}$ is biased with $bias(\tilde{\beta}_{PSA}(\lambda)) = -\lambda(X'X+\lambda D^{-1})^{-1}(\beta-\beta_0)$ and covariance matrix $V(\tilde{\beta}_{PSA}(\lambda)) = X'X(X'X+\lambda D^{-1})^{-2}\sigma^2$. Let $\dot{\alpha}_{P\times P} = (X'X + \lambda D^{-1})^{-1}X'X$, the MSE of $\tilde{\beta}_{PSA}$ can be calculated as:

$$\begin{aligned} MSE(\tilde{\beta}_{PSA}) &= E[(\tilde{\beta}_{PSA}-\beta)'(\tilde{\beta}_{PSA}-\beta)] \\ &= \sigma^2\sum_{i=1}^{p}\frac{\dot{\delta}_i}{(\dot{\delta}_i+\lambda D^{-1})^2} + (\beta-\beta_0)'((X'X)^{-1}+\lambda D^{-1})^{-2}(\beta-\beta_0), \end{aligned}$$

The next step is to estimate the posterior covariance. The work is still in progress.

Another extension is to apply this idea in a likelihood framework, adding the prior information as a penalty term. In fact, likelihood may be penalized for many reasons, such as variable selection and smoothing. If this work is succeeded, the PSA may be extended to many likelihood based on analysis, such as the Mixed model analysis (as mensioned in Chapter 1).

3.5 Discussion

In a comment paper (to Dr. Leo Breiman's paper), Dr. Brad Efron wrote that "from the point of view of statistical development the twentieth century might be labeled '100 years of unbiasedness'. Following Fisher's lead, most of our current statistical theory and practice revolves around unbiased or nearly unbiased estimate ... the theory requires a modestly high ratio of signal to noise, sample size to number of unknown parameters, to have much hope of success...".

In fact, if the sample size is large enough so that all parameters in an appropriate statistical model can be estimated precisely and accurately, we may not bother to develop methods like the PSA. Unfortunately, small sample size (so called "small n") is a typical problem in genetic data analysis. The performance of many current statistical methods based on asymptotic

theory may be poor when sample size is small. As a result, little information can be extracted from the data due to the large variance. In this case, some external independent information is helpful to find and stabilize the signals in the data. In other cases, it is desired that the independent knowledge is integrated in the data analysis simply because the knowledge is important or is the well accepted "truth".

The PSA is a tool which integrates prior moment information. In general, the PSA trades off variance for possibly increased bias. The degree of bias depends on the quality of the prior information, i.e., the better the prior, the smaller the bias. The PSA is also self-adjusted. If the prior information is far away from the truth, the PSA estimator protects itself by staying around the least square estimator in the parameter space. This desired property is ensured by the cross validation procedure.

The PSA is used for estimation. It may not be appropriate to be used for statistical testing (testing the prior information, which may be taken as the null hypothesis), since obviously the results will be in favor of the null hypothesis. However, since the tuning parameter λ measures the depth of the shrinkage, it may be used as an indicator to examine if null hypothesis holds. A large λ implies the data agree with the null hypothesis; while a small λ implies the data are against the null hypothesis.

Appendix

Appendix 2.1
Re-parameterize the Markov chain parameters by the LD measures

Goal: to re-parameterize the 7 $MC2$ parameters in terms of the 7 LD measures, so that the Markov chain likelihood can be expressed with LD measures, and the 2 constraints $D^*_{123} = 0$ and Eq.1 can be tested.

It is difficult to re-write the Markov chain parameters with LD measure directly. We use the haplotype frequencies as a bridge. Firstly, haplotype frequencies are expressed in terms of the LD measures; then the Markov chain parameters are expressed in terms of the haplotype frequencies. Specifically, for three binary markers (M_1, M_2, M_3), the four haplotype frequencies for the first two markers are $P_{00}, P_{01}, P_{10}, P_{11}$; the eight 3-marker haplotype frequencies are $P_{000}, P_{001}, P_{010}, P_{011}, P_{100}, P_{101}, P_{110}, P_{111}$; the seven LD measures are $(P_1, P_2, P_3, D_{12}, D_{23}, D_{13}, D^*_{123})$; the seven $MC2$ parameters are the three chain initial parameters

$$P(H_{1,2} = (1,1)), P(H_{1,2} = (1,0)), P(H_{1,2} = (0,1)),$$

and the four conditional probabilities

$$P^{(3,H_{(1,2)})}_{(1,(1,1))}, P^{(3,H_{(1,2)})}_{(1,(1,0))}, P^{(3,H_{(1,2)})}_{(1,(0,1))}, P^{(3,H_{(1,2)})}_{(1,(0,0))}.$$

It is straightforward to express the haplotype frequencies in terms of the

LD measures:

$$\begin{aligned}
P_{11} &= P_1P_2 + D_{12}, \\
P_{10} &= P_1(1-P_2) - D_{12}, \\
P_{01} &= (1-P_1)P_2 - D_{12}, \\
P_{111} &= P_1D_{23} + (1-P_2)D_{13} + P_3D_{12} + P_1P_2P_3 + D^*_{123}, \\
P_{101} &= -P_1D_{23} + P_2D_{13} - P_3D_{12} + P_1(1-P_2)P_3 - D^*_{123}, \\
P_{011} &= (1-P_1)D_{23} - (1-P_2)D_{13} - P_3D_{12} + (1-P_1)P_2P_3 - D^*_{123}, \\
P_{001} &= -(1-P_1)D_{23} - P_2D_{13} + P_3D_{12} + (1-P_1)(1-P_2)P_3 + D^*_{123}.
\end{aligned}$$

Then the $MC2$ parameters can be expressed as:

$$\begin{aligned}
P(H_{1,2} = (1,1)) &= P_{11}, \\
P(H_{1,2} = (1,0)) &= P_{10}, \\
P(H_{1,2} = (0,1)) &= P_{01}, \\
P^{(3,H_{(1,2)})}_{(1,(1,1))} &= P_{111}/P_{11}; \\
P^{(3,H_{(1,2)})}_{(1,(1,0))} &= P_{101}/P_{10}; \\
P^{(3,H_{(1,2)})}_{(1,(0,1))} &= P_{011}/P_{01}; \\
P^{(3,H_{(1,2)})}_{(1,(0,0))} &= P_{001}/(1-P_{11}-P_{10}-P_{01});
\end{aligned}$$

Now the likelihood can be formulated in LD measures.

Appendix 3.1 The Proof of the Theorem of Existence

Theorem of Existence: If $0 < \lambda < 2\sigma^2/(\beta-\beta_0)(\beta-\beta_0)' = c$, *then* $MSE(\hat{\beta}_{PSA}) < MSE(\hat{\beta}_{LS})$

The following proof is a generalized version of the original proof by Theobald (where $\beta_0 = 0$) (Theobald, 1974). The proof has two steps. In the first step, we show:

Theorem 1: if $0 < \lambda < 2\sigma^2/(\beta-\beta_0)(\beta-\beta_0)' = c$, *then* $M(\hat{\beta}_{LS})-M(\hat{\beta}_{PSA})$ *is positive definite, where* M *is the second-order moment matrix and* $M(\hat{\beta}) = E(\hat{\beta}-\beta)(\hat{\beta}-\beta)'$.

Given $bias(\hat{\beta}_{PSA}(\lambda)) = E(\hat{\beta}_{PSA}-\beta) = -\lambda(X'X+\lambda I_p)^{-1}(\beta-\beta_0)$, and covariance matrix $V(\hat{\beta}_{PSA}(\lambda)) = X'X(X'X+\lambda I_p)^{-2}\sigma^2$, the second-order moment matrix M can be written as:

$$V(\hat{\beta}_{PSA}(\lambda)) = V(\hat{\beta}_{PSA}(\lambda)) + bias(\hat{\beta}_{PSA}(\lambda))bias(\hat{\beta}_{PSA}(\lambda))'$$

Note that when $\lambda = 0$, the point shrinkage method equals the least square method, or $M(\hat{\beta}_{LS} = M(\hat{\beta}_{PSA}(\lambda = 0))$. Next, I will show $M(\hat{\beta}_{LS})-M(\hat{\beta}_{PSA})$ is positive definite if $0 < \lambda < c$.

$M(\hat{\beta}_{LS}) - M(\hat{\beta}_{PSA}(\lambda)) = (X'X)^{-1}\sigma^2 - X'X(X'X+\lambda I)^{-2}\sigma^2 - \lambda^2(X'X+\lambda I)^{-1}(\beta-\beta_0)(\beta-\beta_0)'(X'X+\lambda I)^{-1} = \lambda(X'X+\lambda I)^{-1}\{[(X'X)^{-1}+2I]\sigma^2 - \lambda(\beta-\beta_0)(\beta-\beta_0)'\}(X'X+\lambda I)^{-1}$.

If $0 < \lambda$, $M(\hat{\beta}_{LS})-M(\hat{\beta}_{PSA})$ is positive definite if and only if $[(X'X)^{-1}+2I]\sigma^2 - \lambda(\beta-\beta_0)(\beta-\beta_0)'$ is positive definite. A sufficient condition for this is $2\sigma^2 - \lambda(\beta-\beta_0)(\beta-\beta_0)'$ is positive definite. Since the eigenvalues for $\lambda(\beta-\beta_0)(\beta-\beta_0)'$ are 0 and $2\sigma^2 - \lambda(\beta-\beta_0)'(\beta-\beta_0)$. Thus, as long as $0 < \lambda < 2\sigma^2/(\beta-\beta_0)(\beta-\beta_0)' = c$, $2\sigma^2 - \lambda(\beta-\beta_0)(\beta-\beta_0)'$ is positive definite. Since β, β_0 and σ^2 are all constants, so is c. #

The next step is to show:

Theorem 2: $M(\hat{\beta}_{LS})-M(\hat{\beta}_{PSA})$ *is positive definite* $\Rightarrow M(\hat{\beta}_{PSA}) < M(\hat{\beta}_{LS})$, *where MSE is defined as* $MSE(\hat{\beta}_{PSA}) = E(\hat{\beta}_{PSA}-\beta)'(\hat{\beta}_{PSA}-\beta)$

Proof:

Since $MSE(\hat{\beta}_{PSA}) = tr(M(\hat{\beta}_{PSA}))$ and $MSE(\hat{\beta}_{LS}) = tr(M(\hat{\beta}_{LS}))$, so:

$$MSE(\hat{\beta}_{LS}) - MSE(\hat{\beta}_{PSA}) = tr(M(\hat{\beta}_{LS})) - tr(M(\hat{\beta}_{PSA}))$$

In Theorem 1, it has been proved that $M(\hat{\beta}_{LS}) - M(\hat{\beta}_{PSA})$ is positive definite if $0 < \lambda < 2\sigma^2/(\beta-\beta_0)(\beta-\beta_0)' = c$. It follows that, under the same condition, $tr(M(\hat{\beta}_{LS}) - M(\hat{\beta}_{PSA})) > 0$.

So, in conclusion, $0 < \lambda < 2\sigma^2/(\beta-\beta_0)(\beta-\beta_0)' = c \Rightarrow M(\hat{\beta}_{LS}) - M(\hat{\beta}_{PSA})$ is positive definite $\Rightarrow MSE(\hat{\beta}_{LS}) - MSE(\hat{\beta}_{PSA}) > 0$ #

References

References for chapter 1

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing," Journal of the Royal Statistical Society. B, 57, 289 -300.

Box , G. and Tiao, G. (1973). "Bayesian Inference in Statistical Analysis". New York: John Wiley and Sons

Chhabra, S., Shockley, K., Conners, S., Scott, K., Wolfinger, R. and Kelly, R. (2003). "Carbohydrate-induced differential gene expression patterns in the hyperthermophilic bacterium Thermotoga maritime". Journal of Biological Chemistry 278, 7540-7552

Chu, T., Weir, B., and Wolfinger, R. (2002). "A systematic statistical linear modeling approach to oligonucleotide array experiments". Mathematical Biosciences 176, 35-51

Chu, T., Weir, B., and Wolfinger, R. (2004). "Comparison of Li-Wong and Loglinear Mixed Models for the Statistical Analysis of Oligonucleotide Arrays". Bioinformatics 20, 500-506

Efron, B., Tibshirani, R., Goss, V., Chu, G. (2000) "Microarrays and their use in a comparative experiment", Technical Report, Stanford University

Hochberg, Y. and Westfall, P. (2000). "On Some Multiplicity Problems and Multiple Comparisons Procedures in Biostatistics". Handbook of Statistics, Elsevier Sciences 18, 75-113

Hoerl, A. and Kennard, R. (1970). "Ridge regression: Biased estimation for nonorthogonal problems". Technometrics, 12(1):55-67.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U. and Speed, T. (2003). "Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data". Biostatistics 4, 249-264

Jin, W., Riley, R., Wolfinger, R., White, K., Passador-Gurgel, G., and Gibson, G. (2001). "The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster". Nature Genetics 29, 389 - 395

Kerr, M., Martin, M. and Churchill, G. (2000). "Analysis of variance for gene expression microarray data". Journal of Computational Biology 7, 819-837

Kerr, M., Afshari, C., Bennett, L., Bushel, P., Martinez, J., Walker, N., and Churchill, G. (2002). "Statistical analysis of a gene expression microarray experiment with replication". Statistica Sinica 12, 203-217

Li, C. and Wong, W. (2001). "Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection". Proceedings of the National Academy Science 98, 31-36

Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., Brown, H. (1996) "Expression monitoring by hybridization to high-density oligonucleotide arrays", Nat. Biotechnol. 14, 1675.

Lonnstedt, I and Speed, T. (2001). "Replicated Microarray Data". Statistical Sinica 12, 31-46

Nature Genetics Editors (eds.), (2003). "Chipping Forecast II". Nature Genetics Supplement 32, 461-552

SAS online help: Proc Mixed/prior, SAS institute, Cary, NC

Smyth, G (2004). "Linear models and Empirical Bayes methods for assessing differential expression in microarray experiments". Statistical Applications in Genetics and Molecular Biology 3, Article 3

Speed, T. (2003). "Statistical Analysis of Gene Expression Microarray Data". Chapman & Hall, Boca Raton

Storey, J. and Tibshirani R. (2003). "SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays". In The Analysis of Gene Expression Data: Methods and Software, by G Parmigiani, ES Garrett, RA Irizarry and SL Zeger (editors). Springer, New York.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso". Journal of the Royal Statistical Society, Ser. B, 58, 267-288.

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

Wolfinger, R. Tobias, R. and Sall, J. (1994). "Computing Gaussian likelihoods and their derivatives for general linear mixed models", SIAM Journal on Scientific Computing, 15, 1294-1310.

Wolfinger, R., Kass, R. (2000) "Nonconjugate Bayesian Analysis of Variance Component Models". Biometrics 56, 768-774

Wolfinger, R. Gibson, G., Wolfinger, E., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., Paules, R. (2001) "Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models". Journal of Computational Biology 8, 625-637

References for chapter 2

Altmller, J., Palmer, L.J., Fischer, G., Scherb, H. and Wjst, M. (2001) "Genomewide scans of complex human diseases: true linkage is hard to find". Am. J. Hum. Genet., 69, 936-950

Bennett, J. H., (1954) "On the theory of random mating". Ann. Eugen. 18: 311-317

Csiszar, I. and Shields, P. (1999) "Consistancy of the BIC order estimator". Electronic Research Announcements of the American Mathematical Society 5, 123-127

Daly, M.J. et al. (1998) "Genehunter 2.0–a complete linkage analysis system". Am. J. Hum. Genet. 63, A286

Daly, M., Rioux, J., Schaffner, S., Hudson T., Lander, E., (2001) "High-resolution haplotype structure in the human genome". Nat Genet 29: 229-232

DeLuca, M., Roshina, N., Geiger-Thornsberry, G., Lyman, R., Pasyukova, E., Mackay, T. (2003) "Dopa-decarboxylase affects variation in Drosophila longevity". Nature Genetics 34: 429-433.

Devlin, B., Risch, N., and Roeder, K. (1996) "Disequilibrium mapping: Composite likelihood for pairwise disequilibrium". Genomics 36: 1-16

Excoffier, L. and Slatkin, M. (1995) "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population". Mol. Biol. Evol. 12, 921-927.

Finesso, L. (1992) "Estimation of the order of a finite Markov chain, in Recent Advances in the Mathematical Theory of Systems, Control, and Network Signals" Proc. MTNS-91, H. Kimura and S. Kodama, Eds., Mita Press, 643-645

Geiringer, H., (1944) "On the probability theory of linkage in Mendelian heredity". Ann. Math. Stat. 15: 25-57

Gorelick, R. and Laubichler, M (2004) "Decomposing Multilocus Linkage Disequilibrium". Genetics 166: 1581-1583

Hastings, A. (1984) "Linkage disequilibrium, selection and recombination at three loci". Genetics 106: 153-164

HapMap website: http://www.hapmap.org/whatishapmap.html.en.

Hawley, M. and Kidd, K. (1995). "HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes". J. Hered. 86, 409-411

Hudson, R. (2002) "Generating samples under a Wright-Fisher neutral model". Bioinformatics 18:337-338

Jorde, L, (2000) "Linkage Disequilibrium and the Search for Complex Disease". Genes. Genome Research 10, 1435-1444

Kaplan, N. and Morris, R. (2001), "Prospects for association-based fine mapping of a susceptibility gene for a complex disease", Theor. Popul. Biol. 60, 181-191

Katz, R. (1981), "On some criteria for estimating the order of a Markov chain". Technometrics 23, 243-249

Kerem, B., et al, (1989)"Identification of the cystic fibrosis gene: genetic analysis", Science 245, 1073-1080

Li N, Stephens M. (2003) "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data". Genetics.165(4):2213-33

Lin DY. (2004) "Haplotype-based association analysis in cohort studies of unrelated individuals". Genetic Epidemiology, 26:255-264

Long, A.D. and Langley, C.H. (1999), "The power of association studies to

detect the contribution of candidate genetic loci to variation in complex traits", Genome Research 98, 720-731

Long, J. C., Williams, R. and Urbanek, M. (1995). "An E-M algorithm and testing strategy for multiple-locus haplotypes". Am. J. Hum. Genet. 56, 799-810

Lonjou C, Zhang W, Collins A, Tapper WJ, Elahi E, Maniatis N, Morton NE (2003) "Linkage disequilibrium in human populations". Proc Natl Acad Sci USA 100: 6069-6074.

Maniatis, N., Collins, A., Xu, C.-F., McCarthy, L., Hewett, D., Tapper, W., Ennis, S., Ke, X. and Morton, N. E. (2002) "The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis". Proc. Natl. Acad. Sci. USA 99, 2228-2233.

McPeek, M and Strahs, A. (1999) "Assessment of Linkage Disequilibrium by the Decay of Haplotype Sharing, with Application to Fine-Scale Genetic Mapping". Am. J. Hum. Genet. 65:858-875

Morton, N., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P. and Collins, A. (2001) "The optimal measure of allelic association" Proc. Natl. Acad. Sci. USA 98, 5217-5221

Riordan, J. et al, (1989)"Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA", Science 245, 1066-1073

Slatkin, M (1972) "On treating the chromosome as the unit of selection". Genetics 72: 157-168

Terwilliger, J. (1995) "A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci". Am. J. Hum. Genet. 56: 777-787

Wang, T, Ph.D. dissertation, 2001, NCSU.

Weir, B.S. (1996) "Genetic data analysis II". Sinauer, Sunderland, Mass.

Weir, B.S. and Cockerham, C.C. (1979) "Estimation of linkage disequilibrium in randomly mating populations". Heredity 42: 105-111

Xiong, M. and Guo, S. (1997) "Fine-scale genetic mapping based on linkage disequilibrium: theory and applications". Am. J. Hum. Genet. 60: 1513-1531

Zhang, K., Calabrese, P., Nordborg, M. and Sun, F. (2002), "Haplotype block structure and its applications to association studies: power and study designs", Am. J. Hum. Genet. 71, 1386-1394

Zhang, W, Collins, A, Maniatis, N, Tapper, W, Morton, NE. (2002 (a)) "Properties of linkage disequilibrium (LD) maps". Proc. Natl. Acad. Sci. USA 99:17004-17007

Zhang, W, Collins, A, Abecasis, GR, Cardon, LR, Morton, NE. (2002 (b)) "Mapping quantitative effects of oligogenes by allelic association". Ann. Hum Genet. 66:211-21

Zhang, X.H., Roeder, K., Wallstrom, G., and Devlin, B. (2003) "Integration of association statistics over genomic regions using Bayesian adaptive regression splines". Human Genomics 71891-21:17

Zhao LP, Li SS, Khalid N. (2003) "A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies". Am J Hum Genet 72: 1231-1250.

References for chapter 3

Efron, B and Tibshirani, R. (1993) "An introduction to the Bootstrap". London: Chapman and Hill.

Frank, I. and Friedman, J. (1993) "A statistical view of some chemometrics regression tools". Technometrics 35, 109-148

Goldstein, (1998) "Bayes Linear Analysis", Encyclopedia of Statistical Sciences, update volume 3

Hoerl, A.E. and Kennard, R.W. (1970) "Ridge regression: Biased estimation for nonorthogonal problems". Technometrics, 12(1):55-67

Murray, W., Gill, P. and Wright, M. (1981) "Practical Optimization", Chapter 5. New York: Academic Press

Tibshirani, R. (1996) "Regression Shrinkage and Selection via the Lasso", Journal of the Royal Statistical Society, Ser. B, 58, 267-288

Theobald, C. M. (1974) "Generalizations of Mean Square Error Applied to Ridge Regression", Journal of the Royal Statistical Society. Series B (Methodological), Vol. 36, No. 1. 103-106