PARALLEL TESTING, AND VARIABLE SELECTION – A MIXTURE-MODEL APPROACH WITH APPLICATIONS IN BIOSTATISTICS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Haim Y. Bar

January 2012

© 2012 Haim Y. Bar

ALL RIGHTS RESERVED

PARALLEL TESTING, AND VARIABLE SELECTION – A MIXTURE-MODEL APPROACH WITH APPLICATIONS IN BIOSTATISTICS

Haim Y. Bar, Ph.D.

Cornell University 2012

We develop efficient and powerful statistical methods for high-dimensional data, where the sample size is much smaller than the number of features (the so-called 'large p, small n' problem).

We deal with three important problems. First, we develop a mixture-model approach for parallel testing for unequal variances in two-sample experiments. The treatment effect on the variance has received little attention in the statistical literature, which so far focused mostly on the effect on the mean. The effect on the variance is increasingly recognized in recent biological literature, and we develop an empirical Bayes approach for testing differences in variance when the number of tests is large. We show that the model is useful in a wide range of applications, that our method is much more powerful than traditional tests for unequal variances, and that it is robust to the normality assumption.

Second, we extend these ideas and develop a novel bivariate normal model that tests for both differential expression and differential variation between the two groups. We show in simulations that this new method yields a substantial gain in power when differential variation is present. Through a three-step estimation approach, in which we apply the Laplace approximation and the EM algorithm, we get a computationally efficient method, which is particularly well-suited for 'large p, small n' situations.

Third, we deal with the problem of variable selection where the number of

putative variables is large, possibly much larger than the sample size. We develop a model-based, empirical Bayes approach. By treating the putative variables as random effects, we get shrinkage estimation, which results in increased power and significantly faster convergence, compared with simulation-based methods. Furthermore, we employ computational tricks which allow us to increase the speed of our algorithm, to handle a very large number of putative variables, and to control the multicollinearity in the model. The motivation for developing this approach is QTL analysis, but our method is applicable to a broad range of applications. We use two widely-studied data sets, and show that our model selection algorithm yields excellent results.

BIOGRAPHICAL SKETCH

Haim Bar received his Ph.D. in statistics at Cornell University in 2012. He received his M.Sc. in statistics in 2010 (Cornell University), and an M.Sc. in computer science in 2002 (Yale University). He received his bachelor degree in mathematics (Cum Laude) in 1993, at the Hebrew University in Jerusalem.

His professional interests include statistical modeling, shrinkage estimation, high throughput applications in biology (e.g., genomics, brain imaging), Bayesian statistics, and machine learning.

From 1995 to 1997, he was with Motorola, Israel, as a computer programmer in the Wireless Access Systems Division. From 1997 until 2003 he worked for MicroPatent, LLC, where he held the position of Director of Software Development. From 2005 to 2010 he was a teaching assistant at Cornell, and received the 'Outstanding TA Award' in 2006 ('Introduction to Computer Programming') and 2010 ('Statistical Methods').

This dissertation is dedicated to my beloved family. To my wife, Tali, and my wonderful children, Edan and Abigail, for their support and encouragement. It is also dedicated to my mother, Ghizela, and my late father, Dov.

ACKNOWLEDGEMENTS

I would first like to thank my advisors, Professor James Booth and Professor Martin T. Wells, for their guidance and support. Our weekly meetings have always been stimulating, productive and pleasant, and I have benefitted from them immensely. I am also deeply grateful to Professor Robert Strawderman for his help and support as a committee member, the director of graduate studies, a teacher, and a coauthor. I would also like to thank my collaborators, Dr. Dean Lillard, Dr. Miriam Agler-Rosenbaum, Gil Menda, and Professor Ron Hoy. Special thanks also to Dr. Elizabeth Schifano, a classmate and collaborator, and to Professor Francesca Molinari for introducing me to a fascinating research, for the collaboration, and for the generous funding for two semesters. I would also like to thank Françoise Vermeylen, from the Cornell Statistical Consulting Unit, from whom I learned a lot about solving real-life problems in statistics. I am grateful to the dedicated staff at the Department of Statistical Science and the Department of Biological Statistics and Computational Biology at Cornell, and especially to Beatrix Johnson, Todd Cullen, and Diana Drake. I would also like to thank Marketa and Dean Lillard, and Francesca Molinari and Levon Barseghyan for their friendship. Finally, my deepest gratitude to my family - my wife, Tali, and my children, Edan and Abigail, for their love and support.

TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Table of Contents . . . vi
List of Tables . . . viii
List of Figures . . . ix

1 Introduction **1**
1.1 Overview . . . 1
1.2 The General Approach . . . 4
1.2.1 The Statistical Model . . . 4
1.2.2 Estimation and Inference Procedures . . . 6
1.2.3 Computational Considerations and Tools . . . 8

2 A Mixture-Model Approach for Parallel Testing for Unequal Variances **10**
2.1 Introduction and Motivation . . . 10
2.2 The Mixture Model . . . 14
2.3 The EM Algorithm . . . 17
2.3.1 Complete data log-likelihood . . . 17
2.3.2 The E-step . . . 19
2.3.3 The M-Step . . . 19
2.4 Inference . . . 21
2.4.1 The Frequentist and Bayesian Procedures . . . 21
2.4.2 Shrinkage Estimation . . . 21
2.5 Simulation Results . . . 23
2.6 Case Studies . . . 28
2.6.1 Microarray Data . . . 29
2.6.2 Methylation Data . . . 31
2.6.3 fMRI Data . . . 31
2.6.4 Metabolomics Data . . . 33
2.7 Conclusions . . . 34

3 The Bivariate Model – Simultaneous Test for Mean and Variance **35**
3.1 Introduction . . . 35
3.2 The Model . . . 38
3.3 Estimation . . . 40
3.3.1 The Laplace Approximation . . . 40
3.3.2 Incorporating Gene-Specific Priors . . . 43
3.4 Inference . . . 45
3.4.1 The Frequentist Approach . . . 45
3.4.2 A Note on Multiple Testing . . . 46
3.4.3 The Empirical Bayesian Approach . . . 47

3.5 Power and Sample Size 48
3.6 Simulation Study 57
3.7 Case Studies 59
3.7.1 Metabolomics Data 60
3.7.2 Gene Expression Data 62
3.8 Conclusions 64

4 An Empirical Bayes Approach to Variable Selection and QTL Analysis 66
4.1 Introduction and Motivation 66
4.2 A Statistical Model for Automatic Variable Selection 68
4.3 Estimation 70
4.3.1 The Complete Data Likelihood 70
4.3.2 The EM Algorithm 71
4.3.3 When N is Large – the Modified EM Algorithm 73
4.3.4 Additional Implementation Considerations 75
4.4 Case Studies 79
4.4.1 The Ozone Data 79
4.4.2 The Diabetes Data 83
4.5 Extensions and Future Plans 86
4.6 Conclusions 88

Bibliography 90

LIST OF TABLES

3.1 The metabolites that were detected as significantly different between the control and the R132H groups, using the bivariate test, at the 5% FDR threshold. 62

4.1 Goodness of fit of the three models (Full, Selected, and "2.8" from [51]) in terms of AIC and mean absolute error (MAE). For the normal fit, we also compare the adjusted R^2 values 81
4.2 Parameter estimates for the ozone data with our 'Selected' model. 83
4.3 The selected model for the diabetes data. 84
4.4 The final model for the diabetes data, with $R^2 = 0.51$ 85

LIST OF FIGURES

2.1 Power as a function of the inflation factor, λ. The solid and dashed lines correspond to the random and fixed factor models, respectively. The thick and thin lines correspond to the normal and lognormal data, respectively. 24

2.2 Accuracy as a function of the inflation factor, λ. The solid and dashed lines correspond to the random and fixed factor models, respectively. The thick and thin lines correspond to the normal and lognormal data, respectively. 26

2.3 ROC curves, $\lambda = 4$. The solid and dashed lines correspond to the random and fixed factor models, respectively. The dotted lines correspond to the median centered robust version of Levene's test. 'Random classification' is represented by the dot-dashed line. 27

2.4 The distribution of $x_g = \log \frac{s_{2g}^2}{s_{1g}^2}$, case study 1: the Apo-A1 data set. Number of genes $G = 5,548$, sample size $n_1 = n_2 = 8$, FDR threshold=0.05. 30

2.5 Case study 2: methylation data set. Boxplots of $x_g = \log \frac{s_{2g}^2}{s_{1g}^2}$ in the three mixture components. Number of genes $G = 119,260$, sample size $n_1 = n_2 = 3$, FDR threshold=0.05. 32

2.6 Case study 3: fMRI data set. Boxplots of $x_g = \log \frac{s_{2g}^2}{s_{1g}^2}$ in the three mixture components. Number of voxels $G = 36,145$, sample size $n_1 = 29, n_2 = 22$, FDR threshold=0.05. 33

2.7 The distribution of $x_g = \log \frac{s_{2g}^2}{s_{1g}^2}$, case study 4: metabolomics data set. Number of metabolites $G = 225$, sample size $n_1 = n_2 = 10$, FDR threshold=0.05. 34

3.1 Type II probability in the bivariate case – the grey lines correspond to the variance-covariance matrix for nonnull genes, and the red ellipse corresponds to the $(1-\alpha)\%$ quantile for null genes. The power is the integral of the nonnull probability distribution function over the rejection region (outside the ellipse). 50

3.2 The power depends on the gene-specific realization of the error variance, σ_{1g}^2, since each one determines a difference acceptance region (the red ellipses) 51

3.3 The graphic user interface of the power/sample size calculation program 53

3.4 Output from the power/sample size calculation program 54

3.5 Matthews' Correlation Coefficient as a function of ψ (using the same configuration as in Figure 3.4) 56

3.6 The power to detect nonnull genes, using the univariate (LEMMA, [3]) and bivariate methods. In this example, $n_1 = n_2 = 6$ and the error variances are distributed $IG(\alpha = 2.1, \beta = 10/33)$. The differential expression of nonnull genes is distributed $N(3, 1)$, and the proportion of nonnull genes is 0.1. The two boxplots on the left show the power when there is no differential variation. The two boxplots on the right show the power when the variability of the response is 4 times greater in the treatment group for a set of 200 genes of which 100 are differentially expressed. 59

3.7 The bivariate distribution of $z_g = (d_g, x_g)'$ for the metabolite data in [66], comparing between the control group, and mutation R132H. Number of metabolites: $G = 109$. Sample sizes: $n_1 = n_2 = 3$. 61

3.8 The bivariate distribution of $z_g = (d_g, x_g)'$ for the comparison between the breast, and ovarian cancer cell lines. The blue diamonds show the 14 genes that are detected by both the bivariate model, and LEMMA (which assumes variance homogeneity across treatment groups). The red squares represent the 29 genes that were only discovered by the bivariate model, at the 0.05 FDR level. 64

4.1 Ozone data – diagnostics plots of the 'Full' version of the model in 4.26 (including all main effects). Left: deviance residuals vs. fitted values. Right: quantile-quantile plot of the deviance residuals. 82

4.2 Diabetes data – diagnostics plots of the final model in 4.27. Left: residuals vs. fitted values. Right: quantile-quantile plot of the residuals. 86

CHAPTER 1

INTRODUCTION

1.1 Overview

In recent years we have witnessed incredible technological advances which resulted in massive data sets. Furthermore, the rate and fidelity in which data are produced continue to increase, and the consensus is that this trend will persist, with more types of data, and at increasingly finer resolutions. For instance, less than a decade ago, advanced genetic research involved microarrays with thousands of assayed genes. Today, next-generation sequencing, quantitative trait loci (QTL) mapping, metabolomics, proteomics, and gene methylation analysis consist of tens, or hundreds of thousands of features, or variables of interest. In neuroscience, scientists hope that high-resolution technologies such as functional Magnetic Resonance Imaging (fMRI) and Diffusion Tensor Imaging (DTI) will provide insight into cognitive processes, memory, and mental conditions. These technologies usually yield millions of observations, since they consist of repeated measures over tens or hundreds of thousands of voxels in the brain. While these new technologies offer a much more detailed view of complex processes, they also introduce significant computational and statistical challenges. Traditional statistical methods are not adequate for the new paradigm in which the sample size is usually much smaller than the number of features (the 'large p, small n' problem). As a result, to meet these new challenges, statistical and computational methodologies are also evolving rapidly.

Changes in statistical methods are not limited just to new computational techniques. Even the most fundamental concepts are viewed through a differ-

ent prism. For example, for many years *shrinkage estimation* was referred to as 'the James-Stein paradox', and the original paper [50] was viewed mostly as a curious result in theoretical statistics. Today, this method is considered among the most powerful in high-throughput data analysis. More generally, it is now widely recognized is that in order to deal with the 'large p, small n' problem, one has to 'borrow strength' across genes, and an efficient way to do it, is via mixed-models where the differential expression of nonnull genes is assumed to be a realization of a random effect. Such models induce James-Stein shrinkage, and thus increase the power to detect significant genes.

Simulation-based techniques have also evolved rapidly in recent years, thanks to advances in computing and the need to handle a large number of hypotheses. Markov Chain Monte Carlo (MCMC, [20]) and the bootstrap [31] are two notable examples. Many computational problems in statistics can only be solved via approximations and computer simulations. This is especially true for high-dimensional data sets, where closed-form solutions to estimation problems are often not feasible.

In this dissertation we develop efficient and powerful statistical methods for high-dimensional data.

In Chapter 2 we deal with one of the most common research questions – which genes are affected by a treatment. Just like in the traditional two-sample test, the most common interpretation of this question is to estimate the effect of the condition on the *mean* of expression levels. We argue that the effect can also be on the *variance*, and develop a powerful method to detect this effect. The effect on the variance is increasingly recognized in recent biological literature. For instance, Hansen et al. [43] say that: "The increased across-sample variabil-

ity in methylation within the cancer samples of each tumor type compared to normal was even more striking than the differences in mean methylation." We show that our model is useful in a wide range of applications, including gene-expression, gene methylation, metabolomics, and brain imaging. We also show that our method is much more powerful than traditional tests for unequal variances, and that it is also robust to the normality assumptions imposed by the model.

In Chapter 3 we extend the ideas from Chapter 2 and previous work [3], and develop a novel approach to estimate the treatment effect on genes. We introduce a unified, *bivariate* normal model that accounts for both differential expression and differential variation between the two groups. We show that this model fits a wide range of data sets very well. Furthermore, we show in simulations that this new method yields a substantial gain in power when differential variation is present. Through a three-step estimation approach, in which we apply the Laplace approximation and the EM algorithm, we get a computationally efficient method, which is particularly well-suited for 'large p, small n' situations.

Chapter 4 deals with the problem of variable selection where the number of putative variables is large, possibly much larger than the sample size. We develop a model-based, empirical Bayes approach. By treating the putative variables as random effects, we get shrinkage estimation, which results in increased power and significantly faster convergence, compared with simulation-based methods. Furthermore, we employ a couple of computational tricks which allow us to increase the speed of our algorithm, to handle a very large number of putative variables, and to control the multicollinearity in the model. The

motivation for developing this approach is QTL analysis, but our method is applicable to a broad range of applications. We apply it to two widely-studied data sets, and show that our model selection algorithm yields excellent results.

1.2 The General Approach

All three chapters of this dissertation share the same principled approach. In this section we describe the general philosophy and methodology, in order to provide the 'big picture' for the rest of this manuscript. The principles described below can be divided into three categories: (i) the statistical model, (ii) the estimation and inference procedures, and (iii) computational considerations and tools.

1.2.1 The Statistical Model

The model-based approach: in any statistical analysis problem, it is a good idea to start with a model that can reasonably describe the data. Among the models that fit the data well, we tend to prefer the simplest one. This is known in the folklore as "Occam's razor" or "lex parsimoniae" (the law of parsimony). When dealing with very large data sets, this approach has the potential to greatly reduce the dimensionality and complexity of the estimation problem.

Random effects: in the context of detecting differential genes or in model selection problems where the number of candidate variables is very large, we find that the mixed-model approach is particularly appropriate. Rather than estimating thousands of fixed effects (for each gene individually), we assume that the

gene-specific effects are realizations of a parametric distribution. This reduces the number of estimated parameters significantly. In fact, the number of parameters in the mixed-model approach remains fixed, regardless of the number of tests. This feature contributes to an increase in power, and to the scalability of the computational algorithm.

Shrinkage estimation: integrating out the random effects in linear mixed models leads to shrinkage estimation [50] as was shown in [33]. Shrinkage estimators borrow strength across levels, and thus increase the power. This is particularly important when the sample size is small but the total number of observations is large (for example, we typically see microarray data sets consisting of tens of thousand of genes, but only a handful of individuals).

Mixture models: ultimately, our goal is to classify variables. For example, in gene expression analysis we want to identify differentially expressed genes versus 'null' genes. In variable selection algorithms we seek a binary classification of explanatory variables – either a variable is significant to the model, or it is not (and should be excluded). Incorporating these latent classification variables into the model is done via mixture models. Specifically, suppose that there are two classes, namely 'null' and 'nonnull' (or 'alternative', in the traditional two-sample terminology). In our parametric, model-based approach, we assume that the 'nulls' and the 'nonnulls' follow distribution functions f_0 and f_1, respectively. We define Bernoulli random variables $b_g \sim Ber(p)$, and say that the response follows a mixture distribution, so that $y \sim (1 - b_g)f_0 + b_g f_1$.

The advantage of using mixture models becomes apparent in the estimation and inference phase, where we have a single procedure to fit the data and determine the null status of each variable. This, again, contributes to the overall

power, since by accounting for the classification status, or more precisely, the null probability of each variable, we reduce the bias in the estimators.

For the remainder of this section it will be convenient to have a generic mixture model which includes random effects. Let f is the probability density function of the normal distribution, and let $b_g \sim Ber(p)$. We assume that the response is a mixture of two normal distributions:

$$y \sim (1 - b_g) f(\mu_0, \sigma^2_{0,g}) + b_g f(\mu_{1,g}, \sigma^2_{1,g}) . \tag{1.1}$$

We assume that $\mu_{1,g}$ are drawn from a normal distribution, and that $\sigma^2_{j,g}$ are drawn from an inverse gamma distribution, $IG(\alpha, \beta)$, for $j = 0, 1$.

This model seems to fit different types of data sets very well, including the logarithm of gene expression data, metabolite levels, etc. However, the methodology presented here is far more general. In some (non-normal) cases the derivations of the likelihood or the parameter estimates may be more difficult, but the principles presented here (parsimonious hierarchical mixture-models, involving random effects) remain just as useful and effective.

1.2.2 Estimation and Inference Procedures

The EM algorithm: the main objective in our analysis is to estimate the latent null status variables. When dealing with 'missing data' situations, one of the most powerful and computationally efficient methods is the Expectation-Maximization (EM) algorithm [29]. To apply this iterative algorithm, we write the complete data likelihood function, ignoring the fact that the null status variables are missing. Then, in the E-step we plug in their expected values, and

in the M-step we maximize with respect to the parameters in the model. We continue this process until some convergence criterion is met. In our case, the expected values are obtained by a simple application of Bayes rule:

$$Pr\left(b_g = 1\right) \;=\; \frac{p \cdot f(\mathbf{y}; \mu_{1,g}, \sigma^2_{1,g})}{p \cdot f(\mathbf{y}; \mu_{1,g}, \sigma^2_{1,g}) + (1-p) \cdot f(\mathbf{y}; \mu_0, \sigma^2_{0,g})}\,. \tag{1.2}$$

Frequentist inference – the False Discovery Rate: the parametric, model-based approach leads to an explicit likelihood function. Since the likelihood under the null is known, we can readily compute the p-values for each hypothesis, and select the significant ones, while controlling for the false discovery rate using the FDR algorithm [8].

Bayesian inference: the latent variables are estimated by their posterior distribution, and hence we can take an empirical-Bayesian inference approach and declare a test (e.g., gene) significant if its posterior null probability is below a certain threshold.

In practice, the two inferential approaches typically yield similar results. However, there is a philosophical difference, since the FDR-based method is based solely on the null distribution, whereas the empirical-Bayesian inferential procedure takes into account the form of the alternative, or nonnull distribution.

The estimation and inference approach described here falls into the 'empirical Bayes' framework, where the prior distributions are estimated from the data (via the EM algorithm). However, as was demonstrated in [5], the model-based approach and its hierarchical nature lends itself naturally to fully Bayesian implementations, via MCMC simulations [54]. While usually quite slower than the EM algorithm, the fully Bayesian method has two appealing futures. First, it is not necessary to integrate out random effects, which could be hard in some

configurations. Second, MCMC simulations yield posterior distribution, rather than point estimates for each parameter. This could be used to assess convergence, but more importantly to perform sensitivity analysis. However, we find that in terms of the end-result, the EM algorithm and the MCMC method give similar results. Therefore, considering the tradeoff between these two features of MCMC and the speed or the EM algorithm, the latter is preferred in the applications we are dealing with, since the computational efficiency is critical when dealing with so many tests (genes, voxels, etc.)

1.2.3 Computational Considerations and Tools

In the process of developing the models and the estimation procedures, it is important to keep in mind implementation considerations. For example, integrating out random effects (and thus inducing shrinkage estimation) is often an analytically-intractable problem. Furthermore, obtaining maximum likelihood estimators is straightforward in principle, but in practice one has to resort to numerical solutions. This subsection highlights the main computational tools used in this dissertation.

The Laplace approximation: in models such as (1.1), where the error variance is assumed to random, it is often not possible to integrate out the random effect analytically. However, a simple and accurate approximation is obtained via the Laplace approximation [27, 16]. For details specific to the applications in this dissertation, see [3]. Generally, suppose we have an unnormalized probability density, $p(x)$, for $x \in \Omega$, and we need to find its normalizing constant, $K = \int_\Omega p(x)dx$. Also suppose that $p(x)$ is unimodal, and has its mode at x_0. Then,

using a Taylor expansion of $\ln p(x)$, we can get an approximation of K:

$$\tilde{K} \quad = \quad p(x_0)\sqrt{2\pi/c}\,, \tag{1.3}$$

where

$$c \quad = \quad -\frac{\partial^2}{\partial x^2}\ln p(x)|_{x=x_0}\,. \tag{1.4}$$

In other words, we replace the integral with a function of the second derivative of the log-likelihood, evaluated at the mode.

Matrix algebra: many derivations are made simple if the right tools are used. In linear models, and especially normal models, matrix algebra is a very powerful tool. Quite often, a matrix representation of a linear model greatly simplifies the derivation of likelihood functions and maximum likelihood estimators. Many useful formulas in matrix algebra are provided in 'The Matrix Cookbook' [63]. In our application of the EM algorithm in Chapter 4, we also find the book 'Variance Components' [60] to be an indispensable resource. General computational considerations and algorithms pertaining to matrix algebra (as well as other techniques) are provided in [42].

Programming: statistical programs have become an essential component in the statistician's toolbox. In particular, we find that `R` [64] with its rich set of built-in distributions, operators and functions, specialized packages, and its vectorized computation philosophy, is particularly useful. For example, when dealing with complex data or models, it is often impossible to derive closed-form formulas, and the only way to compute estimators, evaluate likelihood functions, etc, is through optimization functions (e.g. `nlminb`.)

CHAPTER 2

A MIXTURE-MODEL APPROACH FOR PARALLEL TESTING FOR UNEQUAL VARIANCES

2.1 Introduction and Motivation

Testing for unequal variances is usually performed in order to check the validity of the assumptions that underlie standard tests for differences between means (the t-test and anova). However, existing methods for testing for unequal variances (Levene's test and Bartlett's test) are notoriously non-robust to normality assumptions, especially for small sample sizes. Moreover, although these methods were designed to deal with one hypothesis at a time, modern applications (such as to microarrays and fMRI experiments) often involve parallel testing over a large number of levels (genes or voxels). Moreover, in these settings a shift in variance may be biologically relevant, perhaps even more so than a change in the mean. This chapter introduces a parsimonious model for parallel testing of the equal variance hypothesis. It is designed to work well when the number of tests is large; typically much larger than the sample sizes. The tests are implemented using an empirical Bayes estimation procedure which 'borrows information' across levels. The method is shown to be quite robust to deviations from normality, and to substantially increase the power to detect differences in variance over the more traditional approaches even when the normality assumption is valid.

Research questions are often framed in terms of the effect a treatment has on a response. When comparing two conditions (say control and treatment) the question is typically interpreted in terms of the difference between the two

means. When the response is continuous, the most widely-used test to detect the effect of the treatment is the two-sample t-test. In this context, a test for unequal variances can be performed to assess the validity of the equal variance assumption. However, unlike the t-test, tests for equality of variances are notoriously non-robust, as highlighted by George Box's famous quote, "To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!" [12]. Reviews of the literature on testing equality of variances can be found in [10], [11] and [40].

Increasingly, however, there are new insights that suggest that biological variance plays an important role in determining cellular and organismal processes. This chapter deals with problems in which testing for unequal variances is of scientific importance in its own right. Furthermore, the focus is on situations in which a large number of parallel tests are conducted. A parsimonious model and empirical Bayes estimation procedure is developed that 'borrow strength' across the levels being tested. This results not only in increased power over standard tests when the normality assumption holds, but also in substantially improved performance when it does not. The wide applicability of the approach is illustrated using four different types of data sets: gene expression, gene methylation, functional Magnetic Resonance Imaging (fMRI), and metabolomics data. In these settings changes in the variance under the treatment are often biologically relevant. Moreover, failing to account for unequal variances can undermine the performance of methods for detecting for changes in the mean. We address this last point in the next chapter.

As a specific example, consider a simple Pavlovian-type learning experiment

in which the response is measured in terms of volume of blood flowing through voxels in the brain. Both control and treatment groups receive a simple visual signal in regular intervals. The subjects in the treatment group also receive an auditory stimulus in addition to the visual signal. Since the stimulus does not require complicated cognitive processing, it is conceivable that the overall mean response levels will not differ between the groups. However, in the treatment group, in preparation for the audio signal relevant areas in the brain might have smaller variability, to ensure the availability of the necessary level of blood for the anticipated task. Similarly, areas not involved in processing the audio signal might exhibit increased variability among the treatment subjects.

Variation, in genetic and phenotypic terms, has been thought to be a component of population fitness and adaptability. One way to interpret the association between expression variance and phenotype is to consider changes in pathways. If the genes in a particular pathway have very low variance, a natural interpretation is that those genes are highly constrained. Ho et al. [45] report that they "found that changes in expression variability are associated with changes in coexpression patterns. Therefore, differential variability is potentially an important manifestation of changes in gene regulation." Hansen et al. [43] say that "the increased across-sample variability in methylation within the cancer samples of each tumor type compared to normal was even more striking than the differences in mean methylation." Other recent examples where biological sources of variation play an important role in determining cellular and organismal processes can be found in [17, 21, 37, 38, 53, 55, 56, 57, 59, 62, 65].

The recent methodology proposed in [56] for assessing the variance of gene expression uses a one at a time analysis of the coefficient of variation. They

compute the coefficient of variation for each gene by dividing the standard deviation of its expression measures across a sample population by its average expression. They then designate low variance genes as those falling below the lower 25th percentile of the genome-wide coefficient of variation distribution based on all donors and high variance genes as those above the 75th percentile; those genes in the range between the 25th and 75th percentile they refer to as the mid variability gene set.

In the microarray context it is now widely recognized that methods that borrow strength across genes, by assuming that the gene-specific variances come form a common distribution, are more powerful for detecting mean treatment effects, [3, 49, 69]. However, these methods all assume variance homogeneity across conditions. We develop a new model-based approach to parallel testing for unequal variances that complements the existing methods for detecting changes in the mean. Although the methodology can be applied in a variety of settings, for simplicity of exposition, we use terminology from the microarray literature, and so the parallel tests concern variability in gene-specific expression in arrays based on samples from control and treatment groups.

Our model assumes that the ratio of the sample variances from the control and treatment groups arises from a three components mixture: a null component in which the ratio is proportional to an F-statistic; and two non-null groups representing inflated and deflated variance in the treatment group relative to the control. The three component mixture is identified by a latent multinomial random variable which is treated as missing data when fitting the model via the EM algorithm. Two variants of the model are considered: one in which the inflation/deflation factors are constant across all the parallel tests; and one

in which they are assumed to come from a log normal distribution. Genes are declared as non-null if their posterior null probability is less than a predefined threshold. Alternatively, frequentist inference can be conducted by controlling the false discovery rate using the estimated null distribution.

Our approach to determining high and low variance is in line with a growing literature on empirical Bayesian analysis of high dimensional data [3]. The hierarchical nature of our proposed method yields shrinkage estimation which results in high power and accuracy, while maintaining a low false discovery rate. Furthermore, as will see in Section 2.5, the inference based on our approach is quite robust.

The chapter is organized as follows. The mixture model is defined in Section 2.2. Section 2.3 outlines the details of the EM algorithm. The Bayesian and frequentist inference procedures are described in Section 2.4. A simulation study demonstrating the improved power, robustness, and accuracy of the method relative to 'one gene at a time' approaches is discussed in Section 2.5. Section 2.6 presents results from four case studies and some concluding remarks are given in Section 2.7.

2.2 The Mixture Model

Denote the (normalized) response for gene g in array j under condition i by y_{ijg}, and suppose that, given the gene-specific variances, σ^2_{1g} and σ^2_{2g},

$$y_{ijg} \sim N(\mu_{ig}, \sigma^2_{ig}) \tag{2.1}$$

independently, for all i, j and g, where $i = 1$ for arrays in the control group and $i = 2$ for the treatment group, $j = 1, \ldots, n_{ig}$, and $g = 1, \ldots, G$. Typically G is in the hundreds or thousands, whereas the sample sizes, n_{ig}, are much smaller, often only in the single digits.

The sample variance for gene g in condition i is given by

$$s_{ig}^2 = \sum_{j=1}^{n_{ig}} (y_{ijg} - \bar{y}_{i\cdot g})^2 / f_{ig}\,, \tag{2.2}$$

where $f_{ig} = n_{ig} - 1$. It follows from the normality assumption (2.1) that the ratio of variances in the control and treatment samples is proportional to a central F-statistic; that is,

$$r_g|\rho_g \sim \rho_g \frac{\chi_{f_{2g}}^2 / f_{2g}}{\chi_{f_{1g}}^2 / f_{1g}}\,, \tag{2.3}$$

where $r_g = s_{2g}^2/s_{1g}^2$ and $\rho_g = \sigma_{2g}^2/\sigma_{1g}^2$. In order to classify the genes as having the same, inflated or deflated variance under treatment we suppose that each ratio, ρ_g, $g = 1, \ldots, G$, is drawn from a three components mixture with probability vector, $\boldsymbol{p} = (p_0, p_1, p_2)$. Associated with each gene is a trivariate latent indicator vector $\boldsymbol{\delta}_g = (\delta_{0g}, \delta_{1g}, \delta_{2g})$ distributed as multinomial$(1, \boldsymbol{p})$ which determines whether the variance in the treatment group is null, inflated or deflated with respect to the control group. More specifically,

$$\rho_g|\delta_g, \lambda_g \sim \tau \lambda_g^{\delta_{1g} - \delta_{2g}}\,, \tag{2.4}$$

where $\lambda_g > 0$ is a gene-specific inflation/deflation factor, and the parameter τ allows for the incorporation of fixed covariate effects into the model. In the simplest case, with no covariates, τ represents a constant multiplicative difference between the variances in the control and treatment groups which is often noticeable in real data. For example, in fMRI data the stimulus presented to the

treatment group may affect subjects overall brain activity, and not just regions in the brain that are associated with the task.

We consider two variants of the model: a *fixed inflation factor* model in which $\lambda_g \equiv \lambda$, where λ is constant across all genes; and a *random inflation factor* model in which the λ_g's are assumed to come from a lognormal distribution. These assumptions both lead to inferences about the variance ratios that borrow strength across the genes, resulting in greater power to detect inflated or deflated variance under treatment.

The assumption of a lognormal distribution for λ_g can be motivated from the perspective of classical shrinkage estimation [50] and its connection to BLUPs arising in linear mixed models [33]. Specifically, consider the variable $x_g \equiv \log(r_g)$. Equations (2.3) and (2.4) imply that

$$x_g = \log \tau + (\delta_{1g} - \delta_{2g}) \log \lambda_g + \xi_{2g} - \xi_{1g} \tag{2.5}$$

where $\xi_{ig} = \log(\chi^2_{f_{ig}} / f_{ig})$, $i = 1, 2$, have known mean and variance given by $E(\xi_{ig}) = \psi(f_{ig}/2) - \log(f_{ig}/2)$ and $\mathrm{Var}(\xi_{ig}) = \psi'(f_{ig}/2)$, ψ and ψ' being the digamma and trigamma functions, respectively. Using independence and applying the delta method implies that $\xi_{2g} - \xi_{1g}$ is approximately normal with mean and variance given by

$$\theta_g = \psi(f_{2g}/2) - \log(f_{2g}/2) - \psi(f_{1g}/2) + \log(f_{1g}/2) \tag{2.6}$$

and

$$\kappa_g^2 = \psi'(f_{1g}/2) + \psi'(f_{2g}/2)\,. \tag{2.7}$$

Thus, if $\log \lambda_g \sim N(\theta, \kappa^2)$, equation (2.5) has the form of a mixture of linear mixed models, and shrinkage estimates of individual components of $\log \lambda_g$ can be estimated by their posterior expectations given the observed x_g, $g = 1, \ldots, G$ [33].

The random inflation factor model assumes that the logarithm of the sample variance is approximately normal. This may not be the case when the error distribution is misspecified, because generally, the distribution of the sample variance may depend on the population mean. Thus, in general it is not clear how the method described in this section will perform when the normality assumption does not hold. However, we see (empirically) in this section that (i) the normality assumption is quite reasonable for a wide range of applications; and (ii) even when the assumption does not hold, our methods performs quite well, and much better than one at a time methods. Furthermore, the effects of non-normality and heterogeneity of variances are investigated in [25, 26, 24]. In these references, it has been observed that the normal approximation to the log of the ratio between variances is very reasonable under very general departures from the null hypothesis of normality and variance homogeneity.

2.3 The EM Algorithm

2.3.1 Complete data log-likelihood

Regarding the latent indicator vector $\boldsymbol{\delta}_g$ as missing data, we obtain the complete data log-likelihood to implement the EM algorithm.

For the fixed inflation factor model where $\lambda_g \equiv \lambda$, the complete data log likelihood (omitting terms that do not depend on unknown parameters) is obtained directly from the identities (2.3) and (2.4) as

$$\sum_{g=1}^{G} \ell_F(r_g) \;=\; \sum_{g=1}^{G} \left\{ \sum_{k=0}^{2} \delta_{kg} \log p_k + \frac{f_{1g}}{2} \log\left(\tau \lambda^{\delta_{1g}-\delta_{2g}}\right) \right.$$

$$
\begin{aligned}
&\quad -\frac{f_{1g}+f_{2g}}{2}\log\left(\tau\lambda^{\delta_{1g}-\delta_{2g}}+r_g f_{2g}/f_{1g}\right)\Big\} \\
&= \sum_{g=1}^{G}\sum_{k=0}^{2}\delta_{kg}\log p_k+\sum_{g=1}^{G}\frac{f_{1g}}{2}\left\{\log\tau+(\delta_{1g}-\delta_{2g})\log\lambda\right\} \\
&\quad -\sum_{g=1}^{G}\frac{f_{1g}+f_{2g}}{2}\Big\{\delta_{0g}\log\left(\tau+r_g f_{2g}/f_{1g}\right) \\
&\qquad\qquad +\delta_{1g}\log\left(\tau\lambda+r_g f_{2g}/f_{1g}\right) \\
&\qquad\qquad +\delta_{2g}\log\left(\tau/\lambda+r_g f_{2g}/f_{1g}\right)\Big\}\,. \qquad (2.8)
\end{aligned}
$$

For the random inflation factor model, using the normal approximation to the log chi-squared distribution, and the mixed linear model representation in (2.5), we obtain

$$
\begin{aligned}
\sum_{g=1}^{G}\ell_R(x_g) &= \sum_{g=1}^{G}\sum_{k=0}^{2}\delta_{kg}\log p_k-\frac{1}{2}\sum_{g=1}^{G}\log[(\delta_{1g}-\delta_{2g})^2\kappa^2+\kappa_g^2)] \\
&\quad -\frac{1}{2}\sum_{g=1}^{G}\frac{[x_g-\mu_g-(\delta_{1g}-\delta_{2g})\theta]^2}{(\delta_{1g}-\delta_{2g})^2\kappa^2+\kappa_g^2} \\
&= \sum_{g=1}^{G}\sum_{k=0}^{2}\delta_{kg}\log p_k \\
&\quad -\frac{1}{2}\sum_{g=1}^{G}\left[\delta_{0g}\log(\kappa_g^2)+\delta_{1g}\log(\kappa^2+\kappa_g^2)+\delta_{2g}\log(\kappa^2+\kappa_g^2)\right] \\
&\quad -\frac{1}{2}\sum_{g=1}^{G}\delta_{0g}\frac{[x_g-\mu_g]^2}{\kappa_g^2}-\frac{1}{2}\sum_{g=1}^{G}\delta_{1g}\frac{[x_g-\mu_g-\theta]^2}{\kappa^2+\kappa_g^2} \\
&\quad -\frac{1}{2}\sum_{g=1}^{G}\delta_{2g}\frac{[x_g-\mu_g+\theta]^2}{\kappa^2+\kappa_g^2}\,, \qquad (2.9)
\end{aligned}
$$

where $\mu_g=\log\tau+\theta_g$ is the expected value of x_g in the null case (i.e. when $\delta_{0g}=1$).

2.3.2 The E-step

The E-step of the EM algorithm involves taking the expectation of the complete data log-likelihood conditional on the observed data. In the context of our mixture model, strict implementation of the E-step requires evaluating the expectation of all components of the complete data log-likelihood that are functions of the latent indicator, $\boldsymbol{\delta}_g$, $g = 1, \ldots, G$. In particular, if the complete data likelihood is linear in the latent indicator, as in (2.8) the E-step reduces to evaluating the posterior probabilities,

$$pr(\delta_{kg} = 1|r_g) = \frac{p_k L_k(r_g)}{\sum_{l=0}^{2} p_l L_l(r_g)}, \tag{2.10}$$

at the current iteration parameter estimates, where $L_k(r_g) = \exp\{\ell_F(r_g)\}$ with $\delta_{kg} = 1$. The same argument holds for the random inflation factor model with ℓ_R replacing ℓ_F in the posterior probability formula (2.10).

2.3.3 The M-Step

Let φ denote the complete vector of model parameters, and let $Q(\varphi, \varphi^{(t)}) = E_{\varphi^{(t)}}[\ell(\{r_g\})]$ denote the Q-function obtained by substituting the estimated posterior probabilities, $\hat{\delta}_{kg}^{(t)}$, after iteration t in (2.8) or (2.9). The M-step at the $(t+1)$st iteration involves maximization of $Q\left(\varphi, \varphi^{(t)}\right)$ with respect to each parameter in φ. That is,

$$\varphi^{(t+1)} = \arg\max_{\varphi} Q\left(\varphi, \varphi^{(t)}\right).$$

Maximization of the Q-function with respect to the multinomial probabilities is the same for both fixed and random inflation factor models, the update at

iteration $t+1$ being

$$\hat{p}_k^{(t+1)} = \frac{1}{G}\sum_{g=1}^{G}\hat{\delta}_{kg}^{(t)}. \tag{2.11}$$

The other parameters updates depend on the assumptions regarding the inflation factors.

M-Step: Fixed Inflation Factor

Differentiating (2.8) results in the following update equations for τ and λ, respectively:

$$\sum_{g=1}^{G}\frac{f_{1g}f_{2g}\left(r_g - \tau\lambda^{\hat{\delta}_{1g}-\hat{\delta}_{2g}}\right)}{f_{2g}r_g + f_{1g}\tau\lambda^{\hat{\delta}_{1g}-\hat{\delta}_{2g}}} = 0. \tag{2.12}$$

and

$$\sum_{g=1}^{G}\frac{f_{1g}f_{2g}(\hat{\delta}_{1g}-\hat{\delta}_{2g})\left(r_g - \tau\lambda^{\hat{\delta}_{1g}-\hat{\delta}_{2g}}\right)}{f_{2g}r_g + f_{1g}\tau\lambda^{\hat{\delta}_{1g}-\hat{\delta}_{2g}}} = 0\,. \tag{2.13}$$

If $\hat{\delta}_{1g} = \hat{\delta}_{2g} = 0$ for all g, set $\hat{\lambda} = 1$.

M-Step: Random Inflation Factor

Differentiating (2.9) results in the update equations for τ, θ and κ^2:

$$\log\hat{\tau} = \frac{\sum_{g=1}^{G}\left[\frac{\delta_{0g}(x_g-\theta_g)}{\kappa_g^2} + \frac{(\delta_{1g}+\delta_{2g})(x_g-\theta_g)+(\delta_{2g}-\delta_{1g})\theta}{\kappa^2+\kappa_g^2}\right]}{\sum_{g=1}^{G}\left(\frac{\delta_{0g}}{\kappa_g^2} + \frac{\delta_{1g}+\delta_{2g}}{\kappa^2+\kappa_g^2}\right)} \tag{2.14}$$

$$\hat{\theta} = \frac{\sum_{g=1}^{G}(\delta_{1g}-\delta_{2g})\frac{x_g-\mu_g}{\kappa^2+\kappa_g^2}}{\sum_{g=1}^{G}\frac{\delta_{1g}+\delta_{2g}}{\kappa^2+\kappa_g^2}}, \tag{2.15}$$

and

$$\hat{\kappa}^2 \;=\; \frac{\sum_{g=1}^{G}\delta_{1g}\left[(x_g-\mu_g-\theta)^2-\kappa_g^2\right] + \delta_{2g}\left[(x_g-\mu_g+\theta)^2-\kappa_g^2\right]}{\sum_{g=1}^{G}(\delta_{1g}+\delta_{2g})}, \tag{2.16}$$

and $\hat{\theta} = \hat{\kappa} = 0$ if $\hat{\delta}_{1g} = \hat{\delta}_{2g} = 0$ for all g.

2.4 Inference

2.4.1 The Frequentist and Bayesian Procedures

Our model-based approach allows us to assess the null status of a hypothesis, either using a frequentist procedure based on false discovery rate (FDR, [8]); or using empirical Bayes inference via the posterior null probabilities.

Under the fixed inflation factor model the statistic, r_g/τ has an F-distribution under the null. In this case the frequentist p-value for gene g is equal to $pr(\tau F < r_{g,obs})$ if $r_g/\tau < 1$ and $pr(\tau F > r_{g,obs})$ if $r_g/\tau > 1$, where $F \sim F(f_{2g}, f_{1g})$. For the random inflation factor model the corresponding p-value is given by $pr\left(|Z| > (x_{g,obs} - \mu_g)/\kappa_g\right)$, where Z is a standard normal variate.

The empirical-Bayesian approach is to classify genes based on the estimated posterior probabilities, $\hat{\delta}_{kg}$, $k = 0, 1, 2$. Thus, a gene is declared nonnull if either $\hat{\delta}_{1g}$ or $\hat{\delta}_{2g}$ exceed a given threshold.

2.4.2 Shrinkage Estimation

For the random factor model, the posterior probability of $\delta_{1g} = 1$ can be rewritten in form

$$pr(\delta_{1g} = 1|x_g) = \frac{1}{\frac{p_0 \cdot L_0(x_g)}{p_1 \cdot L_1(x_g)} + 1 + \frac{p_2 \cdot L_2(x_g)}{p_1 \cdot L_1(x_g)}},$$

where the ratio L_0/L_1 is given by

$$\frac{L_0(x_g)}{L_1(x_g)} = \frac{(2\pi\kappa_g^2)^{-1/2} \exp\left\{-(x_g - \mu_g)^2/2\kappa_g^2\right\}}{[2\pi(\kappa^2 + \kappa_g^2)]^{-1/2} \exp\left\{-(x_g - \mu_g - \theta)^2/2(\kappa^2 + \kappa_g^2)\right\}}$$

$$
\begin{aligned}
&= (1-c_g)^{-1/2} \exp\left\{-\frac{1}{2}\frac{[c_g(x_g-\mu_g)+(1-c_g)\theta]^2}{c_g\kappa_g^2} + \frac{\theta^2}{2\kappa^2}\right\} \\
&\propto (1-c_g)^{-1/2} \exp\left\{-\frac{1}{2}T_g^2\right\}, \qquad (2.17)
\end{aligned}
$$

say, with the constant of proportionality being $\exp(\theta^2/2\kappa^2)$, and where

$$
c_g = \frac{1}{\kappa_g^2}\left(\frac{1}{\kappa_g^2}+\frac{1}{\kappa^2}\right)^{-1} = \frac{1}{1+\kappa_g^2/\kappa^2}\,.
$$

Similarly, for the other likelihood ratio we have

$$
\frac{L_2(x_g)}{L_1(x_g)} = \exp\left\{-\frac{2(x_g-\mu_g)\theta}{\kappa^2+\kappa_g^2}\right\}. \qquad (2.18)
$$

Suppose that $\theta > 0$, so that δ_{1g} is an indicator of inflated variance. Then, L_2/L_1 converges to zero as x_g increases to infinity and so, in the limit, δ_{1g} is solely a function of the ratio, L_0/L_1, and hence of the statistic, T_g. On the other hand, L_2/L_1 converges to infinity as x_g decreases to $-\infty$ so that δ_{1g} converges to zero. This makes sense since, in this case, it is highly unlikely that the gene is in the inflated variance nonnull group. Parallel arguments can be made regarding δ_{2g}.

Note that $x_g - \mu_g$ is the observed difference between the log variances in the control and treatment groups for gene g (after adjusting for the covariate effects), and θ represents the expected difference if the gene has inflated variance under treatment (assuming $\theta > 0$). Thus, the numerator of the statistic, T_g, has the form of classical James-Stein shrinkage estimator of difference in the log variances, with the amount of shrinkage of the observed difference towards θ determined by the ratio of variances κ_g^2/κ^2.

2.5 Simulation Results

We compared the performance of the two estimation procedures in terms of power, accuracy, and false discovery rate with the 'one hypothesis at a time' approach, using the median centered robust version of Levene's test [15, 52]. We chose the Levene test since previous studies have shown it to be relatively robust and powerful [40]. We also compared our estimation and inference procedure with other well-known 'one at a time' methods, like Bartlett's test [7]. The traditional methods that do not borrow strength across levels lack power, especially when the sample sizes are small. For a comprehensive review of 'one at a time' methods and their power and robustness properties, see, for example, [10] or [11].

Similarly, methods that do borrow strength across hypotheses but are looking for significant differences between the means of the two groups (for example, LEMMA [3], Limma [69]) perform very poorly in terms of power and accuracy. It should be noted that by design, the mean-based tests excel when there is a difference in the mean response between the group, and are expected to have low power and accuracy when the mean is not significantly different, but the variance is.

In our simulations we varied the sample sizes, ranging from $n_i = 2$ to $n_i = 30$ and allowed for the two groups to have different sample sizes. We also varied the inflation factor, so that $2 \leq \lambda \leq 10$. We also used a variety of underlying distributions, including normal, Cauchy, lognormal, and exponential, in order to assess robustness. Each simulation configuration was repeated 20 times, and the results are reported in terms of the mean of the 20 experiments. The config-

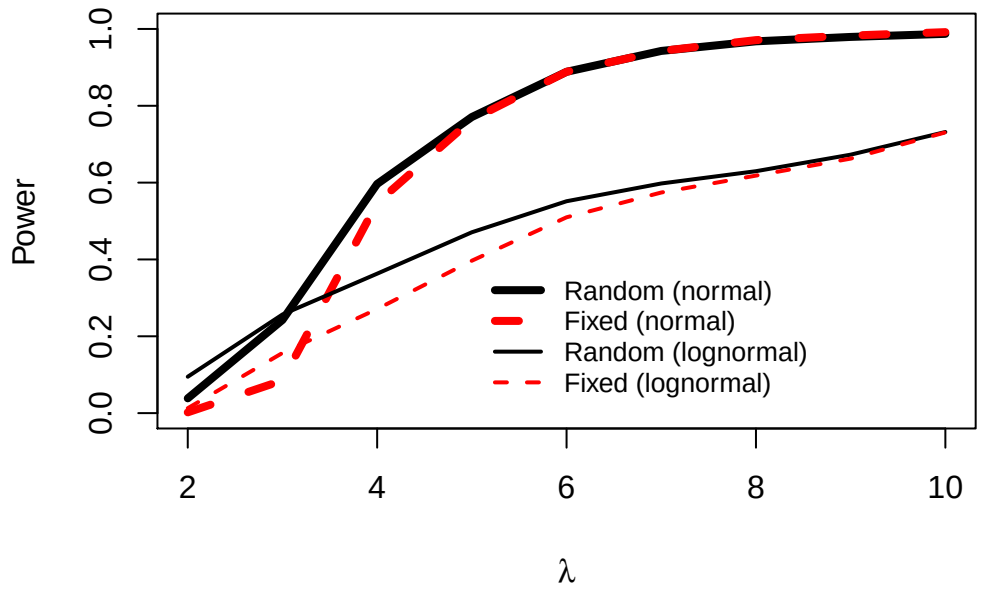

Figure 2.1: Power as a function of the inflation factor, λ. The solid and dashed lines correspond to the random and fixed factor models, respectively. The thick and thin lines correspond to the normal and lognormal data, respectively.

urations reported in this chapter consist of $n_1 = 4$, $n_2 = 7$, the total number of 'genes' is $G = 2000$ and the proportion of the inflated-variance subset is $p = 0.1$. The underlying distributions of the response, y_{ijg} are $N(0, 0.25)$ and $LN(0, 0.25)$, and Cauchy distribution, with location and scale parameters equal to 0 and 0.1, respectively. The results reported here are representative of the wide range of simulation studies that we performed.

Figure 2.1 shows the power of the random inflation factor model (solid line) and the fixed inflation factor model (dashed line) for two configurations: one, for normal data, $y_{null} \sim N(0, 0.25)$ (thick lines), and one for lognormal data, $y_{null} \sim LN(0, 0.25)$ (thin lines). To generate these power plots we used the frequentist-type inference, and controlled for false discovery rate at the 5% level. The 'one hypothesis at a time' approach which uses the Levene test and the

mean-based approach (using `lemma`) yielded no discoveries for any λ (after applying the Benjamini-Hochberg adjustment), and are not included in the plot. The 'random inflation factor' is more powerful than the 'fixed factor' procedure, in both configurations. As expected, as the true inflation factor increases both procedures become more powerful. Also, the power is higher when the underlying data are normally distributed.

Of course, in addition to power we would like the methods to have high level of accuracy (the total percentage of correct classifications, i.e., 100×(True Positive + True Negative)/G). Figure 2.2 shows (for normal and lognormal data) that both methods are quite accurate, and their accuracy increases as the inflation factor increases. In contrast, the conservative one-at-a-time approach, as well as the mean-based methods (not shown in the plot), yield approximately constant level of accuracy (in this case, 0.9, since by not rejecting **any** test, it correctly classifies all the null subset.)

In terms of false discovery rate, the fixed factor approach is more conservative, and has a lower FDR for all λ (but also less power). When the data are normally distributed and $\lambda > 2$ both methods have low false discovery rate. For smaller values of λ the false discovery rate is higher, especially with the random factor method (approximately 0.25, vs. 0.1 with the fixed factor method.) When the data are not normally distributed, $y \sim LN(0, 0.25)$, the false discovery rate increases quite dramatically for small values of λ, especially with the random factor method.

ROC curves (of the average true positive rate versus the false positive rate) are given in Figure 2.3 for the normal, lognormal, and Cauchy data, when the inflation factor is $\lambda = 4$. The three ROC plots are confined to a false positive rate

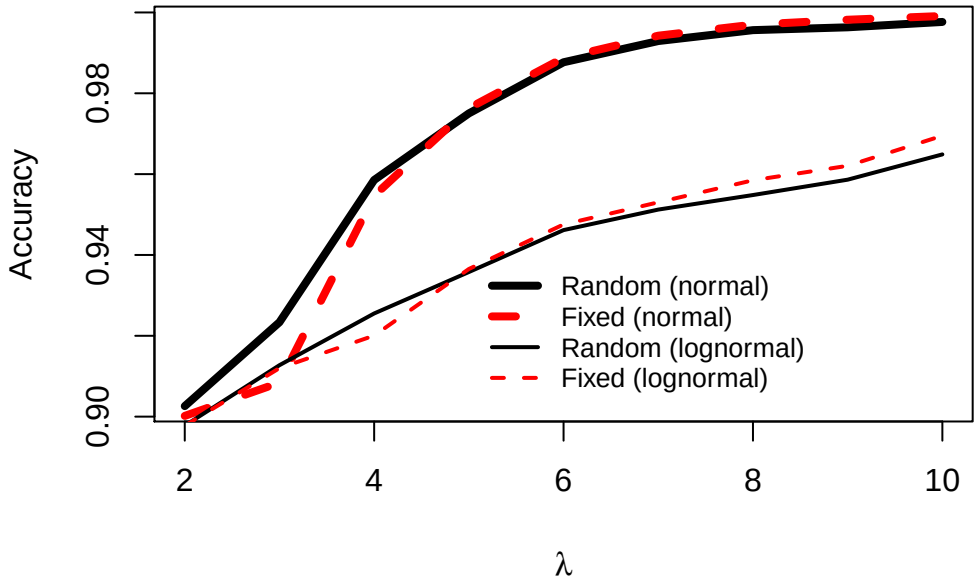

Figure 2.2: Accuracy as a function of the inflation factor, λ. The solid and dashed lines correspond to the random and fixed factor models, respectively. The thick and thin lines correspond to the normal and lognormal data, respectively.

of less than or equal to 0.2 since higher error rates than this would clearly be undesirable. In all cases the random factor model has the best performance. For example, when the data are normal, at a false positive level of 0.05 the average true positive rates are approximately 0.2, 0.5, and 0.65 for the (median-centered) Levene, fixed inflation factor, and random factor methods, respectively.

When the data are normal or lognormal, both the random and fixed factor models are much better than the median centered robust version of Levene's test. In particular, the middle plot shows that Levene's test is not at all robust to the normality assumption, as its ROC curve falls below the 'random classification' line (in grey). In contrast, under the lognormal data generation scheme, both the random and fixed factor models are quite robust. Furthermore, for all the simulated distributions the performance of our methods improves as the

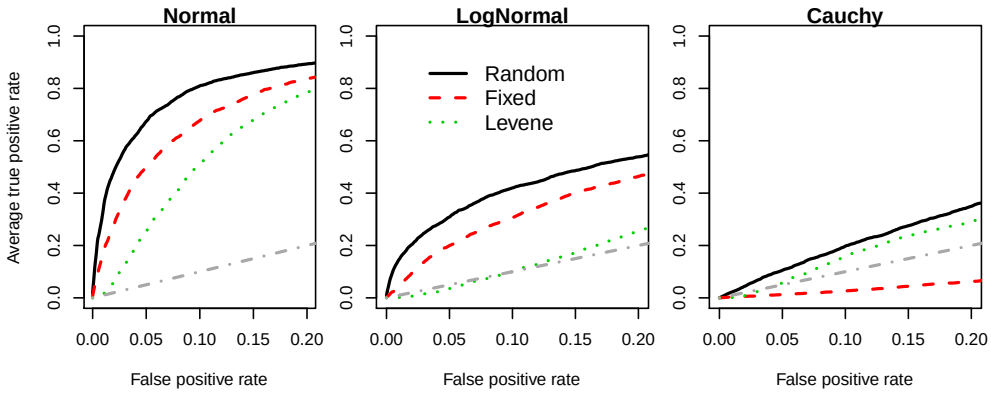

Figure 2.3: ROC curves, $\lambda = 4$. The solid and dashed lines correspond to the random and fixed factor models, respectively. The dotted lines correspond to the median centered robust version of Levene's test. 'Random classification' is represented by the dot-dashed line.

inflation factor increases, but the Levene method does not exhibit any improvement (not shown in the plot).

The fixed factor method performs very poorly with Cauchy data (right panel). In fact, it is even worse than the Levene method. An explanation is that the estimate of λ is not consistent because the mean of the Cauchy distribution does not exist, so the fixed factor model is clearly not appropriate in this extreme case. In contrast, the random factor model allows for variability in the distribution of the inflation factor, and is able to detect a reasonable number of the genes with differential variance while maintaining a low false positive rate.

Note that the interpretation of the ROC plot requires care. It seems that for the normal data the 'one at a time' method has comparable performance to that of our model-based approach, since the Levene-based ROC curve is below the other two thick curves, but above the diagonal (in grey). However, it merely depicts that for a certain threshold of the p-values, the number of true positives exceed the number of false positives. In practice, the thresholds used to plot

the ROC curve for the Levene test are much too high to be practical in real-life applications, since, as we discussed above, the 'one at a time' method yields no discoveries at any reasonable FDR threshold when the number of tests is large.

2.6 Case Studies

We consider four different statistical applications to genetics and molecular biology to demonstrate the wide range of data sets to which our method can be applied. In all cases we find strong evidence that there is a subset of the data in which the variance in the treatment group is significantly higher or lower than in the control group but there is no significant difference between the means. The first case study involves a gene expression data set. The second deals with epigenetic data (methylation), while the third uses data from a brain imaging experiment (functional MRI data). The final example concerns metabolomics data.

The results in this section illustrate two things that are relevant to our previous derivations. First, we see that the observed distributions of the statistics r_g and x_g in the applications considered are very close to the ones in our model. In particular, the normal approximation of x_g appears to be very appropriate. Second, when the overall mean does not change due to the treatment, but the variance does, our method is able to detect it. In that sense, it complements the mean-based methods, which would (most likely) fail to detect the change in variance, unless it is coupled with a significant change in the mean response.

2.6.1 Microarray Data

Callow et al. [18] used gene targeting in embryonic stem cells to produce mice lacking apolipoprotein A-1, a gene known to play a critical role in high density lipoprotein (HDL) cholesterol levels. In our analysis, we used the data and normalization method provided with the `limma` R package [70], which consists of 5,548 ESTs, from eight control (wild type "black six") mice and eight "knockout" (lacking ApoA1) mice. Common reference RNA was obtained by pooling RNA from the control mice, and was used to perform expression profiling for all 16 mice. Using the `lemma` package [4], which is designed to detect genes that are differentially expressed, 9 genes are detected (with a 0.2 posterior probability threshold) including the ApoA1 gene and others closely related to it. The same set of the top eight genes were also identified as nonnull (among others) when using other (mean-based) packages like `limma` and `locfdr` [36]. These genes were confirmed to be differentially expressed in the knockout versus the control line by an independent assay.

Applying the method in this chapter while controlling the false discovery rate at 5% we find 21 genes in which the variance in the treatment group was significantly higher than in the control, and 21 genes in which the variance was significantly smaller in the treatment group. Most of these genes had very small mean-response difference (defined as $d_g = \bar{y}_{2\cdot g} - \bar{y}_{1\cdot g}$) and were not detected by any mean-based method, or by 'one at a time' test for unequal variance.

Figure 2.4 shows the distribution of the statistics $\{x_g\}$. The scatter plot on the left shows the genes with significantly higher and lower variance, marked by upper red or lower blue triangles, respectively. The scatter plot and the histogram (right) show that the normal approximation fits the distribution of x_g

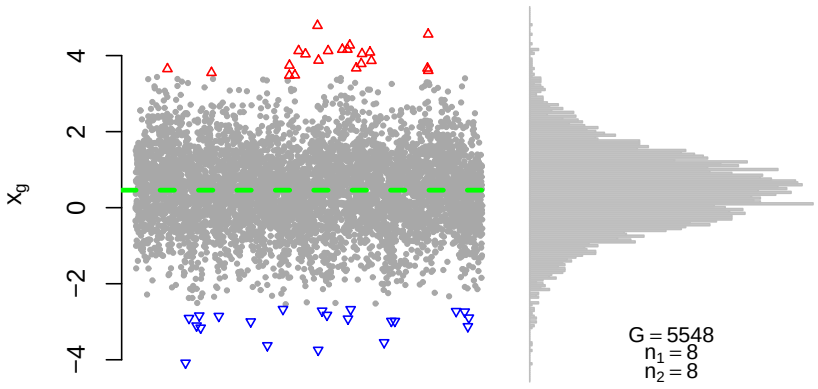

Figure 2.4: The distribution of $x_g = \log \frac{s_{2g}^2}{s_{1g}^2}$, case study 1: the Apo-A1 data set. Number of genes $G = 5,548$, sample size $n_1 = n_2 = 8$, FDR threshold=0.05.

very well. The green dashed line represents the overall mean of x_g (which, in our previous notation, we referred to as $\log(\tau)$.)

We investigated the functional status of the genes that had deflated and inflated variances using the National Institute of Health Gene tool and Genomenet (`http://www.ncbi.nlm.nih.gov/gene` and `http://www.genome.jp/`, respectively). It turns out that the inflated variance genes mostly have to do with cell signaling, while the deflated variance genes seem more related to tighter regulation of a lipid metabolism gene network. Given that the gene of primary focus in the study, ApoA1, encodes apolipoprotein A-I, which is the major protein component of high density lipoprotein (HDL) in plasma these results are biologically plausible.

2.6.2 Methylation Data

DNA methylation plays an important role in regulation of gene expression. Recent studies have shown that hyper- or hypomethylation are associated with cancer (either as a causal effect or as an early indicator of the disease). In the following analysis, we used an unpublished data set with 119,260 genes, and three subjects in each group. Using the mean-based approach [3] we did not find any significantly hyper or hypomethylated genes. However, applying the methods developed in this chapter we found a total of 153 genes with inflated methylation, and 150 with deflated methylation (at the FDR level of 5%). In contrast, traditional 'one at at time' methods yield no discoveries, after accounting for multiple testing.

The observed mean differences (d_g) are rather small, but the observed log-ratio between the mean squared errors were very large (in absolute value) for some genes. Figure 2.5 shows the boxplots of the three mixture components. The distribution of x_g in the null component is approximately normal with mean 0, and the significant genes have $|x_g| > 7$. Recall that x_g is on the logarithmic scale, so for the significant genes this corresponds to at least four orders of magnitudes in the ratio between the mean squared errors between the two groups.

2.6.3 fMRI Data

Functional magnetic resonance imaging (fMRI) is used to measure the change in blood flow in the brain during certain neural or cognitive activity. In this example we use data from a Pavlovian-type experiment, in which both groups were shown a visual cue, but for the treatment group it was immediately fol-

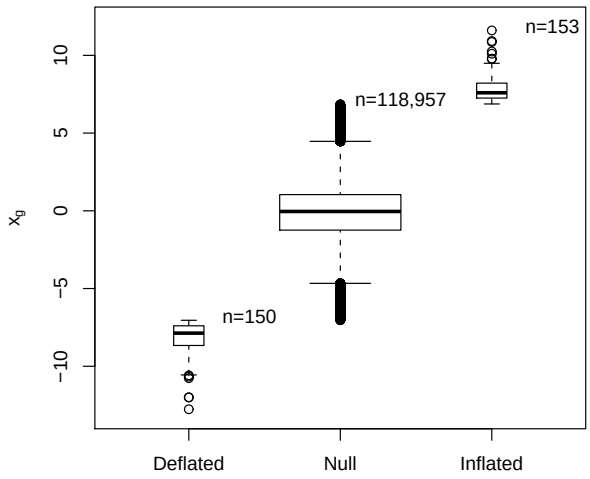

Figure 2.5: Case study 2: methylation data set. Boxplots of $x_g = \log \frac{s_{2g}^2}{s_{1g}^2}$ in the three mixture components. Number of genes $G = 119,260$, sample size $n_1 = n_2 = 3$, FDR threshold=0.05.

lowed by an auditory signal [71]. One of the goals of the experiment was to test whether after several training cycles there will be a difference in the response to the visual cue between the two groups, and if so, in which region of the brain. According to the Pavlovian paradigm, it is expected that once trained, the treated subjects will respond to the visual cue as if they receive the auditory cue. For more details about the experiment, see the 'Supporting Online Material' document in [71].

Again, no voxels were found to have significantly different mean levels of response when using mean-based methods. However, we do find many voxels which exhibit significantly different levels of variability. Figure 2.6 shows the boxplots of the three mixture components that our method identified from a total of 36,145 voxels. A total of 1,276 voxels had a significantly increased variance in the treatment group, and 1,142 voxels had a significantly decreased variance in the treatment group.

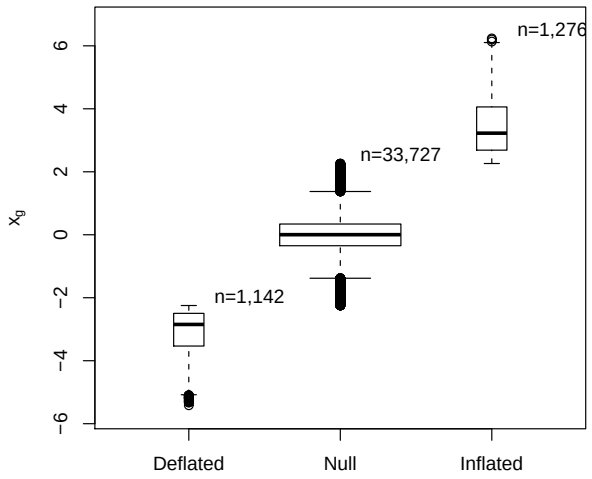

Figure 2.6: Case study 3: fMRI data set. Boxplots of $x_g = \log \frac{s_{2g}^2}{s_{1g}^2}$ in the three mixture components. Number of voxels $G = 36,145$, sample size $n_1 = 29, n_2 = 22$, FDR threshold=0.05.

2.6.4 Metabolomics Data

Our final example uses data from the area of metabolomics ("the study of unique chemical fingerprints that specific cellular processes leave behind"). In this (unpublished) experiment, two groups of pregnant women were treated with two different levels of choline. The levels of nearly 250 metabolites were measured during the first and the twelfth weeks of the pregnancy. Here, we analyze the effect of the treatment on metabolite levels after 12 weeks (taking week 0 as the baseline for each woman). Once again, testing for differences in mean response levels between the groups yields no discoveries. However, with our method we found four metabolites whose variance increased significantly due to the treatment, and seven whose variance decreased (see Figure 3.7).

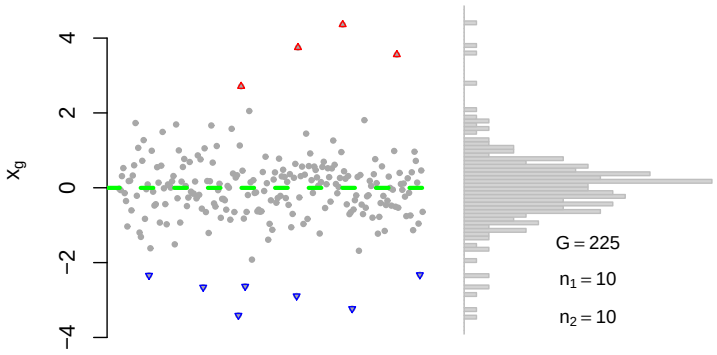

Figure 2.7: The distribution of $x_g = \log \frac{s^2_{2g}}{s^2_{1g}}$, case study 4: metabolomics data set. Number of metabolites $G = 225$, sample size $n_1 = n_2 = 10$, FDR threshold=0.05.

2.7 Conclusions

This chapter introduced a new model and an estimation procedure (based on the EM algorithm) for parallel testing for inequality of variances. The model borrows strength across the entire data, resulting in increased power and accuracy, while maintaining a low false discovery rate. Simulations show that the method performs well even when the number of tests is very large and the sample sizes are small, and that it is quite robust to deviations from normality. The analysis of four different data sets shows that the model assumptions are realistic, that the method is broadly applicable, and that it complements methods that test for differences in means.

CHAPTER 3

THE BIVARIATE MODEL – SIMULTANEOUS TEST FOR MEAN AND VARIANCE

3.1 Introduction

Recent advances in technology allow researchers to measure responses at increasingly finer resolutions. For example, microarrays are used to measure expression levels for thousands of genes, and functional Magnetic Resonance Imaging (fMRI) is used to measure volume of blood-flow in tens or hundreds of thousands of voxels in the brain. The large number of responses, combined with the high cost of such experiments introduces a severe multiple-testing problem, since the number of hypotheses is much larger than the sample size (the so-called 'large p, small n' problem). Although the method developed in this chapter can be used in a broad range of applications, we use the terminology of gene expression experiments in the rest of this chapter, for convenience of the exposition.

To address the problem of multiple testing, methods that 'borrow strength' across genes have been developed. Some of the most powerful methods gain power by cleverly reducing the dimensionality of the problem. To do that, they assume a certain parsimonious model that governs the distribution of the responses (see, for example [3, 49, 69]). This model-based approach yields shrinkage estimation of a small number of parameters, involving all the data, and thus borrowing strength across genes. One aspect in which these methods often differ, is in the way the random error variance is modeled. The most powerful among them assume that the error variances are also generated by a random

process, and they estimate the hyper-parameters of the (assumed) underlying distributions. However, these methods all assume variance homogeneity across treatment groups. In the previous chapter we introduced a different approach to modeling the distribution of the error variances. According to our model, the error variance in the control group follows an inverse gamma distribution, but in the treatment group the error variances follow a three-component mixture distribution. In the null component the error variances follow the same inverse gamma distribution, up to a constant factor. In the other two components, the error variances are additionally inflated or deflated by a random factor.

Here, we introduce a unified model that accounts for both differential expression and differential variation between the groups. Combining the approach in [3] for estimation of differential expression with the model for differential variation in Chapter 2, we derive a bivariate normal model for the mean-difference and the logarithm of the ratio of the error variances. According to the model, genes in the treatment group may exhibit higher or lower expression levels compared with the control, or the two groups may differ in their gene-specific error variances.

To fit the bivariate model to gene expression data, we use an empirical Bayes approach and the EM algorithm. To set up the EM algorithm, we define the 'missing data' as pairs of independent indicator variables that encode the differential expression and variation status of each gene. The nonnull components are assumed to be realizations of normal distributions. Therefore, to compute the complete data log likelihood, these random effects, as well as the random error variance, have to be integrated out. Although an analytical integration is not tractable, we develop a three step procedure which is both computationally

efficient, and powerful. First, to estimate the differential variation parameters we show that they do not depend on the differential expression parameters, and use the method in Chapter 2. Second, we define the re-weighted mean squared errors, which take the differential variation into account, and estimate the parameters of the error variance distribution, similarly to [3]. Finally, we apply the Laplace approximation to estimate the differential expression parameters, by plugging in the posterior mode of the gene-specific error variances.

The null status of genes with respect to differential expression, variation, or both, is determined by the posterior distribution of the aforementioned indicator variables. Alternatively, the bivariate model allows for a simple frequentist inference approach, while controlling the false discovery rate.

We show that the unified model increases the power to detect differentially expressed genes, because the shrinkage estimation of the error variance is more accurate than existing methods. Furthermore, unlike existing methods, the bivariate model detects differential variation, which is sometimes of scientific importance in its own right.

This chapter is organized as follows. In Section 3.2 we introduce the bivariate mixture model. In Section 3.3 we derive the estimation procedure based on the Laplace-approximated EM algorithm. Section 3.4 deals with the frequentist and Bayesian inference procedures, and in Section 3.5 we describe a method for power/sample size estimation. In Section 3.6 we describe results from a simulation study. Section 3.7 includes two case studies, and we conclude with Section 3.8.

3.2 The Model

Using the same notation and assumptions from Chapter 2, suppose that the response for gene g for subject j in treatment i, follows a normal distribution:

$$y_{ijg} \sim N\left(\mu_{ig}, \sigma^2_{ig}\right), \tag{3.1}$$

with $i = 1, 2$, $j = 1, \ldots, n_{ig}$ and $g = 1, \ldots, G$. We denote the sample mean of gene g in treatment group i by m_{ig}. We also assume that σ^2_{1g} are distributed inverse-gamma, with shape and scale parameters α, β.

Under model (3.1) the within-group sample mean is

$$m_{ig} = \frac{1}{n_{ig}} \sum_{j=1}^{n_{ig}} y_{ijg} \sim N\left(\mu_{ig}, \sigma^2_{ig}/n_{ig}\right). \tag{3.2}$$

Denote the observed difference between the treatment groups for gene g by $d_g \equiv m_{2g} - m_{1g}$. We assume that d_g are drawn from a mixture of three distributions, with probabilities $\mathbf{q} = (q_0, q_1, q_2)$. In particular, we assume that for non-differentially expressed genes $\mu_{2g} - \mu_{1g} = \nu$, and for differentially expressed genes $\mu_{2g} - \mu_{1g} = \nu + \left(\gamma_{1g} - \gamma_{2g}\right)\psi_g$, where

$$\psi_g | \boldsymbol{\gamma}_g \sim N(\psi, \sigma^2_\psi), \tag{3.3}$$

$$\boldsymbol{\gamma} = (\gamma_{0g}, \gamma_{1g}, \gamma_{2g}) \sim \mathit{multinomial}(1, \mathbf{q}). \tag{3.4}$$

In Chapter 2 we defined the sample variance for gene g in group i, $s2_{ig}$ (2.2), and we defined λ_g which we assumed satisfies $\log \lambda_g \sim N(\theta, \kappa^2)$. We also defined $\boldsymbol{\delta}_g = (\delta_{0g}, \delta_{1g}, \delta_{2g})$, a vector distributed as multinomial$(1, \boldsymbol{p})$ which determines whether the variance in the two groups is the same (up to a constant factor, τ), or if it is inflated or deflated in the treatment group (relative to the control).

Define the random variable $z_g = (d_g, x_g)'$, where x_g is as defined in (2.5). Then, since m_{ig} and s_{ig}^2 are independent, the distribution of z_g, conditional on $\psi_g, \boldsymbol{\gamma}_g, \lambda_g$ and $\boldsymbol{\delta}_g$, is bivariate normal:

$$z_g \sim N\left(\begin{pmatrix} \nu + \left(\gamma_{1g} - \gamma_{2g}\right)\psi_g \\ \theta_g + \log\tau + \left(\delta_{1g} - \delta_{2g}\right)\log\lambda_g \end{pmatrix}, \begin{pmatrix} \frac{\sigma_{1g}^2}{n_{1g}} + \frac{\sigma_{2g}^2}{n_{2g}} & 0 \\ 0 & \kappa_g^2 \end{pmatrix}\right) \tag{3.5}$$

Note that for null genes $\gamma_{0g} = \delta_{0g} = 1$, and the distribution of z_g is given by

$$z_g|\left(\gamma_{0g} = \delta_{0g} = 1\right) \sim N\left(\begin{pmatrix} \nu \\ \theta_g + \log\tau \end{pmatrix}, \begin{pmatrix} \frac{\sigma_{1g}^2}{n_{1g}} + \frac{\tau\sigma_{1g}^2}{n_{2g}} & 0 \\ 0 & \kappa_g^2 \end{pmatrix}\right). \tag{3.6}$$

We will use this later to construct the test to determine the null status of genes.

It is also worth noting in other hierarchical models that incorporate random effects to detect differentially expressed genes (e.g. LEMMA [3], LIMMA [69]) the variance of d_g in the nonnull distribution is larger than in the null group. This is simply a result of the commonly used assumption that the error variance distribution is the same across the two conditions. Thus, when integrating out the random effect, ψ_g, the total variance is the sum of the random error and the variance of the random effect. Occasionally, this does not seem to be justified when analyzing real data. In contrast, our model allows for some (nonnull) genes to have a deflated variance in the treatment group (by a factor of λ_g). Consequently, the total variance of d_g for such genes is smaller compared with the null genes. Even for null genes, if the data are generated by a process in which the overall inflation factor τ is not 1 and it is not accounted for by the estimation procedure, the null variance may be biased. In particular if the error variance of the null is over estimated, the test procedure will lack power to detect nonnull genes.

3.3 Estimation

3.3.1 The Laplace Approximation

The bivariate distribution of z_g given in (3.5) involves unobserved variables $\vartheta_g = \{\boldsymbol{\gamma}_g, \boldsymbol{\delta}_g, \psi_g, \lambda_g\}$. Thus, to estimate the set of parameters in the model, we turn to the EM algorithm [29], with ϑ_g playing the role of the 'missing data'. Denote the set of parameters in the model by $\varphi = \{\mathbf{p}, \mathbf{q}, \nu, \tau, \psi, \sigma_\psi^2, \theta, \kappa^2, \alpha, \beta\}$. In the E-step of the algorithm, we compute the expectation of the log likelihood, conditional on the current parameter values at the t-th iteration, and the observed data:

$$Q\left(\varphi|\varphi^{(t)}\right) = E_{\vartheta_g|z_g,\varphi^{(t)}}\left[\log L\left(\varphi; z_g, \vartheta_g\right)\right]. \tag{3.7}$$

The complete data log likelihood is a non-linear function of the latent indicator variables $\boldsymbol{\gamma}_g$ and $\boldsymbol{\delta}_g$, making the integration with respect to these variables analytically intractable. We solve this problem by updating $\boldsymbol{\gamma}_g$ and $\boldsymbol{\delta}_g$ by their posterior expectations using Bayes rule in each iteration. Using these posterior expectations, we obtain estimates for $\mathbf{p}$ and $\mathbf{q}$. For $k = 0, 1, 2$,

$$\hat{p}_k^{(t+1)} = \frac{1}{G}\sum_{g=1}^{G}\hat{\delta}_{kg}^{(t)}, \tag{3.8}$$

$$\hat{q}_k^{(t+1)} = \frac{1}{G}\sum_{g=1}^{G}\hat{\gamma}_{kg}^{(t)}. \tag{3.9}$$

In order to estimate the other parameters in φ we need to integrate out the random variables ψ_g, λ_g, and σ_{1g}^2 in (3.7). Now, under the assumption of variance homogeneity across treatment groups, where $\tau = 1$ and $\lambda_g = 1$ for all g, the distribution in (3.5) reduces to a univariate mixture model for the differential expression (d_g). The reduced model, in which $\sigma_{\epsilon,g}^2 \equiv \sigma_{1g}^2 = \sigma_{2g}^2$, and

$\sigma^2_{\epsilon,g} \sim invGamma(\alpha, \beta)$ is identical to the 'RR' model in [3]. Integration in (3.7) with respect to $\sigma^2_{\epsilon,g}$ is also intractable. To address this problem [3] apply the Laplace approximation, replacing the integral with respect to $\sigma^2_{\epsilon,g}$ with a function of its posterior mode. This yields an approximated version of complete data likelihood, which is very accurate and computationally efficient. This is followed by straightforward integration with respect to ψ_g, which is assumed to be normally distributed. However, in the more general case, the random variables $\{\sigma^2_{2g}\}$ depend on the variational null-status of the genes ($\boldsymbol{\delta}_g$), and are not assumed to follow the same distribution as $\{\sigma^2_{1g}\}$. Therefore, a direct extension of the Laplace approximation approach is not possible.

Furthermore, the parameter τ and latent variables $\log \lambda_g$ enter into the full (bivariate) complete data loglikelihood in a nonlinear way that makes exact application of the EM algorithm intractable. Thus, we take a three-step estimation approach, involving two simple and accurate approximations.

Step I: The distribution of x_g does not depend on estimating the mean parameters. Also, recall that κ^2_g and θ_g are known, since they depend only on the sample sizes. Hence, to compute $Q\left(\varphi|\varphi^{(t)}\right)$, we first integrate the complete data log likelihood of just x_g with respect to λ_g. Then, we obtain maximum likelihood estimates for the variance-inflation parameters τ, θ, κ^2, and $\boldsymbol{\delta}_g$, based only on the statistics x_g. The method is described in detail in [6].
Note that the distribution of z_g as given in (3.5) suggests that d_g also contains information about τ, θ, κ^2, and $\boldsymbol{\delta}_g$. However, since G is large, the approximated EM estimates of these parameters based on just x_g, are very accurate.

Step II: Define $f_g = n_{1g} + n_{2g} - 2$, and the re-weighted mean square error,

$$s_g^2 = \frac{1}{f_g}\left[\sum_{j=1}^{n_{1g}}\left(y_{1jg} - \bar{y}_{1\cdot g}\right)^2 + \sum_{j=1}^{n_{2g}}\frac{\left(y_{2jg} - \bar{y}_{2\cdot g}\right)^2}{\tau\lambda_g^{\delta_{1g}-\delta_{2g}}}\right]. \tag{3.10}$$

Plugging in the estimate for τ and the posterior means of λ_g and δ_{kg} from Step I into (3.10), we obtain an approximation for s_g^2. Now, $(x_g, \log\lambda_g)'$ follows a bivariate normal distribution with mean $\mathbf{m}$ and variance-covariance matrix S, such that $\mathbf{m} = (\log\tau + (\delta_{1g} - \delta_{2g})\theta, (\delta_{1g} - \delta_{2g})\theta)'$, $S_{1,1} = \kappa^2 + \kappa_g^2$, and $S_{1,2} = S_{2,1} = S_{2,2} = \kappa^2$. Hence, conditional on x_g, the posterior mean of λ_g is $\mathcal{E}\left(\lambda_g|\exp(x_g)\right) = \exp\left[\mathcal{E}\left(\log\lambda_g|x_g\right) + \frac{1}{2}Var\left(\log\lambda_g|x_g\right)\right]$. Let $\hat{\eta}_g = \hat{\kappa}^2/(\hat{\kappa}^2 + \kappa_g^2)$. Then, the posterior mean of λ_g can be written as:

$$\lambda_g|x_g, \boldsymbol{\delta}_g = \exp\left[(1-\eta_g)(\delta_{1g} - \delta_{2g})\theta + \eta_g(x_g - \log\tau - \theta_g + \kappa_g^2/2)\right]. \tag{3.11}$$

Now, the normality assumption in (3.1) implies that $s_g^2|\sigma_{1g}^2 \sim \sigma_{1g}^2\chi_{f_g}^2/f_g$. Combining this with the assumption $\sigma_{1g}^2 \sim IG(\alpha,\beta)$, we obtain a likelihood function for estimation of α and β. Maximum likelihood estimates are derived, for example, in [3] (Section 3). Alternatively, the method of moments results in closed-form estimates of α and β. Similar derivations appear in [69] based on the argument that the mean square errors (assuming variance homogeneity) follow a scaled F-distribution. Let $M_1 = \sum_{g=1}^G s_g^2/G$ and $M_2 = \sum_{g=1}^G (s_g^2)^2/G$. Then

$$\hat{\alpha} = \frac{2M_2 - (1+2/f)M_1^2}{M_2 - (1+2/f)M_1^2} \tag{3.12}$$

$$\hat{\beta} = \frac{1}{M_1(\alpha-1)}. \tag{3.13}$$

Step III: To obtain estimates for the mean parameters, ν, ψ, σ_ψ^2 and $\boldsymbol{\gamma}_g$, we need to integrate out the random variables ψ_g in the likelihood function of d_g. To do that, we apply the Laplace approximation by replacing the integral with

respect to σ^2_{1g} with the posterior mode, which we compute using the estimates from Step II. The resulting estimation equations for ν, ψ, σ^2_ψ and $\boldsymbol{\gamma}_g$ are the same as in [3], except for the plug-in estimator for σ^2_{1g}, which now takes into account the possibility that the variance in the treatment group is not the same as in the control, and that some genes may have inflated/deflated variance. Specifically, the posterior mode of the random error variance conditional on s^2_g, denoted by $\tilde{\sigma}^2_{1,g}$ satisfies

$$\tilde{\sigma}^2_{1,g} = \frac{s^2_g f_g/2 + 1/\hat{\beta}}{f_g/2 + \hat{\alpha} + 1}\,. \tag{3.14}$$

We now let $\tilde{\sigma}^2_g = \tilde{\sigma}^2_{1,g}\left(\frac{1}{n_{1g}} + \frac{\hat{\tau}\hat{\lambda}^{\hat{\delta}_{1g}-\hat{\delta}_{2g}}}{n_{2g}}\right)$ and write the Laplace approximated complete data log-likelihood function of d_g:

$$\begin{aligned}
\sum_{g=1}^{G} \ell(d_g) \;=\; & \sum_{g=1}^{G}\sum_{k=0}^{2} \gamma_{kg} \log q_k \\
& -\frac{1}{2}\sum_{g=1}^{G}\left[\gamma_{0g}\log(\tilde{\sigma}^2_g) + (\gamma_{1g}+\gamma_{2g})\log(\sigma^2_\psi + \tilde{\sigma}^2_g)\right] \\
& -\frac{1}{2}\sum_{g=1}^{G}\gamma_{0g}\frac{(d_g-\nu)^2}{\tilde{\sigma}^2_g} - \frac{1}{2}\sum_{g=1}^{G}\gamma_{1g}\frac{(d_g-\nu-\psi)^2}{\sigma^2_\psi+\tilde{\sigma}^2_g} \\
& -\frac{1}{2}\sum_{g=1}^{G}\gamma_{2g}\frac{(d_g-\nu+\psi)^2}{\sigma^2_\psi+\tilde{\sigma}^2_g}.
\end{aligned} \tag{3.15}$$

See [3] for the detailed derivation of estimates of the mean parameters from equation (3.15).

3.3.2 Incorporating Gene-Specific Priors

High-throughput experiments (as in genomics, metabolomics, and fMRI for brain imaging) often involve thousands of tests and there is no prior knowledge about which gene is differentially expressed. However, as more information is collected from other experiments one may have a good idea whether

some genes are more likely to be differentially expressed. Model (3.1) allows to incorporate gene-specific priors to account for such knowledge. For simplicity, our focus in this subsection is on priors on the differential expression latent variables, $\boldsymbol{\gamma}_g$, which appear in the likelihood of d_g, but the same derivations apply to the differential variational latent variables $\boldsymbol{\delta}_g$.

Suppose that the user specifies a subset of genes that are thought to be differentially expressed. We denote the set by A_1 and its complement by $A_0 \equiv \{1, \ldots, G\} \setminus A_1$. To incorporate this prior information, we change the assumption regarding the distribution of the latent variables, $\boldsymbol{\gamma}_g$, so that if $g \in A_1$ then $\boldsymbol{\gamma}_g \sim multinomial(1, \mathbf{q}_{A_1})$, and if $g \in A_0$ then $\boldsymbol{\gamma}_g \sim multinomial(1, \mathbf{q}_{A_0})$. The first term in log-likelihood function in (3.15) changes slightly, and we have

$$\sum_{g \in A_1} \sum_{k=0}^{2} \gamma_{kg} \log q_{A_1 k} + \sum_{g \in A_0} \sum_{k=0}^{2} \gamma_{kg} \log q_{A_0 k}.$$

The estimation of the latent variables $\boldsymbol{\gamma}_g$ is based on their posterior probabilities, conditional on whether the gene is in A_0 or A_1:

$$pr(\gamma_{kg} = 1 | d_g, g \in A_0) = \frac{q_{A_0 k} L_k(d_g)}{q_{A_0 0} L_0(d_g) + q_{A_0 1} L_1(d_g) + q_{A_0 2} L_2(d_g)},$$
$$pr(\gamma_{kg} = 1 | d_g, g \in A_1) = \frac{q_{A_1 k} L_k(d_g)}{q_{A_1 0} L_0(d_g) + q_{A_1 1} L_1(d_g) + q_{A_1 2} L_2(d_g)}.$$

To estimate the vectors $\mathbf{q}_{A_0}$ and $\mathbf{q}_{A_1}$, we simply average the posterior probabilities in group A_0 and A_1, respectively:

$$\hat{q}_{A_0 k}^{(t+1)} = \frac{1}{|A_0|} \sum_{g \in A_0} \hat{\gamma}_{kg}^{(t)}$$
$$\hat{q}_{A_1 k}^{(t+1)} = \frac{1}{|A_1|} \sum_{g \in A_1} \hat{\gamma}_{kg}^{(t)}.$$

This formulation generalizes trivially to any partition of the set of genes $\{1, \ldots, G\}$.

3.4 Inference

3.4.1 The Frequentist Approach

As we pointed out earlier, the null distribution of z_g is known, and given by equation 3.6. Let the null mean and variance-covariance matrix be $\boldsymbol{\mu}_{0g}$ and Σ_{0g}, respectively. Thus, applying theorem 5.2.1a in [58] we construct a test for the hypothesis that a gene is in the null group (that is, it is neither differentially expressed, nor is its variance inflated or deflated in the treatment group.) Specifically, under the (gene-specific) null distribution,

$$(z_g - \boldsymbol{\mu}_{0g})' \Sigma_{0g}^{-1} (z_g - \boldsymbol{\mu}_{0g}) \sim \chi_2^2. \tag{3.16}$$

Sometimes one has to consider specific alternatives in order to know more precisely why a gene has been rejected from being null. For example, is it differentially expressed, or is it because of differential variation, or perhaps both? This can be achieved by means of linear contrasts. In this setting the null hypothesis takes the form $H_0 : \mathbf{R}\boldsymbol{\mu} = \mathbf{r}$ where $\mathbf{R}$ is the contrast matrix. The test statistics are $(\mathbf{R}z_g - \mathbf{r})'(\mathbf{R}\Sigma_{0g}\mathbf{R}')^{-1}(\mathbf{R}z_g - \mathbf{r})$, and under the null they are distributed χ_1^2.

For example, to test for differentially expressed genes, we set $\mathbf{R} = (1, 0)$ and $\mathbf{r} = (0, 0)'$. Similarly, to test for differential variation we set $\mathbf{R} = (0, 1)$ and $\mathbf{r} = (0, 0)'$.

Another possible alternative hypothesis is that a gene is both differentially expressed and has inflated variance in the treatment group, and furthermore, that for nonnull genes $\boldsymbol{\mu}_{A,g} = (c_1, c_2)'$ for some known constants, c_1, c_2. For ex-

ample, it is common practice to require that for nonnull genes the minimum log fold-change is greater than some threshold, say, $c_1 = 1.5$. In this case, the appropriate contrast matrix is $\mathbf{R} = (1/c_1, -1/c_2)$. (See theorem 5.3.1a in [58]).

To account for the large number of hypotheses, we apply the Benjamini-Hochberg procedure [8] to the set of G p-values obtained from these chi-square tests, and control the false discovery rate at the desired level. If in addition we test L contrasts, we divide the desired false discovery threshold level by L before applying the Benjamini-Hochberg procedure to the p-values of each contrast.

3.4.2 A Note on Multiple Testing

The Benjamini-Hochberg procedure [8] guarantees that the expected false discovery rate is controlled at a given threshold. By focusing on the false discovery rate, rather than trying to control the Type-I error for each test, this method provides a more powerful approach in modern applications. Previous multiple testing approaches (e.g. Bonferroni [46], Dunnett [30], Hsu [48]) are suitable for cases in which the total number of tests is small to moderate, but typically yield no discoveries when the number of hypotheses is large, as is the case in most experiments in genomics.

Denote the number of null and nonnull genes by N and m, respectively, and let $M = N + m$. We show that even the FDR procedure becomes weaker as $m/M \to 0$. Suppose that the researcher includes a large number of null genes, and a small number of nonnull genes, of which g_* has the smallest p-value, denoted by p_*. The Benjamini-Hochberg procedure declares as significant the first k (sorted) p-valued, for which $p_{(i)} < i \cdot \alpha/M$, and in particular the smallest p-

value has to satisfy $p_{(1)} < \alpha/M$. So for a sufficiently large N, we have $\alpha/M < p_*$, and $N \approx M$. Now, the distribution of the smallest order statistics among the null genes is known, and we have $P\left(p_{(1)} > \frac{\alpha}{M}\right) = \left(1 - \frac{\alpha}{M}\right)^M \approx exp(-\alpha) \approx 1 - \alpha$. In other words, for a sufficiently large number of null genes, the probability that the procedure will not detect any genes, and therefore not detect g_*, is $1 - \alpha$.

We return to these observations in Section 3.5, where we elaborate on power and sample size.

3.4.3 The Empirical Bayesian Approach

The bivariate model and the EM-based estimation procedure allow for empirical-Bayesian inference in terms of posterior probabilities or the 'local fdr' [35]. Specifically, each gene can be classified into one of nine possible categories, depending on the values of the pairs $(\gamma_g, \delta_g) \in \{-1, 0, 1\} \times \{-1, 0, 1\}$, and a gene is in the null group if and only if $\gamma_g = \delta_g = 0$. The posterior null probability of a gene is given by

$$P_{0,g} \equiv \frac{p_0 f_0(z_g)}{\sum_{c,d} p_{(c,d)} f_{(\gamma_g=c,\delta_g=d)}(z_g)} \equiv \frac{p_0 f_0(z_g)}{f(z_g)} \tag{3.17}$$

where c and d equal -1, 0 or 1; $p_{(c,d)}$ is the joint probability of $(\gamma_g = c, \delta_g = d)$; and $f_{(\gamma_g=c,\delta_g=d)}(z_g)$ is the probability distribution function at z_g under the bivariate model (3.1) with $\gamma_g = c$ and $\delta_g = d$. We say that gene g is nonnull if its posterior null probability is less than a given threshold, u.

Using the cumulative distribution functions instead of the p.d.f, the statistical inference can be done based on the 'Bayesian FDR', defined in [34], so that $Fdr(z_g) \equiv p_0 F_0(z_g)/F(z_g)$.

The posterior probability formulation allows us to classify genes by their latent variable status. In particular, for any gene that is declared nonnull by the criterion $P_{0,g} < u$, we infer that they are differentially expressed if $\gamma_g \neq 0$, and/or have inflated/deflated variance in the treatment group if $\delta_g \neq 0$.

We now return to the issue of multiple testing, this time in the context of a Bayesian analysis. Müller et al. [61] say that "Posterior inference adjusts for multiplicities, and no further adjustment is required", but add that in order for this statement to hold, the prior probability of non-differential expression must be positive and non-degenerate for each gene. This is obviously the case in our model, where the prior is estimated by borrowing information across genes, and is not assumed to be fixed throughout the estimation algorithm.

Both inference methods (the frequentist and the empirical-Bayesian) use the same random-effects model that induces shrinkage, and increases power. The frequentist inference depends only on the null distributions, whereas the empirical-Bayesian approach takes into account the distributions of the nonnull genes, as well as the proportion of null group.

3.5 Power and Sample Size

Recent advances in bioinformatics prompted an impressive growth in data throughput, and a radical change in the statistical analysis paradigm, since analysts are now required to perform thousands of simultaneous tests using relatively small sample sizes. The large number of tests makes it essential to control the false discovery rate (FDR). Over the last couple of decades, methods that involve 'borrowing information across genes' greatly improved the power to

detect the so-called 'nonnull genes', while maintaining a low FDR.

Paradoxically, despite the consensus in the literature that testing 'one gene at a time' has a substantially lower power compared to modern methods that borrow information across genes, when designing an experiment with thousands of tests, the sample-size estimation is still, in many applications, based on the traditional 'one at a time' approach. Determining the required sample size based on a single gene, while adjusting the critical value to account for multiple testing, yields inflated, and usually unrealistic estimated sample sizes.

The cost of new high-throughput technologies and the need to ensure minimal power require new experiment-design methods that will provide a better sample size estimation.

We develop power and sample size estimation procedures that parallel the estimation and inference method described in this chapter. Specifically, we enable 'borrowing information across genes' in the design phase, by assuming a parametric model for the null and nonnull genes, and by assuming that gene-specific parameters are actually realizations of prior distributions. We note that in general, the only meaningful sample size (or power) estimation procedures are ones which rely on the same model assumptions and estimation procedures that are actually used in the discovery phase.

In the remainder of this section we assume that the sample size is n in both groups. Under the null model the bivariate mean is $(\nu, \log\tau)$ and the variance matrix is $diag(\sigma^2_{1_g}(1+\tau)/n, 2\psi'((n-1)/2))$.

Determining the power or sample size under the bivariate model is complicated for two reasons.

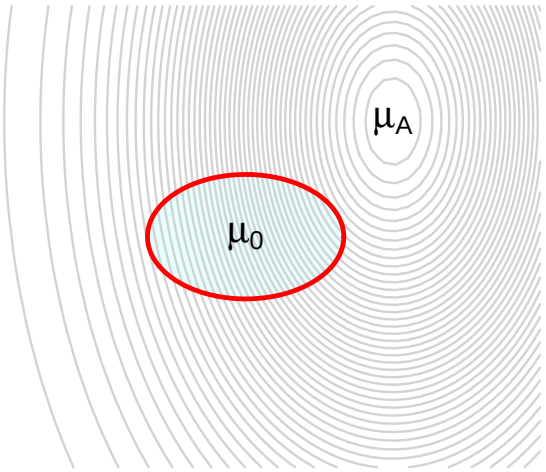

Figure 3.1: Type II probability in the bivariate case – the grey lines correspond to the variance-covariance matrix for nonnull genes, and the red ellipse corresponds to the $(1-\alpha)\%$ quantile for null genes. The power is the integral of the nonnull probability distribution function over the rejection region (outside the ellipse).

First, unlike the well-known, univariate situation, computing the probability of a type-II error in the bivariate case is known to be a hard problem for which there is no closed-form solution. To help illustrate this difficulty, consider Figure 3.1, where $\boldsymbol{\mu}_0$ is the hypothesized mean under the null, i.e. $(\nu, \log \tau)'$; and $\boldsymbol{\mu}_A$ is the hypothesized mean for nonnull genes. The variance-covariance matrix under a normal model is translated graphically into ellipse-shaped contour lines, which determine the quantiles of the bivariate distribution. For instance, the grey lines correspond to the variance-covariance matrix for nonnull genes, and the red ellipse corresponds to the $(1-\alpha)\%$ quantile for null genes (and note that The variance-covariance matrix for null genes is not assumed to be the same as for nonnull.) Computing the probability of a type-II error amounts to integrating the probability distribution function of the alternative in the acceptance region of the null (i.e., inside the red ellipse). This requires the computation of incom-

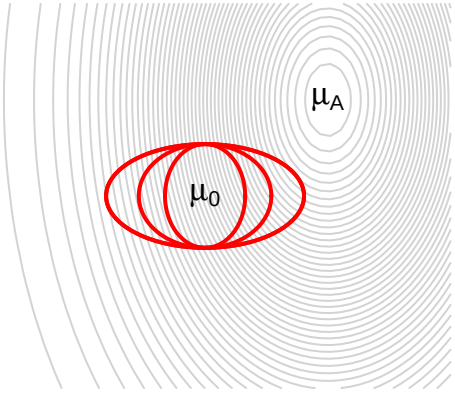

Figure 3.2: The power depends on the gene-specific realization of the error variance, σ^2_{1g}, since each one determines a difference acceptance region (the red ellipses)

plete elliptic integrals within the acceptance region of the null. To address this difficulty, we can compute the Type II probability numerically by summing the probability density function of the alternative on a fine grid over the acceptance region.

The second complication is due to the fact that the null distribution depends on a random variable, namely σ^2_{1g}. This is illustrated in Figure (3.2), where three different realizations of σ^2_{1g} are represented by the red, thick ellipses. Clearly, the probability of Type II error will be different for every realization, and we need to account for that in our power or sample size computation. One way of dealing with it is to take the mean (or the mode) of the random distribution from which the error variances are drawn.

Instead of attempting to evaluate the integral analytically or numerically, we take a Bayesian approach as in [72]. In this setting, the researcher specifies the parameter values $(\alpha, \beta, \nu, \tau, \psi, \sigma^2_\psi, \theta, \kappa^2)$ and the prior distributions in (3.5), and we generate K independent values of σ^2_{1g}, ψ_g, and λ_g. For each $k = 1, \ldots, K$ we

compute the *posterior sample size*, $\{n_k\}$, if the desired probability of Type II error is specified, or the *posterior power*, $\{\beta_k\}$, if the sample size is given.

The graphical user interface for the power or sample size analysis is shown in Figure 3.3. It is written in `R` [64] (using the package `tcltk`) and runs on Windows and Linux. Note that unlike the traditional power calculation tools, we have an element of randomness here since the error variance is specified as an inverse gamma distribution, rather than a fixed value, and the gene-specific differential and log-variational expression are normally distributed. The user specifies both the differential and log-variational expression mean and variance parameters and the proportions of nonnull genes. The user can choose a variable and specify it in terms of a range of values, in the form `from:by:start`. For instance, in this figure the mean differential expression is given by `0.5:0.5:4`, so the power analysis is performed for $\psi = 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4$.

The 'Simulation variables' section of the power analysis interface contains three variables. The first is simply the number of simulated genes. The second is the number of bootstrap iterations. Since we simulate data rather than solve for power analytically, we need to be able to estimate the variability in the results that is due to the simulation. Therefore, we use non-parametric bootstrap, and obtain confidence intervals or predictive distributions for any estimated quantity (e.g., power, accuracy, false discovery rate). Finally, the third field in this section is the FDR threshold that is used to determine the null status of the genes.

Figure 3.4 shows an example of the output from the power/sample size analysis tool. The parameters were set up as follows. The sample sizes are $n = 3$ for both groups. The inverse Gaussian mean and variance are 0.5 and 0.1 respec-

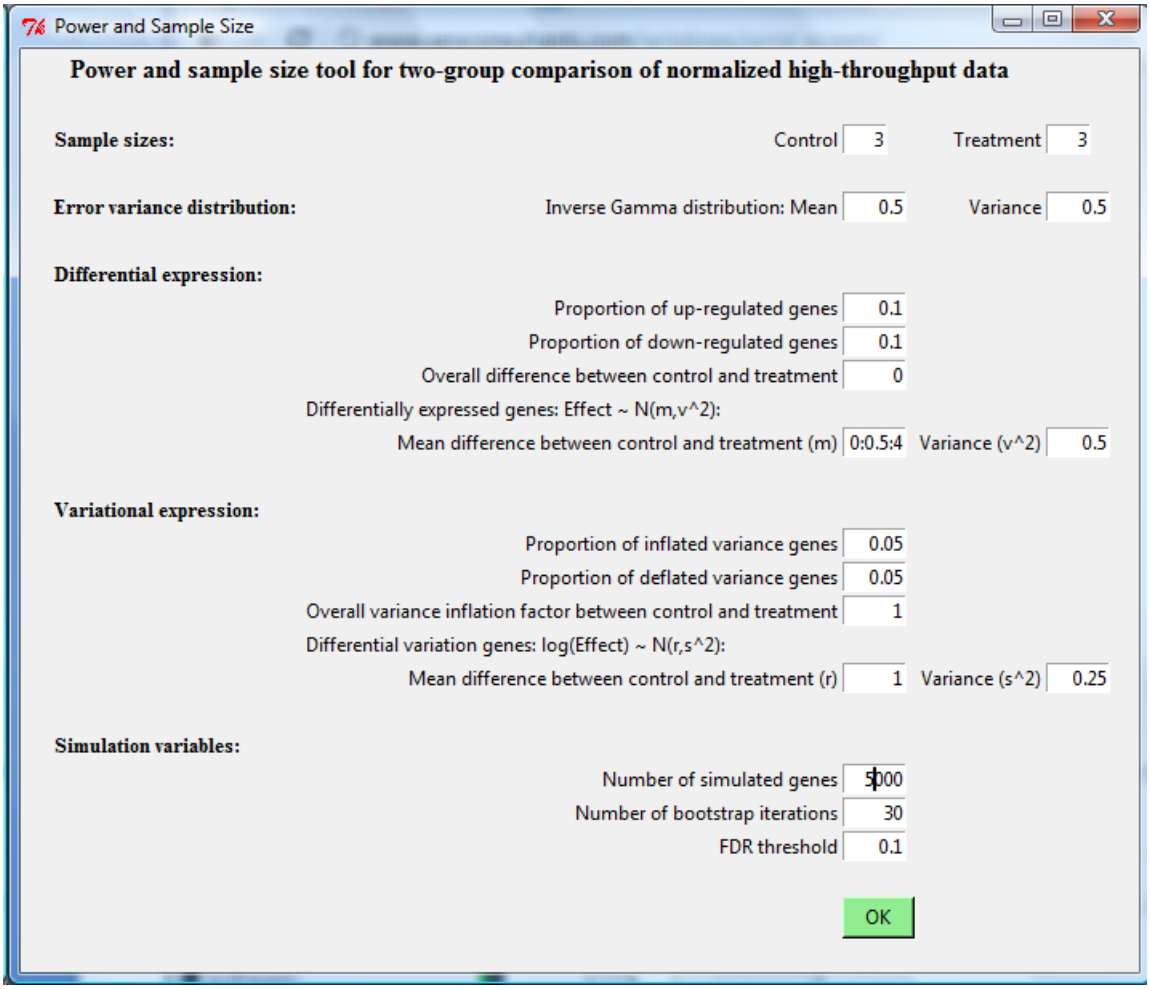

Figure 3.3: The graphic user interface of the power/sample size calculation program

tively. The nonnull probabilities are $p_1 = p_2 = q_1 = q_2 = 0.1$, and note that this means that on average, 39.6% of the genes will be nonnull. The differentially expressed genes are distributed $N(\nu + \psi, 0.5)$ where $\nu = 0$ and $\psi = 1, 1.5, 2, 2.5, 3$. The overall variance inflation factor is $\tau = 1$ and $\lambda_g \sim N(1, 0.25)$. We simulated 20,000 genes, and set the false discovery rate threshold at 0.1. We ran the simulation with 30 bootstrap iterations to assess the variability in the power, accuracy, and false discovery rate. Here we show the plots for one of these bootstrap iterations.

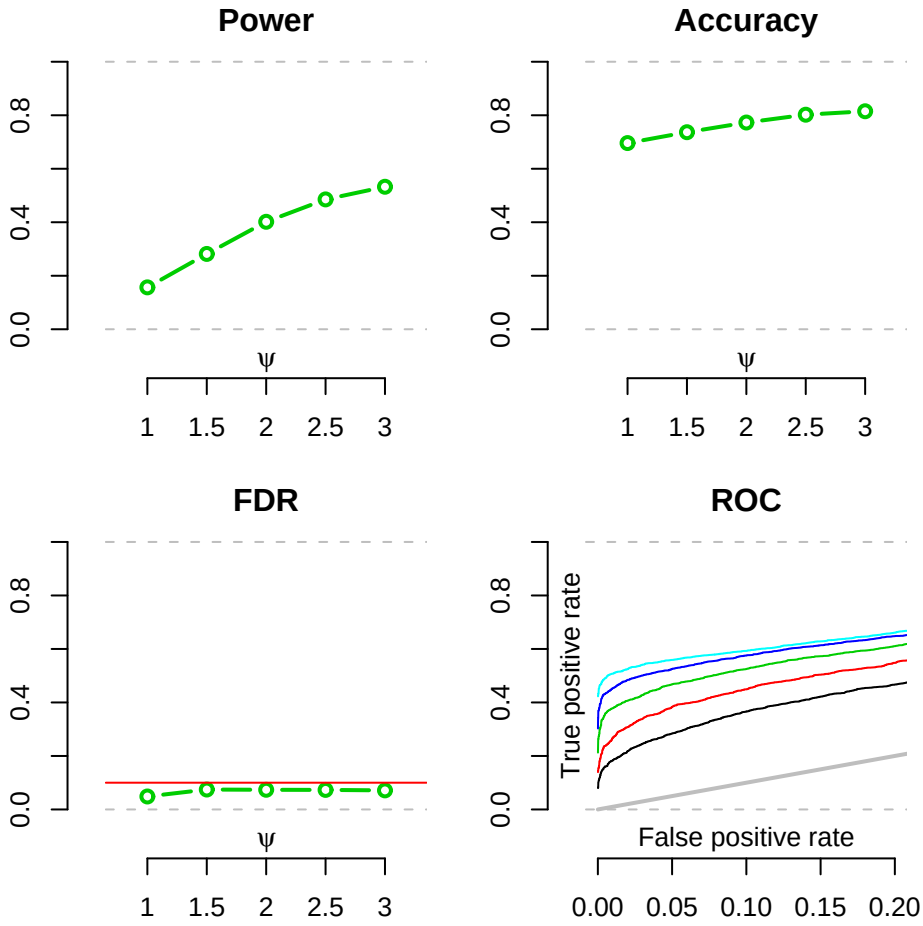

Figure 3.4: Output from the power/sample size calculation program

The output contains four plots. The 'Power' panel shows that for this parameter set-up, the power is less than 0.2 for detecting log fold change of 1, and it increases to almost 0.6 for $\psi = 3$. The 'Accuracy' panel shows the fraction of correct classifications. Recall that approximately 60% of the genes in this set-up are null, so a very conservative method is expected to have an accuracy of about 0.6. As ψ increases in the given range, so does the accuracy, to levels of approximately 0.8 for large ψ.

The 'FDR' panel shows that the actual false discovery rate is below the selected threshold. Finally, the ROC plot shows the true positive rate vs. the false

positive rate for each ψ, from 1 (bottom, black) to 3 (top, light blue). The straight gray line shows the random classification line. In this scenario, all the plots are above the gray line, and the area under the curve increases with ψ.

Yet another measure of the quality of the classification of our bivariate method, is the so-called MCC, or Matthews' Correlation Coefficient [2], also known as the ϕ coefficient. Unlike the FDR, power and accuracy measures, it takes into account all the values in the 'confusion matrix' (namely, TP, FP, FN, and TN, are all included in the computation of MCC). The MCC is given by

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}},$$

and if the denominator is 0 we set MCC=0.

MCC is often preferred to other measures because it can be used even if the proportion of null and non-null genes are very different. For example, when most of the genes are null, a very conservative method will have high accuracy (because if it does not reject anything, it correctly classifies the majority of the genes). However, this does not convey any information about the power of the method.

The MCC is a number between -1 and 1. It is, as the name suggests, a correlation coefficient between the true and predicted binary classifications. The fact that the MCC is on a scale that does not depend on the total number of observations in each class (null or non-null) allows for an intuitive interpretation. If MCC=1, then the procedure provides perfect classification. If MCC=-1, then the procedure gives the inverse prediction (i.e., it is wrong 100% of the time). If MCC=0, then the classification procedure is as good random classification.

Figure 3.5 depicts the MCC as a function of ψ for the same configuration

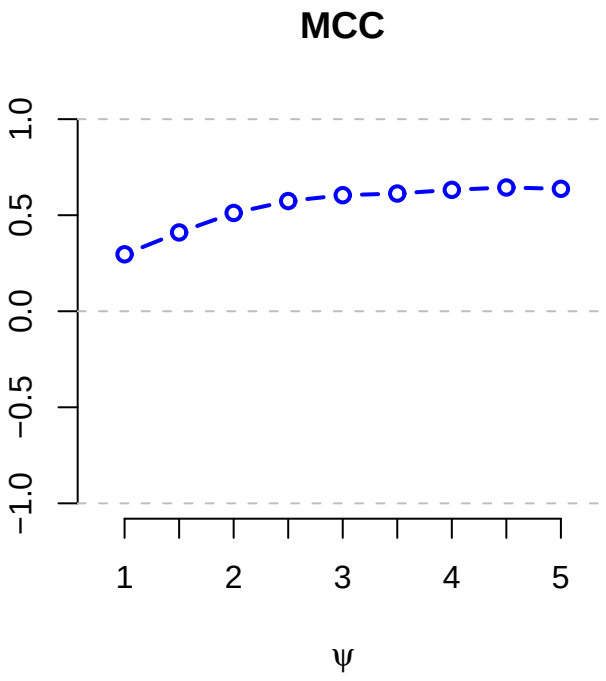

Figure 3.5: Matthews' Correlation Coefficient as a function of ψ (using the same configuration as in Figure 3.4)

used in Figure 3.4. It shows that even when the mean differential expression is small, the bivariate model achieves a fairly high MCC, and it increases with ψ.

Obviously, power and sample size calculations depend on the user's assessment of certain parameters, and these are sometimes available from previous experiments, but other times, they are not. In any case, it is recommended to use this interface to perform sensitivity analysis and see how different (reasonable) configurations affect the power, accuracy, false discovery rate, or the area under the ROC curves. For example, if we use the same configuration, but increase the estimated error variance to have mean 1 and variance 0.5, the estimated power is smaller for all ψ in the range. By varying the parameters, one can choose a sample size which is much more likely to yield the desired power, than when

relying on 'one gene at a time' power estimation methods.

3.6 Simulation Study

We performed an extensive simulation study to evaluate the performance of the bivariate method. We considered several performance metrics, including: (i) accuracy of the parameter estimates under the assumed bivariate mixture model; (ii) power to detect nonnull genes even when the data generation process differs from the assumed model; (iii) true positive rate vs. false positive rate; (iv) total number of correct classifications.

We use the simulation procedures in [3], where the total number of genes is 2,000, and for each configuration we created 30 data sets. The configurations differ in the proportion of null genes, the error variance variability, the differential expression and differential variation parameters, and the sample sizes. Since previous work demonstrated that the most powerful methods to detect differential expression to date are ones that model the differential expression as a random effect, and thus induce shrinkage [3, 49, 69], the focus of this section is on the comparison between one such univariate method (LEMMA), and the bivariate approach. In particular, we show that if there is no differential variation in the data our method performs as well as the most powerful univariate methods. However, if there is differential variation, the bivariate model is more powerful.

In the simulation described in the remainder of this section we have $n_1 = n_2 = 6$ and for the inverse gamma parameters we set $\alpha = 2.1$ and $\beta = 10/33$. This configuration corresponds to $E(\sigma^2_{1g}) = 1$ and high error variance variability.

The differential expression of nonnull genes is distributed $N(3, 1)$. Of the 2,000 genes, 200 were set as differentially expressed. In this configuration, we have no differential variation (that is, $\tau = 1$ and $\lambda_g = 1$). We then introduced additional variability to a subset of 200 genes, of which 100 are differentially expressed. For this subset of genes, the variance in the treatment group was, on average, four times greater than in the control group. That is, in the second configuration we have $\tau = 1$ and $\lambda_g = 4$.

In our analysis, genes are declared as nonnull if their posterior null probability for the mean effect is less than 0.1. The comparison between the power obtained by the univariate and bivariate methods, for the two configurations ($\lambda_g = 1$ and $\lambda_g = 4$), is depicted in Figure 3.6. The plot clearly shows that when there is no differential variation, the bivariate method has the same performance as the univariate method, but when differential variation is present, the bivariate method increases the power, while the univariate method decreases it.

We obtain very similar results with other performance measures, such as the accuracy (total percentage of correct classifications), false discovery rate, and the area under the ROC curve. The simulation study also shows that the estimation procedure obtains accurate parameter estimates when the data are generated under the bivariate model. We also see that inference based on the bivariate model is robust to the normality assumptions. Since these results are consistent with the observations in [3] and 2, which are special cases of the bivariate model in this paper, we do not go into further details here.

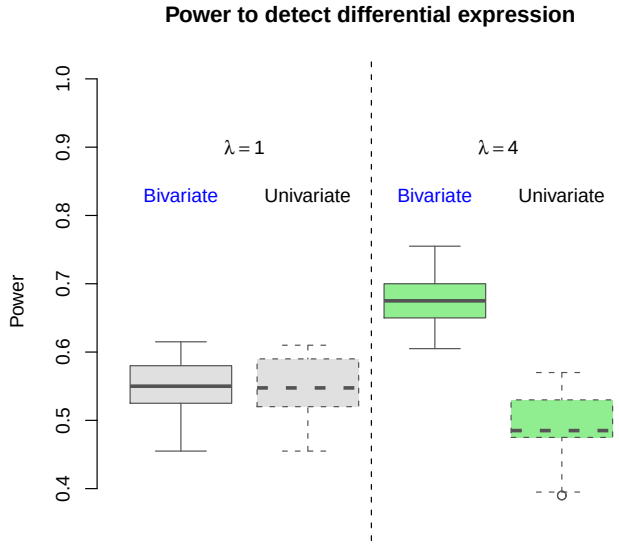

Figure 3.6: The power to detect nonnull genes, using the univariate (LEMMA, [3]) and bivariate methods. In this example, $n_1 = n_2 = 6$ and the error variances are distributed $IG(\alpha = 2.1, \beta = 10/33)$. The differential expression of nonnull genes is distributed $N(3, 1)$, and the proportion of nonnull genes is 0.1. The two boxplots on the left show the power when there is no differential variation. The two boxplots on the right show the power when the variability of the response is 4 times greater in the treatment group for a set of 200 genes of which 100 are differentially expressed.

3.7 Case Studies

We applied the bivariate model to several data sets, using different types of data (including gene expression, methylation, fMRI, and metabolomics data). The bivariate model seems to fit these data types very well. In this section we report the results for two data sets. The first involves metabolomics data from [66]. The second example involves gene expression data from the 'Ross' data set in [67], which is provided with the `MADE4` R-package [23].

3.7.1 Metabolomics Data

In recent years the study of chemical processes involving metabolites has become a popular complement to gene-expression analysis. Metabolites are the product of cellular processes, and they include several pathways, like amino acid (e.g., creatine, glutamine), lipid (e.g., choline, 2-hydroxyglutarate), and energy (citrate, pyrophosphate), just to name a few. Cancer researchers noted that there are differences in cellular metabolism between normal and cancer cells [28]. Here, we use a data set from [66] where over 200 metabolites in human oligodendroglioma (HOG) cells were profiled to determine the effects of expression of IDH1 and IDH2 mutants on cellular metabolism using a glioma cell line.

The data used here are from 'Dataset S2' in [66], which consists of 114 metabolites, and measured in four groups, called 'vector control', 'IDH1-WT', 'IDH1-R132H' and 'Fresh media'. Each group had 3 replicates. This data set has been re-scaled to have median equal to 1, and missing values were imputed with the minimum. In our analysis we compared the R132H group with the control. We removed one metabolite (2-hydroxyglutarate, 2HG) which had much higher levels in the 'IDH1-R132H' group than in the control (the values in the 'IDH1-R132H' group are 136.71, 118.87, 120.40, much higher than the supposed median of 1). When included in the analysis, it is obviously detected as significantly different between the two groups. We excluded four additional metabolites from the analysis, since they had zero within-group variability in one of the groups.

The bivariate analysis is depicted in Figure 3.7. Ten metabolites are detected at the 5% FDR threshold (shown as red squares in the plot). These ten metabolites with their corresponding statistics (d_g, x_g, and the bivariate Chi-

square statistic X^2) appear in Table 3.1. The five metabolites in bold, together with 2HG were detected by the analysis in [66], and they write that "these six metabolites are a subset of those that were altered in lysates of cells expressing either IDH1-R132H or IDH2-R172K."

Our method detects five additional metabolites. It is interesting that [66] does not mention glycylleucine, since, using our method, it has the second largest Chi-square statistic. We also note that dihomo-linolenate clearly exhibits significant differential variation ($x_g = 7.49$). This could be an artefact of the small sample size, but recall that the bivariate model accounts for that through the estimation of θ_g and κ_g^2.

N-acetylmethionine has a relatively small mean-difference (-0.59) compared

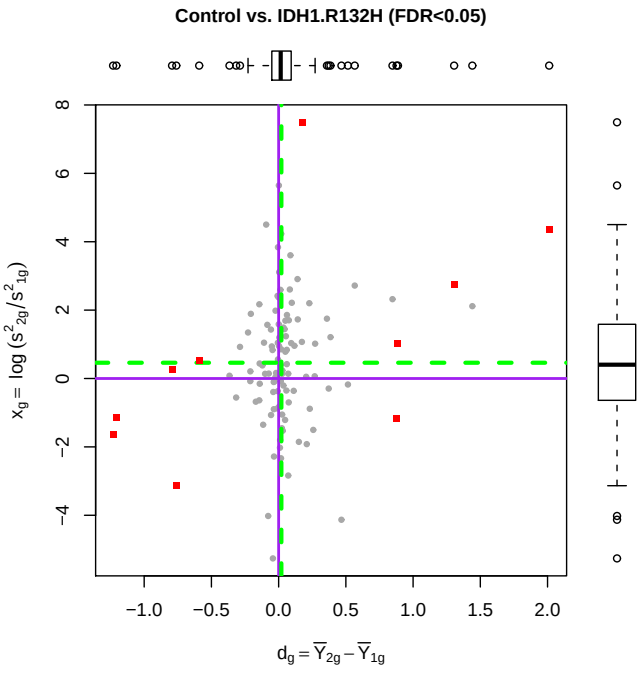

Figure 3.7: The bivariate distribution of $z_g = (d_g, x_g)'$ for the metabolite data in [66], comparing between the control group, and mutation R132H. Number of metabolites: $G = 109$. Sample sizes: $n_1 = n_2 = 3$.

to the other on the list, but its bivariate statistic (18.56) is highly significant. Furthermore, its adjusted p-value when testing for differential expression (in terms of the statistic d_g) is 0.0002. This is a clear case in which the resulting shrinkage estimation which takes into account the differential variation, increases the power to detect significant mean-differences that are otherwise very hard to detect. Similarly, even pyrophosphate which has the highest (Benjamini-Hochberg) adjusted p-value in this set according to the X^2 statistic ($p = 0.008$), is significant according to its d_g statistic, with $p = 0.02$ after adjustment for multiple testing.

3.7.2 Gene Expression Data

The data, originally described in [67] and later in [22], contains gene expression profiles from the NCI60 microarray expression project. In this project cDNA microarrays were used to assess gene expression profiles in 60 human cancer

Metabolite	Pathway	d_g	x_g	X^2
3-methyl-2-oxobutyrate	Amino acid	**-0.76**	-3.14	31.99
3-methyl-2-oxovalerate	Amino acid	**-0.79**	0.26	36.06
4-methyl-2-oxopentanoate	Amino acid	**-1.23**	-1.63	80.81
dihomo-linolenate (20:3n3 or n6)	Lipid	0.18	7.49	16.45
glycerol 3-phosphate (G3P)	Lipid	1.31	2.76	20.52
glycerophosphorylcholine (GPC)	Lipid	**0.88**	-1.16	26.80
glycylleucine	Peptide	-1.21	-1.14	70.15
kynurenine	Amino acid	**0.89**	1.03	33.94
N-acetylmethionine	Amino acid	-0.59	0.52	18.56
pyrophosphate (PPi)	Energy	2.01	4.36	14.34

Table 3.1: The metabolites that were detected as significantly different between the control and the R132H groups, using the bivariate test, at the 5% FDR threshold.

cell lines as part of the National Cancer Institute's drug discovery program. The data were normalized and filtered, and the final data set in [23] contains 1,375 genes. The 60 cancer cell lines in this data set are: BREAST (8), CNS (6), COLON (7), LEUK (6), MELAN (8), NSCLC (9), OVAR (6), PROSTATE (2), and RENAL (8).

Here, we show the comparison between the breast, and ovarian cancer cell lines. In this data set, the overall inflation factor is $\tau = 0.53$, which means that the variance in the breast cancer cell lines is, on average, twice the variance in the ovarian cancer cell line. Hence, we fit the data using two models: (a) the bivariate model; and (b) a univariate random effects model that assumes variance homogeneity across treatment groups. The latter is a special case of the bivariate model, where $\tau = 1$, and $\lambda_g = 1$ for all genes. This reduces to the LEMMA model in [3]. For both models, we use the 5% FDR threshold.

Figure 3.8 plots x_g versus d_g, along with their univariate boxplots. The dashed green lines represent the means of d_g and x_g for the null distribution. The blue diamonds show the 14 genes that are detected by either model (a or b). The red squares represent the 29 genes that are only detected by the bivariate model (a). In total, the bivariate model detects 43 genes, more than three times the number of discoveries made under the assumption of variance homogeneity across treatment groups.

The plot also shows that the bivariate model is very realistic. For example, the boxplot on the right-hand side suggests that the three-group normal mixture model fits the distribution of x_g very well. More generally, the bivariate model predicts that the null group has an elliptic shape, per equation (3.6). This seems to be the case for all the (normalized) data sets we investigated, including not

only gene expression data, but also gene methylation, fMRI, and metabolomics. Finally, the plot supports the model assumption of non-zero mean under the null $(\nu, \theta_g + \log \tau)'$.

3.8 Conclusions

In this chapter we introduced a novel approach for detecting treatment effects in high-throughput data. The bivariate model extends powerful methods for detecting differential expression by considering the effect of the treatment on

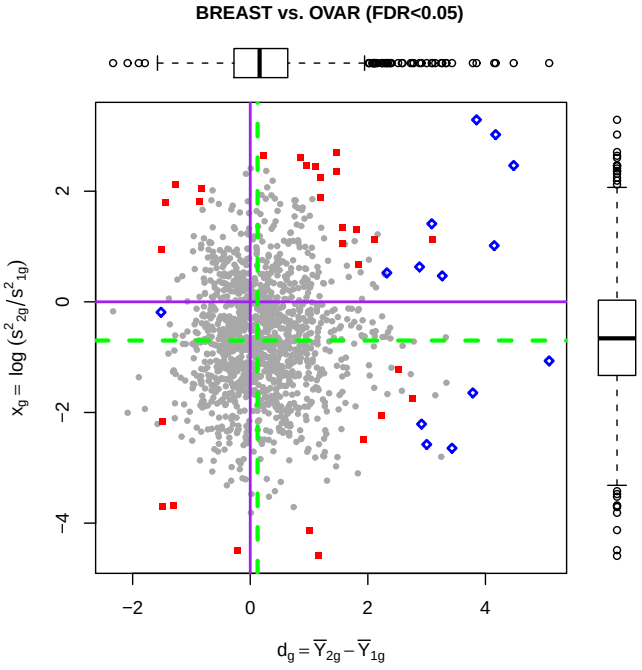

Figure 3.8: The bivariate distribution of $z_g = (d_g, x_g)'$ for the comparison between the breast, and ovarian cancer cell lines. The blue diamonds show the 14 genes that are detected by both the bivariate model, and LEMMA (which assumes variance homogeneity across treatment groups). The red squares represent the 29 genes that were only discovered by the bivariate model, at the 0.05 FDR level.

the mean and the variance. This not only enables the detection of differential variation, but it also increases the power to detect differential expression. We show in simulations that the new method yields a substantial gain in power when differential variation is present. Our case studies show that the model is realistic in a wide range of data sets.

Modeling the differential mean and variance as random effects results in shrinkage estimation, which is known to increase the power, since the estimators borrow strength across genes. Furthermore, through a three-step estimation approach, in which we apply the Laplace approximation, and by using the EM algorithm, we get a computationally efficient method, which is particularly well-suited for 'large p, small n' situations.

Finally, we intend to extend the bivariate approach to a broader set of models. In particular, we will develop similar models and estimation procedures for count data. This involves an extension of the bivariate model to the Generalized Linear Models (GLM) or to the quasi-likelihood frameworks.

CHAPTER 4

AN EMPIRICAL BAYES APPROACH TO VARIABLE SELECTION AND QTL ANALYSIS

4.1 Introduction and Motivation

This chapter focuses on variable selection in normal linear regression models when there are a large number of candidate explanatory variables, most of which have little or no effect on the dependent variable. We propose an empirical Bayes, model-based approach to variable selection which we implement via a fast EM algorithm.

Traditional regression problems typically involve a small number of explanatory variables and an analyst can make educated decisions as to which ones should be included in the regression model, and which should not. However, the new age of high speed computing and recent analytic needs and technological advances in genetics, for example, have dramatically changed this paradigm. It is common practice to use linear regression models to estimate the effects of hundreds or even thousands of predictors on a given response. These modern applications present major challenges. First, there is the so-called 'large p, small n' problem, since the number of predictors, e.g. genetic markers in a Quantitative Trait Loci (QTL) study, often greatly exceeds the sample size. Methods controlling the experiment-wise false discovery rate in one predictor at a time analyses often result in few or no discoveries. Second, the model space is huge. For example, for a modest QTL study with 1000 markers, there are 2^{1000} possible models. This renders traditional search-based algorithms impractical.

Automated methods for variable selection in normal linear regression models have long been studied in the literature [47]. Recent work on this topic includes [13, 19, 41]. Virtually every statistical package contains an implementation of standard stepwise methods. These methods typically add or remove one variable from the model in each iteration, based on sequential F-tests, or based on the change in other goodness-of-fit type scores, including adjusted R-square, Akaike information criterion (AIC) [1], Bayesian information criterion (BIC) [68], or Mallows' Cp. Other approaches use the false discovery rate (FDR) procedure [8].

AIC and BIC belong to a family of criteria that take into account two components: the likelihood of the model, and a term that penalizes complex models. The Cp statistic is similar in nature, in that it involves a penalty term ($2p-n$), but it depends on the residuals sum of squares. It is often used as a stopping rule for stepwise variable selection procedures. Other approaches include variations of the LASSO [73], which minimize the residuals sum of squares, subject to an L1 constraint, namely that the sum of the absolute values of parameter estimates is bound. In other words, the constraint ensures that the number of non-zero parameter estimates is controlled.

Our method is more related to Bayesian approaches, which include [19] and [41]. However, we use an empirical Bayes approach and the EM algorithm, rather than Gibbs sampling. Our model-based approach allows for a fully-Bayesian implementation, but the EM algorithm, combined with a computational trick yield much improved performance. This is particularly important in modern applications in which the running time of an MCMC sampler is too long for many data sets.

This chapter is organized as follows. We introduce the model-based approach in Section 4.2. In Section 4.3 we derive the complete data likelihood function, and the EM algorithm procedure. In Section 4.4 we illustrate our method using well-known data sets and compare our results with others in the literature. In Sections 4.5 we discuss extensions and future plans, and we conclude with Section 4.6

4.2 A Statistical Model for Automatic Variable Selection

Denote the (continuous) responses by $y_i, i = 1, \ldots, N$. Suppose that for each response we have J measurements , $x_{ij}, j = 1, \ldots, J$, of covariates of interest (e.g. sex, population, age) which we want to include in the regression model. We denote the mean effect of the j-th covariate by β_j.

Suppose that there are K putative variables $z_{ik}, k = 1, \ldots, K$, of which only a small subset should be included in the model. Let z_{ik} be the value of the k-th putative covariate of the i-th subject. Here we assume that K is large, and we have no information on which of these covariates should be included in the regression model.

We assume that the response y_i can be modeled using an additive combination of the covariates:

$$y_i = \sum_{j=1}^{J} x_{ij}\beta_j + \sum_{k=1}^{K} z_{ik}\gamma_k u_k + \varepsilon_i \tag{4.1}$$

where

$$\begin{aligned} u_k &\overset{iid}{\sim} N\left(\mu, \sigma^2\right) \\ \gamma_k &\overset{iid}{\sim} multinomial\,(0, 1, -1; p_0, p_1, p_2) \end{aligned}$$

$$\varepsilon_i \overset{iid}{\sim} N\left(0, \sigma_e^2\right).$$

The multinomial random variables γ_k take the value 1 or -1 if and only if the k^{th} putative variable z_{ik} is included in the model. Its sign indicates whether the mean effect of the k^{th} variable on the response is positive or negative.

In this context, the problem of variable selection can therefore be seen as an estimation procedure, where our main interest is in the latent variables $\{\gamma_k\}$.

It is convenient to express the model in matrix notation. Denote the $N \times K$ matrix (z_{ik}) by $\mathbf{Z}$, and write $\mathbf{\Gamma} \equiv diag(\gamma_1, \gamma_2, \ldots, \gamma_K)$, and $\boldsymbol{\mu} = \mu \mathbf{1}_K$. Let $\mathbf{Z}_k$ denote the k^{th} column of $\mathbf{Z}$. Also, denote the $N \times J$ matrix $\mathbf{X} = (x_{ij})$, and the $J-$vector of fixed effects $\boldsymbol{\beta} = (\beta_1, ..., \beta_J)'$. Then

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{\Gamma}\mathbf{u} + \boldsymbol{\varepsilon} \tag{4.2}$$

$$\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}_N, \sigma_e^2 \mathbf{I}_N\right) \tag{4.3}$$

$$\mathbf{Z}\mathbf{\Gamma}\mathbf{u} \sim N\left(\mathbf{Z}\mathbf{\Gamma}\boldsymbol{\mu}, \sigma^2 \mathbf{Z}\mathbf{\Gamma}^2\mathbf{Z}'\right). \tag{4.4}$$

This is similar to the usual mixed-model representation, but with two notable differences. First, our model includes the diagonal matrix, $\mathbf{\Gamma}$, which is used to select the columns from $\mathbf{Z}$. Second, the mean of the random effect terms is not zero. Note that in the usual mixed model context, the mean of the random effect is not identifiable separately from the overall mean, and therefore it is assumed to be 0. However, in mixture models (e.g. [3]) this is not the case, and in fact, not only are the two means identifiable, letting μ be non-zero increases the power to detect significant variables. Furthermore, the non-zero mean allows us to separate the significant covariates into two groups (positive and negative mean effect), and increase the power even more (compared with the two-group mixture model).

We assumed here that the variables $\{\gamma_k\}$ are independent and identically distributed. Clearly, this may be unrealistic and we need to consider possible correlation between variables. We return to this point in Section 4.3.4, and show how we mitigate the problem of multicollinearity in the selected model. We also discuss how one can incorporate additional information about specific putative variables.

Including interaction terms in this framework is straightforward. To add an interaction between $\mathbf{u}_k$ and $\mathbf{u}_m$ we simply augment $\mathbf{Z}$ by adding a column which contains the element-wise product of the k^{th} and m^{th} columns.

Finally, categorical variables are represented by $s-1$ binary columns in the matrix $\mathbf{Z}$, where s is the number of possible levels for that putative variable.

4.3 Estimation

4.3.1 The Complete Data Likelihood

We employ an empirical Bayes approach in which the parameters $\theta = \{\boldsymbol{\beta}, \mu, \sigma_e^2, \sigma^2\}$ are estimated via a modified EM algorithm, and upon convergence we select a column $\mathbf{Z}_k$ to be included in the model if the estimated posterior probability of its latent indicator, γ_k, is greater than a predefined threshold. The complete data likelihood, $f_C(\mathbf{y}, \boldsymbol{\Gamma})$, is obtained by integrating out the random effects, $\{u_k\}$. Then the Q-function for the EM algorithm is given by $Q(\theta|\theta^{(t)}) = E_{\theta^{(t)}}\{\log f_C(\mathbf{y}, \boldsymbol{\Gamma})|\mathbf{y}\}$.

We treat γ_k as missing values and denote $\gamma_{[0]} = \sum_{k=1}^{K} I[\gamma_k = 0]$, $\gamma_{[1]} =$

$\sum_{k=1}^{K} I[\gamma_k = 1]$, $\gamma_{[2]} = K - \gamma_{[0]} - \gamma_{[1]}$, where $I[\cdot]$ is the indicator function. Also denote $\mathbf{V} = \mathbf{Z}\boldsymbol{\Gamma}$. We can write the complete data likelihood conditional on σ_e^2 and integrate the random variable $\mathbf{Vu}$:

$$\left[\mathbf{y}|\sigma_e^2, \sigma, \boldsymbol{\Gamma}\right] = p_0^{\gamma_{[0]}} p_1^{\gamma_{[1]}} p_2^{\gamma_{[2]}} \int \left[\mathbf{y}|\mathbf{Vu}, \sigma_e^2\right] \left[\mathbf{Vu}|\sigma^2\right] d\mathbf{Vu}.$$

This leads to the following log-likelihood function:

$$\begin{aligned} \ell &= \gamma_{[0]} \log(p_0) + \gamma_{[1]} \log(p_1) + \gamma_{[2]} \log(p_2) - \frac{N}{2} \log(2\pi) \\ &\quad -\frac{1}{2} \log\left|\sigma_e^2 \mathbf{I}_N + \mathbf{Z}\sigma^2 \boldsymbol{\Gamma}^2 \mathbf{Z}'\right| \\ &\quad -\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{V}\boldsymbol{\mu})' (\sigma_e^2 \mathbf{I}_N + \mathbf{Z}\sigma^2 \Gamma^2 \mathbf{Z}')^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{V}\boldsymbol{\mu}) \ . \end{aligned} \tag{4.5}$$

Note that the likelihood function is simply the probability distribution function of a multivariate normal random variable with mean $\mathbf{X}\boldsymbol{\beta} + \mathbf{V}\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}^{\dagger} = \sigma_e^2 \mathbf{I}_N + \mathbf{Z}\sigma^2 \Gamma^2 \mathbf{Z}'$, multiplied by the prior probability of the latent variables.

4.3.2 The EM Algorithm

To derive the equations for the maximum likelihood estimates, we start with the mean parameters, $\mu, \boldsymbol{\beta}$. Denote $\mathbf{W} = [\mathbf{X}, \mathbf{Z}\boldsymbol{\Gamma}\mathbf{1}_K]$. The mean of the multivariate normal distribution in the complete data likelihood is $\mathbf{W}\tilde{\boldsymbol{\beta}}$ where $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}', \mu)'$. Then the MLE for $\tilde{\boldsymbol{\beta}}$ is given by

$$(\hat{\boldsymbol{\beta}}', \hat{\mu})' = (\mathbf{W}'(\boldsymbol{\Sigma}^{\dagger})^{-1}\mathbf{W})^{-1}\mathbf{W}'(\boldsymbol{\Sigma}^{\dagger})^{-1}\mathbf{y} \tag{4.6}$$

To estimate the variance parameters, we use the following equations (see Section 8.3.b in [60]). Using the values from the t^{th} iteration of the EM algorithm,

define

$$\tau_e = trace(\sigma_e^2 \mathbf{I}_N - \sigma_e^4(\mathbf{\Sigma}^\dagger)^{-1}) + \sigma_e^4(\mathbf{y} - \mathbf{W}\tilde{\boldsymbol{\beta}})'(\mathbf{\Sigma}^\dagger)^{-2}(\mathbf{y} - \mathbf{W}\tilde{\boldsymbol{\beta}}), \tag{4.7}$$

and the $t+1$ update to σ_e^2 is

$$\sigma_e^2 = \frac{\tau_e}{N}. \tag{4.8}$$

Similarly, let

$$\tau_r = trace(\sigma^2 \mathbf{I}_K - \sigma^4 \mathbf{V}'(\mathbf{\Sigma}^\dagger)^{-1}\mathbf{V}) + \sigma^4(\mathbf{y} - \mathbf{W}\tilde{\boldsymbol{\beta}})'(\mathbf{\Sigma}^\dagger)^{-1}\mathbf{V}\mathbf{V}'(\mathbf{\Sigma}^\dagger)^{-1}(\mathbf{y} - \mathbf{W}\tilde{\boldsymbol{\beta}}) \tag{4.9}$$

and the $t+1$ update to σ^2 is

$$\sigma^2 = \frac{\tau_r}{rank(\mathbf{V})}. \tag{4.10}$$

For p_0, p_1, p_2 we use Lagrange multipliers: $p_0 + p_1 + p_2 = 1$ and $\gamma_{[0]} + \gamma_{[1]} + \gamma_{[2]} = K$, and obtain

$$p_i = \frac{\gamma_{[i]}}{K} \tag{4.11}$$

so $\hat{p}_1, \hat{p}_2$ estimate the proportion of putative variables included in the model.

For the estimation of the latent variables γ_k we use Bayes rule to compute the posterior probability that putative variable k is included in the model:

$$Pr(\gamma_k = 0) = \frac{p_0^{(t)} f\left(\mathbf{y}; \gamma_k = 0, \gamma_{-k} = \gamma_{-k}^{(t)}\right)}{\sum_{s=-1,0,1} p_{i(s)}^{(t)} f\left(\mathbf{y}; \gamma_k = s, \gamma_{-k} = \gamma_{-k}^{(t)}\right)} \tag{4.12}$$

$$Pr(\gamma_k = 1) = \frac{p_1^{(t)} f\left(\mathbf{y}; \gamma_k = 1, \gamma_{-k} = \gamma_{-k}^{(t)}\right)}{\sum_{s=-1,0,1} p_{i(s)}^{(t)} f\left(\mathbf{y}; \gamma_k = s, \gamma_{-k} = \gamma_{-k}^{(t)}\right)} \tag{4.13}$$

$$Pr(\gamma_k = -1) = \frac{p_2^{(t)} f\left(\mathbf{y}; \gamma_k = -1, \gamma_{-k} = \gamma_{-k}^{(t)}\right)}{\sum_{s=-1,0,1} p_{i(s)}^{(t)} f\left(\mathbf{y}; \gamma_k = s, \gamma_{-k} = \gamma_{-k}^{(t)}\right)} \tag{4.14}$$

where $f(\cdot)$ is the likelihood in (4.5) given the current parameter estimates, and $i(s) = 0, 1, 2$ for $s = 0, 1, -1$, respectively. The notation $\gamma_{-k} = \gamma_{-k}^{(t)}$ means that

to update the k^{th} variable in the diagonal matrix $\mathbf{\Gamma}$ we hold all the other ones constant, at their value from the previous iteration.

We set $\gamma_k = 0$ if $Pr(\gamma_k = 0)$ is greater than a certain threshold. Otherwise, if $Pr(\gamma_k = 1) > Pr(\gamma_k = -1)$ we set $\gamma_k = Pr(\gamma_k = 1)$ and if $Pr(\gamma_k = 1) \leq Pr(\gamma_k = -1)$ we set $\gamma_k = -Pr(\gamma_k = -1)$.

In other words, we include the k-th covariate if and only if $Pr(\gamma_k = 0)$ is less than a certain threshold. This will have significant computational benefits when N and K are large, but only a small number of covariates have a significant effect on the response. We elaborate on this in the next subsection. We refer to the variables that are excluded from the model as '*null*'.

4.3.3 When N is Large – the Modified EM Algorithm

Application of the EM algorithm is not entirely straightforward, for two reasons. First, the log complete data likelihood is a non-linear function of the latent variables, making the E-step analytically intractable. We solve this problem by updating the γ_k's by their posterior expectations using Bayes rule, as we showed in the previous subsection.

A second problem stems from the modeling of the putative variables as random effects. When we integrate out the random effect, the variance-covariance matrix of the posterior likelihood contains a large $(N \times N)$ matrix of the form $\mathbf{I}_N + \frac{\sigma^2}{\sigma_e^2}\mathbf{Z}\mathbf{Z}'$, which has to be inverted to compute the iterative maximum likelihood estimates. To address this computational problem we use the Woodbury identity [42], and express $f_C(\mathbf{y}, \mathbf{\Gamma})$ in terms of the $K \times K$ matrix $\Sigma_K^* = \mathbf{I}_K + \frac{\sigma^2}{\sigma_e^2}\mathbf{Z}'\mathbf{Z}$.

This simplifies the computation because the $(k,l)th$ element of $\mathbf{Z}'\mathbf{Z}$ is given by $\langle \mathbf{z}_k, \mathbf{z}_l \rangle \gamma_k \gamma_l$, where $\langle , \rangle$ denotes the inner product of two vectors. In contrast, the elements of $\mathbf{Z}\mathbf{Z}'$ involve all the $\gamma_k s$. We set $\gamma_k^{(t)} = 0$ if the posterior expectation of the k^{th} latent variable in the t^{th} iteration is below a given threshold. Since only a small number of the putative variables are truly associated with the response, the matrix Σ_K^* is relatively sparse and much easier to invert.

Thus, to deal with the inversion of the $N \times N$ matrices $\mathbf{\Sigma}^\dagger$ and $\mathbf{W}'(\mathbf{\Sigma}^\dagger)^{-1}\mathbf{W}$ when N is large we obtain the following form of $(\mathbf{\Sigma}^\dagger)^{-1}$:

$$(\mathbf{\Sigma}_K^\dagger)^{-1} \equiv (\mathbf{\Sigma}^\dagger)^{-1} = \frac{1}{\sigma_e^2}\mathbf{I}_N - \frac{\sigma^2}{\sigma_e^4}\mathbf{Z}\mathbf{\Gamma}\left(\mathbf{I}_K + \frac{\sigma^2}{\sigma_e^2}\mathbf{\Gamma}\mathbf{Z}'\mathbf{Z}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}\mathbf{Z}'. \tag{4.15}$$

For matrices A, B of dimensions $K \times N$, the following identity holds:

$$\left|I_K + AB^T\right| = \left|I_N + B^T A\right|,$$

so we can rewrite the log-likelihood function:

$$\begin{aligned} \ell \;=\; & \gamma_{[0]}\log(p_0) + \gamma_{[1]}\log(p_1) + \gamma_{[2]}\log(p_2) - \frac{N}{2}\log(2\pi) \\ & -\frac{N}{2}\log\left(\sigma_e^2\right) - \frac{1}{2}\log\left|\mathbf{I}_K + \frac{\sigma^2}{\sigma_e^2}\mathbf{\Gamma}\mathbf{Z}'\mathbf{Z}\mathbf{\Gamma}\right| \\ & -(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{V}\boldsymbol{\mu})'\,(\mathbf{\Sigma}_K^\dagger)^{-1}\,(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{V}\boldsymbol{\mu})\;. \end{aligned} \tag{4.16}$$

Suppose that there are L variables for which $\gamma_k \neq 0$, and let $\mathbf{\Gamma}_L$ be the corresponding $L \times L$ matrix ($\mathbf{\Gamma}_L$ is obtained by eliminating all the 0 columns and rows in $\mathbf{\Gamma}$). Let $\mathbf{Z}_L$ be the sub-matrix obtained by eliminating the $K - L$ columns that correspond to $\gamma_k = 0$. Then we can rewrite (4.15):

$$(\mathbf{\Sigma}_K^\dagger)^{-1} \equiv (\mathbf{\Sigma}^\dagger)^{-1} = \frac{1}{\sigma_e^2}\mathbf{I}_N - \frac{\sigma^2}{\sigma_e^4}\mathbf{Z}_L\mathbf{\Gamma}_L\left(\mathbf{I}_L + \frac{\sigma^2}{\sigma_e^2}\mathbf{\Gamma}_L\mathbf{Z}_L'\mathbf{Z}_L\mathbf{\Gamma}_L\right)^{-1}\mathbf{\Gamma}_L\mathbf{Z}_L'. \tag{4.17}$$

We now denote $\mathbf{V} = \mathbf{Z}_L\mathbf{\Gamma}_L$ and $\mathbf{W} = [\mathbf{X}, \mathbf{Z}_L\mathbf{\Gamma}_L\mathbf{1}_L]$. Updating equations

4.6,4.7,4.9, and 4.16 is computationally simpler, since it involves the inversion of $L \times L$ matrices, and we assume that L is smaller than N and K.

4.3.4 Additional Implementation Considerations

We address a number of important implementation considerations. First, we often expect to have groups of highly correlated variables. This is particularly true when dealing with QTL data, where it is known that loci that are physically close, tend to be correlated. In general, one finds a region of loci (rather than a unique locus) that have a significant effect on the quantitative trait. Failing to account for the correlation is likely to cause a multicollinearity problem, and in particular may inflate the standard errors of the parameters. This, in turn, may result in failure to detect important covariates.
To account for correlation we include a variable, u_k, in the model if and only if

1. its posterior null probability is less than a certain threshold, and
2. the model does not include any other variable, u_j, such that $cor(\mathbf{Z}_j, \mathbf{Z}_k)$ is greater than a given threshold.

We define a correlation-based distance measure between any two columns in the matrix $\mathbf{Z}$. Let r_{jk} be the correlation coefficient between the vectors $\mathbf{Z}_j$ and $\mathbf{Z}_k$, and let R^2_{jk} be the corresponding coefficient of determination. Then, let

$$d_{jk} = 1 - R^2_{jk} \tag{4.18}$$

and define

$$C_k = \prod_{j \neq k} (1 - Pr(\gamma_j \neq 0) + d_{jk} Pr(\gamma_j \neq 0)) . \tag{4.19}$$

The adjusted posterior probabilities are now defined as follows:

$$P_0(k) \equiv 1 - C_k + C_k Pr(\gamma_k = 0) \tag{4.20}$$

$$P_1(k) \equiv C_k Pr(\gamma_k = 1) \tag{4.21}$$

$$P_{-1}(k) \equiv C_k Pr(\gamma_k = -1)\,. \tag{4.22}$$

Recall that in order to reduce the dimensionality of the matrices, when $Pr(\gamma_k = 0)$ (as defined in 4.12) is large enough, we set $\gamma_k = 0$. Similarly, for the simplicity of the following arguments, if d_{jk} is less than a specified threshold we set it to 0, and otherwise we set it to 1. This means that if variables j and k are highly correlated we set the distance between them to 0, and otherwise we set it to 1. Of course, after we do this rounding, d_{jk} is not a proper distance metric, because $d_{jk} = 0$ does not imply $\mathbf{Z}_j = \mathbf{Z}_k$, but it is still a distance function.

Consider now the k^{th} variable and the term C_k. Any other variable, j, that is not in the model, has $Pr(\gamma_j \neq 0) = 0$ so these have a unity multiplicative contribution to C_k. The majority of variables are in this category. A non-null variable ($Pr(\gamma_k \neq 0) > 0$) contributes a factor of $1 - Pr(\gamma_j \neq 0) + d_{jk} Pr(\gamma_j \neq 0)$ to C_k. For most of the variables we have $d_{jk} = 1$, and hence they contribute a factor of 1. When $d_{jk} = 0$ we get a factor of $1 - Pr(\gamma_j \neq 0)$. In other words, if a variable that is not highly correlated with the k^{th} variable is already in the model, it does not affect the posterior probability that the k^{th} variable is in the model. But, if a variable that is highly correlated with the k^{th} variable in the model, it reduces the posterior probability that $\gamma_k \neq 0$ is in the model by a factor of $1 - Pr(\gamma_j \neq 0)$. In the extreme case where $Pr(\gamma_j \neq 0) = 1$ we get $P_0(k) = 1$ (that is, the k^{th} variable does not enter the model).

The rounding of d_{jk} is not essential since we assume that $Pr(\gamma_j \neq 0) = 0$ for most of the variables. For j such that $Pr(\gamma_j \neq 0) > 0$ we can make similar arguments about the multiplicative contribution of variable j to C_k. However, since the computation of d_{jk} requires $O(K^2)$ operations and K might be very large, it is sometimes practical to assume that most d_{jk} are 0, and compute it only for pairs for which we have reason to believe are highly correlated. For example, when we deal with genetic data, we will set $d_{jk} = 0$ if j and k correspond to genes that lie on different chromosomes.

Incorporating the term C_k into the posterior probability helps to control the correlation between variables that are included in the model. Hence, it may be seen as a form of clustering, where we pick a single representative from a cluster, which is determined by the threshold we put on d_{jk}. Instead of picking a representative, we can take an average of the cluster (representing its 'center'), or we could perform Principal Component Analysis (PCA) on the variables in the cluster and include the first component in the model.

To incorporate the distance measure into the linear model, we replace the matrix Γ with the $K \times K$ matrix $\Delta = (\delta_{jk})$ where along the diagonal we have $\delta_{jj} = \gamma_j$, but in the upper triangle, i.e. for $j < k$ we have $\delta_{jk} = -\gamma_j\gamma_k(1 - d_{jk})$. As in the matrix Γ, in the lower triangle ($j > k$) we have $\delta_{jk} = 0$. It is easy to check that no term, z_j, appears more than once in the linear model, and if a pair of variables have distance 0, at most one of them appears in the linear model. This prevents the algorithm from including highly-correlated terms in the model. Like Γ, the matrix Δ is also very sparse. Most elements in the matrix are 0 because most of the γ_k's are 0. Of the remaining off-diagonal elements, some are 0 because the distance between z_j and z_k is 1. While Γ is diagonal by definition, the matrix

Δ contains an upper triangular matrix, but with dimension much smaller than $K \times K$. Thus, the computational advantages that we described earlier, still hold, since inverting or multiplying Δ is straightforward, due to its simple, sparse structure.

When we have a choice which of two highly correlated variables to include in the model, we may choose, for example, the one with more observations, or the one with higher variance.

We can also consider other distance functions, d_{jk}. We used R^2, which means that positive and negative correlations are treated the same way. It is reasonable to require that a pair of negatively correlated variables will be considered farther than positively correlated ones. Hence, one may define

$$d_{jk} = \frac{1 - r_{jk}}{2}. \tag{4.23}$$

Thus, positively correlated variables have distance close to 0, while negatively correlated variables have distance close to 1. A similar alternative is the Bhattacharyya distance [9], defined as

$$d_{jk} \quad = \quad \cos B_{jk} = \sum_{i=1}^{N} \left(Z_{ij} Z_{ik} \right)^{1/2} . \tag{4.24}$$

Another possible modification of our model is to incorporate external information about the putative variables. Specifically, the update to the posterior probabilities, $Pr(\gamma_k = s)$, in the general case is done by considering global values of p_0, p_1, and p_2. However, when more information is available for certain variables we can account for that in the prior probability distribution. One way do it is to let p_0, p_1, and p_2 depend on covariates and estimate them with a multinomial logistic regression model. Alternatively, we can partition the covariates as we did in Section 3.3.2, and assign each subset a different multinomial prior.

Finally, when higher-order term are included in the model, we find that for the stability of the estimation procedure it is highly recommended to standardize all the variables. For example, if we let $m_k = \min_i\{Z_{i,k}\}$ and $M_k = \max_i\{Z_{i,k}\}$, then the linear transformation $\mathbf{Z}_k \mapsto (2\mathbf{Z}_k - m_k - M_k)/(M_k - m_k)$ guarantees that the values of the transformed variable are between -1 and 1. Thus, any higher order term involving $\mathbf{Z}_k$ is also between -1 and 1.

4.4 Case Studies

4.4.1 The Ozone Data

We applied our method to the well-known air-pollution data set which was first introduced in [14] to illustrate the ACE procedure. It consists of daily measurements of ozone concentration levels in the Los Angeles basin, collected over 330 days in 1976. There are eight meteorological explanatory variables, labeled $x_1, \ldots, x_8$ by [39] and subsequent authors (e.g., [44]). These variables are

x_1: (vh) Vandenburg height,

x_2: (wind) the wind speed (mph),

x_3: (hum) the humidity (%),

x_4: (temp) the temperature (Fahrenheit),

x_5: (ibh) the temperature inversion base height (feet),

x_6: (dpg) the pressure gradient (mm Hg),

x_7: (ibt) the inversion base temperature (Fahrenheit),

x_8: (vis) the visibility (miles).

We refer mostly to a more recent analysis in [51], subsection 2.4.4, which also uses x_9, the day of the year (doy). Selecting a first-order linear regression model can be done easily by checking all $2^9 = 512$ possible models, but this strategy is not feasible when we wish to include second, or third order terms (with 2^{54} and 2^{219} possible models, respectively.)

We consider models with first and second order terms, with a total of 54 candidate predictors. We compare our results with those in [51], where they compare their selected model (without specifying how it was obtained), with models from [13]. The comparison in [51] is done in terms of the Akaike Information Criterion (AIC), and we do the same here. Their model, labeled "2.8" is

$$Y_{LNP} = Intercept + wind + temp + ibt + vis + doy + vis^2 + doy^2 . \quad (4.25)$$

They fit the model with a GLM with log link, and a gamma error, and obtain AIC score of 1743.3.

Our algorithm converged in a few seconds, and selected the following model:

$$\begin{aligned} Y = {} & Intercept + temp + ibt + vis + doy + vh \cdot temp + vh \cdot dpg + wind^2 + \\ & humid^2 + humid \cdot temp + humid \cdot ibh + ibh^2 + dpg^2 + vis^2 + doy^2 . \end{aligned} \quad (4.26)$$

This model, which we call 'Selected', includes polynomials that are not well-formed in the sense that some terms do not appear in the model as main effects , but they do appear in second order terms. We obtain the 'Full' model by adding the missing first order terms (ibh, humid, vh, dpg, wind). Adding these terms

	Normal			Gamma		
	Full	Selected	"2.8"	Full	Selected	"2.8"
AIC	371.5	367.8	494.8	1630.9	1625.9	1743.4
MAE	0.312	0.313	0.391	0.768	0.772	0.791
R^2	0.83	0.83	0.75	-	-	-

Table 4.1: Goodness of fit of the three models (Full, Selected, and "2.8" from [51]) in terms of AIC and mean absolute error (MAE). For the normal fit, we also compare the adjusted R^2 values

may introduce multicollinearity (which our algorithm is designed to mitigate), but we see in Table 4.1 that in this case the overall fit of the models with or without the missing first order terms, is quite similar.

We fit the three models (Full, Selected, and "2.8" from [51]) with a GLM with log link and a gamma error. We also use a normal fit to the logarithm of the ozone level. Table 4.1 summarizes the goodness of fit of the three models in terms of AIC and mean absolute error (MAE), and for the normal fit, we also provide the adjusted R^2 values. The MAE for the GLM models is computed with the deviance residuals, that account for overdispersion. That is, we use $r_d/\sqrt{\phi}$ where r_d are the GLM model residuals, and ϕ is the estimated overdispersion. Using AIC as our model selection criterion, it is clear that our Selected model has the best fit, under both fitting procedures (`lm` and `glm`). Notice that when we fit a normal model to log(*ozone*) our models explain 83% of the variability, a significant improvement over model "2.8" which explains 75%.

Figure 4.1 shows the diagnostics plots of the residuals. The left panel plot the deviance residuals vs. scaled (log) fitted values. The right panel shows a QQ-plot of the deviance residuals. These diagnostics plots provide further evidence for the adequacy of the model selected by our method.

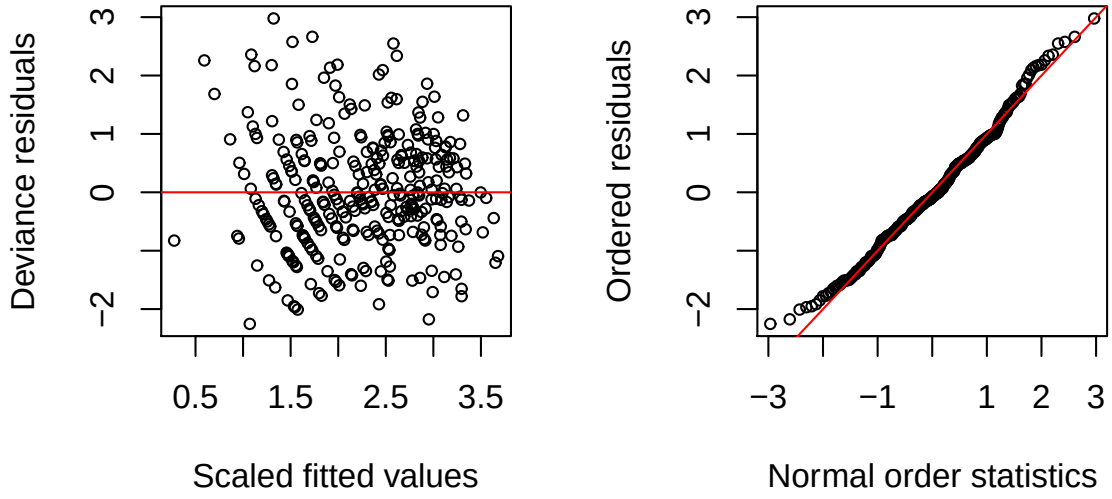

Figure 4.1: Ozone data – diagnostics plots of the 'Full' version of the model in 4.26 (including all main effects). Left: deviance residuals vs. fitted values. Right: quantile-quantile plot of the deviance residuals.

Table 4.2 shows the parameter estimates in our Selected model, using the GLM approach. Note that the estimates are quite different than those in [51] since we transformed the variables. However, the linear transformation does not affect the predicted values and the residuals.

It is interesting that doy^2 has the greatest absolute effect. The log(ozone) level is a second degree polynomial with respect to doy, which has the terms $-0.9012 \cdot doy^2 - 0.2131 \cdot doy$. The maximum of this polynomial with respect to doy corresponds approximately to June 21, the spring solstice. This makes sense, since it is well known that daylight affects ozone levels.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	2.6046	0.0555	46.93	0.0000
temp	0.3601	0.1087	3.31	0.0010
ibt	0.6197	0.0965	6.42	0.0000
vis	-0.1669	0.0431	-3.87	0.0001
doy	-0.2131	0.0306	-6.96	0.0000
$vh \cdot temp$	-0.3769	0.1004	-3.75	0.0002
$vh \cdot dpg$	-0.4489	0.1356	-3.31	0.0010
$wind^2$	0.2733	0.0843	3.24	0.0013
$humid^2$	-0.3179	0.0757	-4.20	0.0000
$humid \cdot temp$	0.4052	0.1096	3.70	0.0003
$humid \cdot ibh$	-0.2442	0.0464	-5.26	0.0000
ibh^2	-0.1842	0.0540	-3.41	0.0007
dpg^2	-0.8021	0.1016	-7.89	0.0000
vis^2	0.2534	0.0706	3.59	0.0004
doy^2	-0.9012	0.0885	-10.18	0.0000

Table 4.2: Parameter estimates for the ozone data with our 'Selected' model.

4.4.2 The Diabetes Data

The 'Least Angle Regression' (LARS) variable selection algorithm was introduced in [32]. As a motivating example, they use the diabetes data set, where the response is a quantitative measure of disease progression one year after baseline. There are ten explanatory variables, including age, sex, body mass index (bmi), average blood pressure (bp), and six blood serum measurements ($s_1, \ldots, s_6$). In total, there are n = 442 diabetes patients. All ten explanatory variables have been standardized to have mean 0 and unit length, and the response was centered around 0, by subtracting the sample mean.

Here, we are interested in comparing our results with their quadratic model which contained a total of 64 putative variables, with 10 main effects, 45 interactions, and 9 squares (not 10, since *sex* is a binary variable, so $sex = sex^2$). This

data set has been recently analyzed using the `care` R package in [75], where a new approach based on the so-called CAR scores to variable selection has been introduced. CAR scores are defined as "the marginal correlations adjusted for correlation among explanatory variables", and according to [75] "the CAR score provides a canonical ordering that encourages grouping of correlated predictors and down-weights antagonistic variables". In our analysis, we used the 'efron2004' data set from the `care` package which contains the original ten variables, and we constructed the second-order terms.

Our algorithm selected the following model:

$$y = Intercept + sex \cdot s_1 + bmi \cdot s_3 + bp \cdot s_6 + s_3 \cdot s_5 + s_5 \cdot s_6 \,.$$

The parameter estimates for this model are given in Table 4.3. The selected model has adjusted $R^2 = 0.51$. In contrast, [32] obtain 'true R^2' of 0.42, and fitting the model with LARS, they get a 'true predictive R^2' of about 0.40. Interestingly, in their simulations they observe that the "proportion explained" of the estimates, as a function of average number of terms reaches a level of over 95% with 6 terms, and a maximum of about 96% with additional variables. Our final model has 6 variables.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1801	0.0626	2.88	0.0042
$sex \cdot s_1$	-4.3729	1.1215	-3.90	0.0001
$bmi \cdot s_3$	-2.0439	0.2549	-8.02	0.0000
$bp \cdot s_6$	-0.7758	0.1500	-5.17	0.0000
$s_3 \cdot s_5$	-1.4497	0.3292	-4.40	0.0000
$s_5 \cdot s_6$	-1.1983	0.1659	-7.22	0.0000

Table 4.3: The selected model for the diabetes data.

We added the missing first order terms to construct the well-formed polyno-

mial model, but because several of the second-order terms are highly correlated with the original variables, some terms were no longer identifiable. (In particular, one reason high correlation exists, is because s_1, s_2, s_3 and s_5 correspond to total cholesterol (TC) , LDL, HDL, and triglycerids (TG) respectively, and LDL is obtained by a linear equation involving the other three, namely LDL=TC-HDL-TG/5).

We used standard variable selection techniques to remove multicollinearity problems from the complete model (with the well-formed polynomials), and obtained the following model:

$$y = Intercept + sex + bmi + bp + s_1 + s_3 + s_5 \,. \tag{4.27}$$

Table 4.4 contains the parameter estimates of our final model. The adjusted R^2 of this model is 0.51, which is quite a bit higher than that obtained by the simulations in [32].

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1927	0.0627	3.07	0.0023
sex	-0.1400	0.0374	-3.75	0.0002
bmi	0.9094	0.1107	8.22	0.0000
bp	0.5204	0.0997	5.22	0.0000
s_1	-0.2488	0.1229	-2.02	0.0436
s_3	-0.4424	0.1281	-3.45	0.0006
s_5	0.9358	0.1294	7.23	0.0000

Table 4.4: The final model for the diabetes data, with $R^2 = 0.51$

Figure 4.2 shows that diagnostics plots for the fitted model, 4.27. The QQ-plot shows that the residuals are approximately normally distributed. The left panel depicts the residuals by fitted values. Although the residuals seem to have mean and variance approximately 0 and 1 respectively, there may be het-

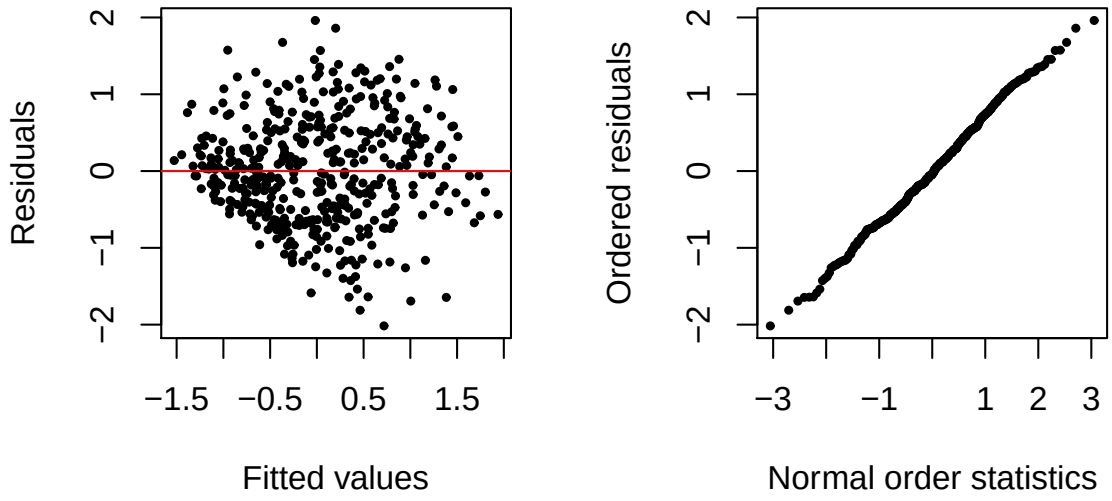

Figure 4.2: Diabetes data – diagnostics plots of the final model in 4.27. Left: residuals vs. fitted values. Right: quantile-quantile plot of the residuals.

eroscedasticity, since the residuals appear to be increasing with the fitted values. This is likely due to the fact that the responses do not appear to be normally distributed, and require a transformation (square root seems appropriate, in this case). However, since we are interested in the comparison between the methods, we do not include the details of the analysis of the transformed data here.

4.5 Extensions and Future Plans

In variable selection algorithms the statistician can divide the *fixed* effects into two groups. The first group consists of variables that are 'locked', in the sense that they will be included in any model being considered. The second group

consists of putative variables, from which only a subset may be selected. To facilitate model selection in situations where the number of putative variables is large, we assumed in model 4.2 that the mean of the response is a linear combination of these two groups of variables. The $\mathbf{X}\boldsymbol{\beta}$ term represents the 'locked' variables, and the $\mathbf{Z\Gamma u}$ term represents the putative variables. We assumed that the putative variables are realizations of a mixture distribution in which one component consists of all the variables that have no effect on the response (the 'null' set, for which the indicator variables γ_k are 0), while the variables that do have an effect on the response are assumed to be realizations of a normal distribution.

This allowed for an efficient estimation of the posterior probability of the indicator variables via the EM algorithm. However, the selected variables are included in the final model as fixed effects. There are situations in which the statistician wants to 'lock' additional *random* effects in the model. For example, in biological applications (e.g. QTL analysis) one may want to include breed, or kinship information as a random effect. This can be easily done, using the same method that we used to estimate the variance parameters, since the update equations for the EM algorithm extend to any number of variance components (see the general formulation of the estimation in Section 8.3.b in [60]). Our model then becomes

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_0\mathbf{v} + \mathbf{Z\Gamma u} + \boldsymbol{\varepsilon} && (4.28)\\ \boldsymbol{\varepsilon} &\sim N\left(\mathbf{0}_N, \sigma_e^2\mathbf{I}_N\right) && (4.29)\\ \mathbf{Z}_0\mathbf{v} &\sim N\left(\mathbf{0}_N, \sigma_0^2\mathbf{I}\right) && (4.30)\\ \mathbf{Z\Gamma u} &\sim N\left(\mathbf{Z\Gamma}\boldsymbol{\mu}, \sigma^2\mathbf{Z\Gamma}^2\mathbf{Z}'\right). && (4.31) \end{aligned}$$

Hence, the statistician can specify the matrix $\mathbf{Z}_0$, and we have just one ad-

ditional parameter to estimate (σ_0^2). This generalizes easily to any number of variance components.

Our model was developed with the goal of analyzing very large QTL data sets in order to detect loci that are associated with biological traits. Therefore, adding interactions between fixed effects and SNPs is a very useful extension. For example, if the quantitative trait is the Forced Expiratory Volume (FEV), it may be the case that the significant loci have a different effect on the response for heavy smokers, than for non-smokers. This requires a relatively simple modification to the model and the estimation procedure. We are also interested in applying our approach to eQTL data, where the responses are normalized gene expression, and to experiments involving repeated measure of the quantitative traits.

Another enhancement that we are currently pursuing is to extend the model to the generalized linear model framework in order to deal with binomial and Poisson responses, as well as censored survival times using the artificial Poisson model as described by [74]. A viable approach is to combine our mixture model framework with iterative estimation procedures such as Double Hierarchical Generalized Linear Models (DHGLM, [51]), which are extremely flexible, and allow to specify non-linear models for the means, variance, and dispersion parameters, separately.

4.6 Conclusions

We developed a model-based, empirical Bayes approach to variable selection. The idea of treating the putative variables as random effects induced shrinkage

estimation, which resulted in increased power and a significantly faster convergence, compared with simulation-based methods. Furthermore, a couple of computational tricks allowed us to increase the speed of our algorithm, to handle a large number of putative variables, and to control the multicollinearity in the model. Through simulations and case studies we confirmed that our approach to variable selection provides excellent results in terms of power, accuracy, and speed.

BIBLIOGRAPHY

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16:412–424, 2000.

[3] Haim Bar, James Booth, Elizabeth D. Schifano, and Martin T. Wells. Laplace approximated em microarray analysis: An empirical bayes approach for comparative microarray experiments. *Statistical Science*, 25(3):388–407, 2010.

[4] Haim Bar and Elizabeth Schifano. *lemma: Laplace approximated EM Microarray Analysis*, 2010. R package version 1.3-1.

[5] Haim Bar and Elizabeth D. Schifano. Empirical and fully bayesian approaches for random effects models in microarray data analysis. *Statistical Modelling*, 11(1):71–78, 2011.

[6] Haim Y. Bar, James G. Booth, and Martin T. Wells. A mixture-model approach for parallel testing for unequal variances. *Statistical Applications in Genetics and Molecular Biology*, 10(1), 2011. Article 55.

[7] M. S. Bartlett. Properties of sufficiency and statistical tests. *Proceedings of The Royal Society of London. Series A, Mathematical and Physical Sciences (1934-1990)*, 160:268–282, 1937.

[8] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate-a practical and powerful approach to multiple testing. *Journal Of The Royal Statistical Society Series B*, 57(3):499–517, 1995.

[9] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

[10] Dennis D. Boos and Cavell Brownie. Bootstrap Methods for Testing Homogeneity of Variances. *Technometrics*, 31(1):69–82, 1989.

[11] Dennis D. Boos and Cavell Brownie. Comparing variances and other measures of dispersion. *Statistical Science*, 19(4):571–578, 2004.

[12] G. E. P. Box. Non-normality and tests on variances. *Biometrika*, 40:318–335, 1953.

[13] L. Breiman. Better Subset Regression Using Nonnegative Garrote. *Technometrics*, 104:373–384, 1995.

[14] L. Breiman and J. Friedman. Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association*, 80:580–598, 1985.

[15] Morton B. Brown and Alan B. Forsythe. Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, 69(346):364–367, 1974.

[16] Ronald W. Butler. *Saddlepoint Approximations with Applications*. Cambridge, 2007.

[17] Long Cai, Nir Friedman, and X. Sunney Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–362, March 2006.

[18] M J Callow, S Dudoit, E L Gong, T P Speed, and E M Rubin. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Research*, 10(12):2022–9, 2000.

[19] G. Casella and E. Moreno. Objective Bayesian Variable Selection. *Journal of the American Statistical Association*, 101:157–167, 2006.

[20] Siddhartha Chib. *Markov chain Monte Carlo methods: computation and inference*, volume 5 of *Handbook of Econometrics*. Elsevier, January 2001.

[21] Alejandro Colman-Lerner, Andrew Gordon, Eduard Serra, Tina Chin, Orna Resnekov, Drew Endy, C. Gustavo Pesce, and Roger Brent. Regulated cell-to-cell variation in a cell-fate decision system. *Nature*, 437(7059):699–706, September 2005.

[22] Aedin Culhane, Guy Perriere, and Desmond Higgins. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, 4(1):59, 2003.

[23] Aedin Culhane, Jean Thioulouse, Guy Perriere, and Desmond Higgins.

MADE4: An R package for Multivariate Analysis of Gene Expression Data. *Bioinformatics*, 21(11):2789–90, 2005.

[24] F. N. David. The z-test and symmetrically distributed random variables. *Biometrika*, 46:123–129, 1959.

[25] F. N. David and N. L. Johnson. A Method of Investigating the Effect of Nonnormality and Heterogeneity of Variance on Tests of the General Linear Hypothesis. *The Annals of Mathematical Statistics*, 22(3):382–392, 1951.

[26] F. N. David and N. L. Johnson. The Effect of Non-Normality on the Power Function of the F-Test in the Analysis of Variance. *Biometrika*, 38(1/2):43–57, 1951.

[27] N. G. de Bruijn. *Asymptotic Methods in Analysis*. Dover: New York, 1981.

[28] Ralph J. DeBerardinis, Anthony Mancuso, Evgueni Daikhin, Ilana Nissim, Marc Yudkoff, Suzanne Wehrli, and Craig B. Thompson. Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49):19345–19350, December 2007.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal Of The Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[30] C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50:1096–1121, 1955.

[31] Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[32] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least Angle Regression. *Annals of Statistics*, 32:407–499, 2004.

[33] Bradley Efron and Carl Morris. Data Analysis Using Stein's Estimator and its Generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

[34] Bradley Efron and Robert Tibshirani. Empirical bayes methods and false

discovery rates for microarrays. *Genetic epidemiology*, 23(1):70–86, June 2002.

[35] Bradley Efron, Robert Tibshirani, J. Storey, and V. Tusher. Empirical bayes analysis of microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[36] Bradley Efron, Brit B. Turnbull, and Balasubramanian Narasimhan. *locfdr: Computes local false discovery rates*, 2008. R package version 1.1-6.

[37] Evan E. Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore, and Joseph H. Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450, June 2010.

[38] Andrew P. Feinberg and Rafael A. Irizarry. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences*, 107(suppl 1):1757–1764, January 2010.

[39] J. Friedman and B. Silverman. Flexible Parsimonious Smoothing and Additive Modeling. *Technometrics*, 31:3–21, 1989.

[40] J. L. Gastwirth, Y. R. Gel, and W. Miao. The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice. *Statistical Science*, 24(3):343–360, oct 2010.

[41] E.I. George and R.E. McCulloch. Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, 88:881–889, 1993.

[42] G.H. Golub and C.F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, MD, US, 1996.

[43] Kasper D. Hansen, Winston Timp, Hector C. Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G. McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep, Eirikur Briem, Kun Zhang, Rafael A. Irizarry, and Andrew P. Feinberg. Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8):768–775, June 2011.

[44] T. Hastie and R. Tibshirani. *Generalized Additive Models.* Chapman and Hall, New York, NY, USA, 1990.

[45] Joshua W.K. Ho, Maurizio Stefani, Cristobal G. dos Remedios, and Michael A. Charleston. Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics*, 24:i390–i398, July 2008.

[46] Y. Hochberg and A. C. Tamhane. *Multiple comparison procedures*. John Wiley & Sons, Inc., New York, NY, USA, 1987.

[47] R. R. Hocking. The Analysis and Selection of Variables in Linear Regression. *Biometrics*, 32:661–675, 1976.

[48] Jason C. Hsu. Simultaneous confidence intervals for all distances from the "best". *Annals of Statistics*, 9:1026–1034, 1981.

[49] J T Gene Hwang and Peng Liu. Optimal tests shrinking both means and variances applicable to microarray data analysis. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article36, 2010.

[50] Willard James and Charles Stein. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, (1):361–379, 1961.

[51] Y. Lee, J.A. Nelder, and Y. Pawitan. *Generalized Linear Models with Random Effects.* Chapman & Hall/CRC, London, UK, 2006.

[52] H. Levene. Robust tests for equality of variances. In I. Olkin, editor, *Contributions to probability and statistics*. Stanford Univ. Press., Palo Alto, CA, 1960.

[53] Jeffrey M. Levsky, Shailesh M. Shenoy, Rossanna C. Pezo, and Robert H. Singer. Single-cell gene expression profiling. *Science (New York, N.Y.)*, 297(5582):836–840, August 2002.

[54] D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. The bugs project: Evolution, critique and future directions (with discussion). *Statistics in Medicine*, 28:3049–3082, 2009.

[55] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, An-

drew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.

[56] Jessica C. Mar, Nicholas A. Matigian, Alan Mackay-Sim, George D. Mellick, Carolyn M. Sue, Peter A. Silburn, John J. McGrath, John Quackenbush, and Christine A. Wells. Variance of Gene Expression Identifies Altered Network Constraints in Neurological Disease. *PLoS Genetics*, 7(8), August 2011.

[57] Jessica C. Mar, Renee Rubio, and John Quackenbush. Inferring steady state single-cell gene expression distributions from analysis of mesoscopic samples. *Genome Biology*, 7:R119+, December 2006.

[58] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analisys*, page 124. CA: Academic, San Diego, 1979.

[59] Nicholas F. Marko, John Quackenbush, and Robert J. Weil. Why is there a lack of consensus on molecular subgroups of glioblastoma? understanding the nature of biological and statistical variability in glioblastoma expression data. *PLoS ONE*, 6(7):e20826, 07 2011.

[60] Charles E. McCulloch, R. Searle Shayle, and George Casella. *Variance Components.* Wiley-Interscience, New York, NY, US, 1992.

[61] Peter Mueller, Giovanni Parmigiani, and Kenneth Rice. Fdr and bayesian multiple comparisons rules. In *Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics*, June 2006.

[62] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden. Regulation of noise in the expression of a single gene. *Nature Genetics*, 31(1):69–73, May 2002.

[63] K. B. Petersen and M. S. Pedersen. The matrix cookbook, Oct 2008. Version 20081110.

[64] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[65] T. Ravasi, C. Wells, A. Forest, D.M. Underhill, B.J. Wainwright, A. Aderem,

S. Grimmond, and D.A. Hume. Generation of diversity in the innate immune system: macrophage heterogeneity arises from gene-autonomous transcriptional probability of individual inducible genes. *Journal of Immunology*, 168(1):44–50, 2002.

[66] Zachary J. Reitman, Genglin Jin, Edward D. Karoly, Ivan Spasojevic, Jian Yang, Kenneth W. Kinzler, Yiping He, Darell D. Bigner, Bert Vogelstein, and Hai Yan. Profiling the effects of isocitrate dehydrogenase 1 and 2 mutations on the cellular metabolome. *Proceedings of the National Academy of Sciences*, 108(8):3270–3275, February 2011.

[67] D T Ross, U Scherf, M B Eisen, C M Perou, C Rees, P Spellman, V Iyer, S S Jeffrey, M Van de Rijn, M Waltham, A Pergamenschikov, J C Lee, D Lashkari, D Shalon, T G Myers, J N Weinstein, D Botstein, and P O Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, 24(3):227–35, 2000.

[68] Gideon E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[69] Gordon K. Smyth. Linear models for empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. Article 2.

[70] Gordon K Smyth. *Limma: linear models for microarray data*, pages 397–420. Springer, New York, 2005.

[71] F. Soliman, C. E. Glatt, K. G. Bath, L. Levita, R. M. Jones, S. S. Pattwell, D. Jing, N. Tottenham, D. Amso, L. H. Somerville, H. U. Voss, G. Glover, D. J. Ballon, C. Liston, T. Teslovich, T. Van Kempen, F. S. Lee, and B. J. Casey. A genetic variant bdnf polymorphism alters extinction learning in both mouse and human. *Science*, 327:863–866, 2010.

[72] D.J. Spiegelhalter, K.R. Abrams, and J.P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, chapter 6. John Wiley & Sons Ltd.

[73] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statististical Society, Series B.*, 58(1):267–288, 1996.

[74] J. Whitehead. Fitting Cox's Regression Model to Survival Data Using GLIM. *Applied Statistics*, 29:268–275, 1980.

[75] Verena Zuber and Korbinian Strimmer. High-Dimensional Regression and Variable Selection Using CAR Scores. *Statistical Applications in Genetics and Molecular Biology*, 10(1):407–499, 2011.