Surrogate Variable Analysis

Jeffrey Tullis Leek

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2007

Program Authorized to Offer Degree: Biostatistics

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Jeffrey Tullis Leek

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

John D. Storey

Reading Committee:

Joshua M. Akey

Norman E. Breslow

John D. Storey

Date: 13 December 2007

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this thesis is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 84106-1346, 1-800-521-0600, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform.

Signature 

Date 12/13/07

University of Washington

Abstract

Surrogate Variable Analysis

Jeffrey Tullis Leek

Chair of the Supervisory Committee:
Associate Professor John D. Storey
Biostatistics

Modern high-throughput molecular biology experiments measure data for thousands of related features and seek to rank those features for association with some variables of experimental or clinical importance. The process of ranking features for association with primary variables is complicated by genetic, environmental, and technical factors that influence hundreds or thousands of features at a time. In high-dimensional experiments these factors are often unknown, unmeasured, or incapable of being tractably modeled. Consistent patterns of variation across features due to unmeasured or unmodeled factors can confound the relationship between the primary variables and the measured features. In this thesis we provide a statistical framework for modeling large-scale noise dependence caused by unmeasured or unmodeled factors in high-throughput data. We argue that estimating the sources of noise dependence is more appropriate than estimating the pairwise covariance between all features when the number of features is large. A direct connection is made with the well-studied problem of multiple testing dependence, which typically focuses on the distribution of P-values from multiple testing procedures. We introduce the concept of surrogate variables, estimable linear combinations of the true unmeasured or unmodeled factors causing noise dependence, that can be included when modeling the relationship between the primary variables and the feature level data. We also propose

algorithms for estimating surrogate variables based on principal component analysis of relevant subsets of features. Under certain conditions accounting for the estimated surrogate variables asymptotically corrects the ranking and error rate estimation in high-throughput data analysis. We also discuss pathological situations when surrogate variables can not be estimated. To illustrate the power of this approach, we apply our estimates of the surrogate variables to improve reproducibility in a large clinical gene expression study of trauma related outcomes.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

This research was supported in part by an ARCS Fellowship, NHGRI training grant T32 HG00035, and funding from the National Institutes of Health (NIH NIGMS U54 GM062119 and R01 HG002913).

Thanks to Professors Joshua Akey, Norman Breslow, Galen Shorack, Eric Schadt, and John Storey for being on my Ph.D. committee. Their support and ideas were invaluable and have lead to great improvements in this work and many new ideas for future research.

I am particularly grateful to John Storey for his encouragement, insight, and support during my graduate education. His excitement and commitment are contagious and he has taught me an incredible amount about how to be a thoughtful statistician and scientist.

Of course, this accomplishment would not have been possible without the love and support of my friends and family. My parents, Max and Susan, have always encouraged me to pursue my interests and taken the time to listen to my ideas and stories. My brother Luke and my sister Anna have always been there for me, even when work kept me from calling as often as I should.

Thanks to Daniel, Quenten and Mark for having my back no matter what is going down. Thanks to the members of Thunderstats, Random Charge, and Wardrobe Malfunction for giving me perspective, and 10 IMA championships.

Finally, I want to thank my wife, Leah Jager. Every success is sweeter, every failure is less painful, and my life is so much richer thanks to her.

Chapter 1

INTRODUCTION

A traditional scientific experiment is designed to quantify associations between a relatively small number of measured variables. One example of a traditional experiment that has had a major impact on human health is the interventional clinical trial. In a clinical trial, patients are randomly assigned to either an experimental treatment or a placebo. After a period of time, an outcome variable quantifying the health of the patients is collected. The goal of the trial is to determine if there is any average difference in the outcome between the treated and control patients. A number of other variables may be measured and accounted for when modeling the relationship between treatment and outcome, but the goal of the experiment is still primarily to measure a single association.

When studying complicated and poorly understood systems the goal is often more exploratory, e.g. ranking a number of measured features for association with a small number of outcome variables. Examples of this new type of exploratory and data-rich experiment include identifying the genetic variations underpinning diabetes [15, 25], selecting the demographic variables that are most influential in determining the success of a marketing campaign [77], or screening chemicals for desirable properties [45]. Each of these experiments measures data for only a small number of outcomes, but simultaneously compares these outcomes to hundreds, thousands or even millions of features with the goal of ranking the features for association.

In the biological sciences, there has been a particularly large growth in these so-called "high-throughput" experiments in the past two decades precipitated by the hu-

man genome project [90, 88]. A focus of many high-throughput experiments has been to quantify variation in the biochemical processes underlying the central dogma of molecular biology, which describes the flow of information in all human cells. Put simply, the central dogma states that DNA sequence is transcribed into RNA molecules which are then translated into proteins that determine the structure and function of the cell [21]. Variations in the process of transcription and translation, along with changes in the structure and interactions between the various constituent molecules (DNA, RNA, protein) in large part characterize the function of a cell in changing environments [2]. However, given the complicated and poorly understood nature of these processes, a complete picture of cellular function must be developed by monitoring variations in a large number of features simultaneously [42]. High-throughput experiments are an efficient study design for this type of large scale analysis.

Commercial high-throughput technologies were first developed less than two decades ago [63, 54]. The invention of the microarray technology is one reason for the explosive growth of high-throughput experiments in the last two decades[73]. The term microarray now refers to a small chip that contains a large number of biological assays that are performed simultaneously. Microarrays have been developed to measure genetic variation [95, 37], gene expression [73, 63], binding sites between proteins and DNA [67], epigenomic variation such as DNaseI hypersensitivity [26], chromosomal copy number variation [78], and protein expression [56] on a genome-wide scale. Table 1 displays the most common high-throughput technologies, the type of variation they capture, and the approximate number of features on a typical microarray.

Just as in a traditional experiment, microarray experiments are performed to quantify associations among variables. Rather than performing thousands of independent experiments, a microarray performs thousands of related experiments simultaneously. Usually there is no specific *a priori* information about the relative importance of the features, so the goal is to rank them based on their association with the outcome or experimental variables. The mostly highly ranked features above some threshold

Table 1.1: A table describing the most common biological high-throughput microarray technologies.

Technology	Type of Features	Typical # of Features
Array CGH	chromosomal copy #	100,000+
SNP Array	SNPs	10,000 - 500,000+
DNA Microarray	gene expression	5,000 - 50,000+
ChIP-Chip	DNA-protein binding	100,000+
DNaseI Array	DNase I hypersensitivity	100,000+
Protein Microarray	protein expression	5,000-10,000+
Tissue Microarray	tissue-specific gene expression	up to 1,000

are followed up in subsequent detailed experiments. Since laboratories have finite resources and only a small number of the identified features can be individually analyzed in validation or follow-up studies, it is important to (1) accurately rank the features according to their true association with the primary outcome variables and (2) accurately estimate the level of noise among the features that appear to show true association.

1.1 Microarrays for Gene Expression

It is estimated that there are approximately 20,000-25,000 protein coding genes in the human genome [19]. For comparison, the relatively simple model organism *Caenorhabditis elegans* has approximately 20,000 protein coding genes [87]. Thus to account for our complexity much effort has turned to understanding the regulation of transcription and translation in humans and model organisms. In fact, the first high-throughput technology to receive widespread use in molecular biology is the gene expression microarray, which measures variation in the process of transcription

for thousands of DNA sequences simultaneously[73, 63]. A number of novel gene expression arrays have been developed that differ in the number and type of mRNA profiled, the design of the array, and the type of samples compared. Even though there are many types of microarrays, the basic process for each is very similar.

A modern gene expression microarray consists of thousands of probes printed on to specific locations on a glass or plastic slide [41]. Each probe is a large number of identical oligonucleotides, or short sequences of nucleotides, that perfectly match a target sequence of DNA. Usually each target DNA sequence codes for all or part of a specific mRNA molecule from the organism being studied. Total RNA samples are isolated from the sample of interest and complementary DNA is created via reverse transcription. The cDNA molecules are then labeled with a fluorescent dye and hybridized to the microarray. If the sequence of the cDNA molecule is complementary to the sequence of the oligonucleotide for a specific probe, then the cDNA will preferentially bind to that probe. The microarray is then scanned using a laser and the amount of fluorescence for each spot is recorded. The intensity of the fluorescence is a measure of the relative abundance of that particular mRNA molecule [13, 63].

There are a number of important computational hurdles in arriving at a meaningful quantitative measure of expression for each gene on a microarray. On a strictly technical level, statistical methods have been developed to address the issues of segmenting and adjusting the background of the image from the microarray scan [97, 1] and creating summary statistics for the level of intensity for each probe [20]. There is also a large literature dedicated to statistical methods for normalizing microarray experiments, with particular emphasis on accounting for technical factors such as array or dye bias that may affect the relationship between mRNA abundance and intensity [92, 97]. The end result of these preprocessing steps is a single quantitative value indicating the relative abundance of mRNA molecules for a given gene measured for each microarray.

In a differential expression experiment, n biological samples are obtained, total

mRNA is extracted and analyzed with gene expression microarrays measuring the relative level of transcription for m genes. The normalized and log transformed expression data are arranged into an $m \times n$ matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)^T$ where the vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})$ represents the expression values for gene i. In addition to the expression data, data for primary variables $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ describing the study design or experimental outcomes is collected. In a microarray experiment, each individual is represented by a set of of parameters $\boldsymbol{\theta}_j$. Given that set of parameters, the gene expression measurements are a random multivariate observation $\mathbf{x}_{\cdot j}|\boldsymbol{\theta}_j$ with mean $\mu(\boldsymbol{\theta}_j)$ and variance $\boldsymbol{\Sigma}(\boldsymbol{\theta}_j)$, where $\boldsymbol{\Sigma}$ is usually not dependent on $\boldsymbol{\theta}_j$.

The goal of the experiment is then to rank the genes for association with the primary variable based on a regression model [75]. In the language of regression, the expression values for a gene, $\mathbf{x}_{i\cdot}$, act as the dependent variable in the regression. Individuals are randomly sampled from a population, sometimes in a stratified manner, as is the case in a randomized experiment with an intervention. For each individual, a set of covariates is measured in addition to the dependent gene expression. The analysis proceeds by regressing each dependent variable, i.e. each gene's expression, on a subset of the independent covariates, potentially adjusting for others just as in a classical regression. In other words, the the analysis of expression data is no different than a study where n patients are randomly sampled, their sex and age is recorded, as well as several dependent variables such as height, weight, and blood pressure and each dependent variable is regressed on age and sex.

There are two main classes differential expression studies performed in practice: *static* and *timecourse*. A static differential expression analysis compares the average level of expression for each gene across a fixed number of conditions. One highly cited static differential expression study compared expression of thousands of genes across breast cancer subtypes defined by *BRCA1* and *BRCA2* mutations and sporadic cases [38]. For each gene, the goal is to determine which genes showed changes in average expression between *BRCA1* and *BRCA2* positive tumors. The primary variable for the

BRCA study is just a single vector $\mathbf{y}$ of indicators of tumor type for each microarray. On the other hand, a timecourse differential expression study seeks to identify genes whose expression varies with time. One recent example of a timecourse differential expression study monitored the expression patterns of genes in kidney samples from patients of various ages [69]. The primary variable in this case is a vector $\mathbf{y}$ consisting of the age of the person sampled for each microarray. The goal is to find the genes that showed any smooth age-related pattern.

1.2 Differential Expression

The statistical methods developed in this dissertation are inspired by the problem of analyzing data from differential expression studies. However, the fundamental mathematical structure translates directly across many types of high-throughput studies. For each gene, a differential expression study fits a model relating the expression for that gene to the primary variables.

$$\begin{aligned} x_{ij} &= f_i(\mathbf{y}_j) + e_{ij} \\ x_{ij}|\mathbf{y}_j &\sim G(x_{ij}|\mathbf{y}_j, \boldsymbol{\theta}_i) \\ f_i(\mathbf{y}_j) &= \mathbf{E}_G[x_{ij}|\mathbf{y}_j, \boldsymbol{\theta}_i] \end{aligned} \tag{1.1}$$

In model 1.1, x_{ij} is the expression for gene i on array j, $f_i(\cdot)$ is a function relating the primary variable to expression for gene i and e_{ij} is random noise with mean zero. For each gene the parameters $\boldsymbol{\theta}_i$ quantify the relationship of interest and statistical inference for $\boldsymbol{\theta}_i$ proceeds by testing the m hypotheses:

$$H_{0i} : \boldsymbol{\theta}_i \in \Theta_0 \quad \text{vs.} \quad H_{1i} : \boldsymbol{\theta}_i \in \Theta_1 \tag{1.2}$$

Note that parameters vary from gene to gene, but the two hypotheses remain the same. This is a common feature of high-throughput studies, where the goal is to test each feature for the same specific signal.

To make these general ideas more concrete, consider the two examples of gene expression studies discussed above. In the *BRCA* experiment, there is only a single primary variable $\mathbf{y} = (y_1, \ldots, y_n)^T$, an indicator of whether tumor sample j is *BRCA1* positive, and $f(\cdot)$ defines a simple linear model:

$$x_{ij} = \mu_i + \beta_i y_j + e_{ij}$$

where β_i is equal to the average change in expression between *BRCA1* and *BRCA2* positive tumors for gene i. The goal of the experiment is then to determine which of the β_i are equal to zero and which are non-zero.

$$H_{0i} : \beta_i = 0 \qquad \text{vs.} \qquad H_{1i} : \beta_i \neq 0$$

In the timecourse experiment, there is again a single primary variable, $\mathbf{y} = (y_1, \ldots, y_n)^T$, giving the age of the patients in the study. We can use the same form with a slightly more complicated $f(\cdot)$ to write down a model relating expression to age:

$$x_{ij} = \mu_i + \sum_{k=1}^{d} \beta_{ik} s_k(y_j) + e_{ij}$$

Here the functions $s_k(\cdot)$ parameterize a continuous model of age. For example, the $s_k(\cdot)$ could be a d-dimensional polynomial of age, e.g. $s_k(y_j) = y_j^k$. The goal is then to determine if any of the genes show non-constant patterns of expression over time, or if any of the β_{ki} are non-zero.

$$H_{0i} : \beta_{ki} = 0 \ \ \forall k \qquad \text{vs.} \qquad H_{1i} : \exists\, \beta_{ki} \neq 0$$

If we let $d = 2$, then we are trying to identify all of those genes who on average fluctuate linearly or quadratically with respect to age.

1.3 Multiple Testing Procedures

Regardless of the specific choice of the model $G(x_{ij}|\mathbf{y}_j, \boldsymbol{\theta}_i)$ and the hypotheses, statistical inference for high-throughput experiments such as gene expression studies is

performed using multiple testing procedures (MTP) [27, 84]. A multiple testing procedure has two steps: (1) rank the features with respect to the strength of evidence for the alternative hypothesis H_{1i} and (2) draw a cutoff and calculate some error measure for that cutoff. Unlike single hypothesis testing, where we intend to reach a single decision about an association between two variables, what really matters in a high-throughput experiment is the ordering of the features from most significant to least [80].

Step one of a multiple testing procedure is performed by evaluating a function of each feature's data that quantifies it's association with $\mathbf{Y}$ and ordering the features based on this value. The most general form for this function is: $\mathcal{S}(\mathbf{x}_i|\mathbf{Y},\mathbf{H}) \geq 0$, which incorporates all of the data collected ($\mathbf{Y}$) and any previous knowledge about the relative ordering of any particular feature ($\mathbf{H}$). In practice, previous knowledge about the ordering of the features $\mathbf{H}$ is rarely used in MTPs, either because that information is unreliable or because it is not clear how to appropriately take advantage of the information quantitatively. The general form of these statistics in this case is simply $\mathcal{S}(\mathbf{x}_i|\mathbf{Y})$.

The function values $\mathcal{S}(\mathbf{x}_i|\mathbf{Y})$ for each gene in a gene expression analysis are based on the parameter estimates from model 1.1. $\mathcal{S}(\mathbf{x}_i|\mathbf{Y})$ can be a standard statistic such as the two-sample t-statistic for the *BRCA* study [43] or the F-statistic comparing the null and alternative model fits for the kidney aging study [85]. It can also be a much more complicated statistic based on the data for all of the genes such as an empirical Bayes statistic [31] or the recently proposed ODP statistics [80, 81]. The genes are then ranked according to the value of $\mathcal{S}$ from largest to smallest.

Step two is performed by comparing the values of $\mathcal{S}(\mathbf{x}_i|\mathbf{Y})$ to the values we would expect under the null hypothesis of no relationship between $\mathbf{x}_i$ and $\mathbf{Y}$ defined by model 1.2. However, when making probabilistic statements in the multiple testing case the usual error measures for single hypothesis testing, such as the type I error rate, are no longer straightforwardly interpretable. For example, if we call all features

are significant with P-values, $p_i \leq \alpha$, and m tests are performed, we would expect to claim $m \times \alpha$ true associations even if there was no relationship between any of the $\mathbf{x}_i$ and $\mathbf{Y}$. If m is on the order of thousands, then even for small values of α we will likely have many false associations. Thus, when performing multiple hypothesis tests, it is more appropriate to consider probability statements about m random values of $\mathcal{S}(\mathbf{x}_i|\mathbf{Y})$.

Once a function value has been calculated for each feature, any cutoff λ defines a subset of features called significant. The potential outcomes when thresholding a set of function values for a MTP are described in Table 1.2. The key quantities for error rate estimation are the total number of features rejected for a specific cutoff (R) and the number of false positives (V).

Table 1.2: A table describing the potential outcomes when thresholding function values in a MTP.

Hypothesis	Accept Null	Reject Null	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

Regardless of the choice of cutoff and ranking function, it is important to estimate the expected number or proportion of false positives among those features we call significant. The three most common error rates considered in practice are the family wise error rate (FWER) [94], the expected number of false positives (EFP)[80], and

the false discovery rate (FDR) [9, 79].

$$\begin{aligned} \text{FWER} &= \mathbf{Pr}(V \geq 1 | \mathcal{S}, \lambda) \\ \text{EFP} &= \mathbf{E}(V | \mathcal{S}, \lambda) \\ \text{FDR} &= \mathbf{E}\left(\frac{V}{R} | \mathcal{S}, \lambda\right) \mathbf{Pr}(R > 0 | \mathcal{S}, \lambda) \end{aligned}$$

When estimating error rates for multiple testing procedures, the goal is to be conservative, i.e. on average we would like to overestimate, rather than underestimate, the true error. If we overestimate the error rate then we will not claim more false associations than claimed. Almost all error rate estimators are designed to be conservatively biased [94, 9, 79]. Ideally the estimators will be only slightly conservative in order to maximize the number of true positives [84]. Another global measure of significance that is of interest is the proportion of truly null hypotheses among those tested, π_0. Estimators of π_0, like estimators of error rates, are also designed to be conservatively biased, again we would like to overestimate the number of features that follow the null hypothesis H_{0i} [79].

1.4 Dependence in High-Throughput Studies

In a high-throughput study, thousands of related features are measured on each sample. In practice the exact relationship among the features is rich and largely unknown. In other words, data from a high-throughput experiment do not represent the result of thousands of independent experiments, rather thousands of closely related experiments subjected to similar conditions. For example, when gene expression information is obtained for n samples, all m genes for each array are affected by the same variables. Similarly, in a whole genome association study the distribution of many SNPs for any individual may be affected by population structure [57].

The interrelationship between features in a high-dimensional study results in dependence. There are two general categories of dependence in high-throughput studies: signal dependence and noise dependence. Signal dependence refers to any structure

among the parameters that define the signal of interest. In the *BRCA1* study any structure among the β_i constitutes signal dependence. That is, the entire set of $\beta_1, \ldots, \beta_m$ has a distribution of values that has some scientifically meaningful interpretation, which should then also be interpretable in the context of how the null and alternative hypotheses were defined. For example, when comparing *BRCA1* to *BRCA2*, there may be an asymmetry to the nonzero β_i, which implies that differential expression tends to have a direction of over-expression in one of the two groups. Signal dependence can be beneficial for multiple testing procedures, and optimality properties for ranking features based on this type of structure have recently been derived [80, 81]. Noise dependence, on the other hand, refers to stochastic dependence between feature level data. It refers to a dependence structure in what we model as being random variables in our analysis. Since addressing noise dependence in high-throughput studies is the primary focus of this dissertation, we give a precise definition as follows.

Definition 1. Noise Dependence *Noise dependence exists between features i and i' in a study if* $\mathbf{Pr}(\mathbf{x}_i, \mathbf{x}_{i'}|\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i'}, \mathbf{Y}) \neq \mathbf{Pr}(\mathbf{x}_i|\boldsymbol{\theta}_i, \mathbf{Y})\mathbf{Pr}(\mathbf{x}_{i'}|\boldsymbol{\theta}_{i'}, \mathbf{Y})$.

In the case of model 1.1 any structure among the e_{ij} constitutes noise dependence. One example of this type of structure is a set of common random variables other than $\mathbf{Y}$ that affects the data for many different features. In statistical terms, these variables are called confounders, and they are often unmeasured or unmodeled in high-throughput experiments.

As a specific example, consider the *BRCA* study described above where $\mathbf{y}$ is an indicator of tumor subtype. Suppose that in addition to changes in expression being associated with tumor subtype, the age of the individuals also has a substantial influence on expression. So some genes exhibit differential expression with respect to tumor subtype, some with respect to age, and some with respect to both. If age is not included in model 1.1 when identifying differential expression with respect to

tumor subtype, this may (1) induce extra variability in the expression levels due to the effect of age, decreasing our power to detect associations with tumor subtype, (2) introduce spurious signal due to the fact that the effect of age on expression may be confounded with tumor subtype, or (3) induce long-range dependence in the apparent noise of the expression data, complicating any assessment of statistical significance for differential expression. In statistical terms, age is a confounder of the relationship between the gene expression values and tumor subtype. There is a large body of work focused on addressing confounding in observational studies, see, for example, [70] and references therein. This work is not designed for the setting of a large number of features m being tested. One approach for addressing confounding is to include the known confounders in the model relating the dependent and independent variables [70]. Unfortunately, even if age were measured, it may be one of dozens of available measured factors, making it statistically intractable to determine which to include in the model. Furthermore, even measured factors such as age may act on distinct sets of genes in different ways, or may interact with an unobserved factor, making the effect of age on expression difficult to model.

1.5 Surrogate Variable Analysis - A New Approach for Addressing Noise Dependence

In light of the difficulties that can be caused by noise dependence in the analysis of high-throughput studies, it is not surprising that a number of statistical methods have been designed to control estimates of significance under dependence. To date there exist two classes of adjustments based on the level of information used: adjustments to the P-values or null statistics distribution [10, 30] and data based adjustments [65]. P-value and null distribution adjustments have received the most attention in the statistical literature [10, 30], while data based adjustments have mostly been developed in the context of correcting for population stratification in association studies [65]. Yet to date there has not been a recognition of the common goal of these two

approaches in correctly performing a large-scale significance analysis in the presence of strong noise dependence. In this dissertation we propose a general framework that unites these two approaches based on the concept of surrogate variables.

To understand our general framework, first consider the general problem of estimating dependence across features. The data for each array from a high-throughput experiment consist of a set of measurements $\mathbf{x}_{\cdot j}$ that are interrelated. One measure of the linear dependence between the features is the covariance matrix, $\mathbf{\Sigma}$, where σ_{ij} is the covariance between the expression values for feature i and feature j. When the data are normally distributed, $\mathbf{\Sigma}$ completely defines the dependence structure between the features. An approach to accounting for noise dependence would be to estimate $\mathbf{\Sigma}$ and use this estimate to adjust downstream multiple testing procedures. This is what one would do when viewing the problem of noise dependence as a standard multivariate inference problem.

Unfortunately, the number of features is very large, so the corresponding covariance matrix has an intractably large number of terms to estimate. In a gene expression experiment, the number of features is on the order of thousands, and thus, a completely unspecified $\mathbf{\Sigma}$ has millions of elements to estimate. However, there are a relatively small number of samples $n \ll m$ that can be used to estimate these millions of parameters. One of the main contributions of this dissertation is to show that it is not necessary to estimate the covariance matrix, $\mathbf{\Sigma}$, which consists of fixed, population level parameters in order to completely account for linear dependence between features. In fact, we show that conditioning on an appropriately specified *random* matrix $\mathbf{V}$ of dimension $n \times k$ where $k \leq n$ is sufficient to eliminate all linear dependence between features, thereby avoiding the need to estimate the $\frac{(m+1)m}{2}$ population parameters of the covariance matrix.

We call any basis for the column space of $\mathbf{V}$ a set of surrogate variables and propose algorithms for estimating these variables. We also show how to directly incorporate $\mathbf{V}$ into a large-scale significance analysis. The identification, estimation,

and incorporation of these surrogate variables is what we call "surrogate variable analysis."

This result is due to a surprising reversal of the so-called curse of dimensionality [8] when analyzing data of the structure found in high-throughput experiments. The curse of dimensionality in high-dimensional data analysis refers to the increased difficulty of estimating parameters in high-dimensional spaces. However, in this case, we show that as the dimension of the data increases, we actually get better estimates of the surrogate variables. We also show that our direct approach for addressing noise dependence is a general framework, encompassing existing ideas such as the data based adjustments that have been proposed in the genetics literature. With these results in hand we show that careful estimation of the surrogate variables allows us to capture dependence in situations where these data based adjustments fail.

1.6 An Outline

A brief outline of this dissertation is as follows. Chapter 2 demonstrates the difficulties caused by noise dependence in high-throughput studies through the use of a simple illustrative example. Chapter 3 introduces surrogate variables through a general theoretical framework of noise dependence for normal high-throughput data and extends this framework to a much broader class of distributions. Chapter 4 proposes new surrogate variable estimation algorithms based on principal components that can be efficiently calculated with the singular value decomposition and contrasts these algorithms with the current state of the art for multiple testing dependence. Chapter 5 presents the results of proof of concept analyses in static, timecourse and genetical genomics differential expression experiments. In Chapter 6 we apply the newly developed surrogate variable estimators to analyze the data from the Inflammation and the Host Response to Injury Glue Grant, a large collaborative study of the gene expression response to blunt trauma. We show that applying the surrogate variable analysis technique we have developed improves reproducibility across four phases of

the study. Chapter 7 focuses on theoretical results related to the general statistical framework for noise dependence and the estimates of the surrogate variables. Chapter 8 provides some brief concluding remarks and ideas for extensions of the surrogate variable concept.

Chapter 2

THE EFFECT OF NOISE DEPENDENCE

In the development of multiple testing procedures, the error terms in model 1.1 are often assumed to be independent, or at most weakly dependent [9, 79, 83, 94]. In other words, it is assumed that there is only one pervasive source of differential expression among the genes. If this assumption is violated, how does that affect the significance results from MTPs? In this chapter we use simple simulated microarray studies to illustrate the effect of noise dependence on MTP. We show that noise dependence can affect both the ranking and error rate estimation of MTP and can substantially alter the global conclusions drawn from such a study.

2.1 Simulation Set-up

This chapter focuses on a simple simulated static differential expression experiment, where it is assumed that patients are randomized to one of two groups. Each simulated data set consists of 1,000 genes on 20 arrays and it is assumed that an equal number of patients are randomized to each group. Thus the outcome variable $\mathbf{y}$ can be written:

$$y_j = \begin{cases} 1 & j = 1, \ldots, 10 \\ 0 & j = 11, \ldots, 20 \end{cases}$$

Two sets of simulations will be presented, the first represents an idealized study where the noise across genes is completely independent and the second incorporates a second unmodeled factor that induces noise dependence across the simulated genes. To get a clear picture of the behavior of rankings and error rate estimates, it will be necessary to create several simulated studies drawn from the same hypothetical population.

The first simulation is drawn from following simple model for the expression x_{ij} of gene i on array j:

$$\begin{aligned} x_{ij} &= \mu_i + \beta_i y_j + u_{ij} \\ u_{ij} &\overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_{\mathrm{i}}^2) \end{aligned} \tag{2.1}$$

The population parameters μ_i, β_i, and σ_i^2 are simulated initially and then held fixed, in order to mimimic a hypothetical population. The population level parameters are drawn from the following distributions.

$$\begin{aligned} \mu_i &\overset{i.i.d.}{\sim} N(0,1) \\ \beta_i &\overset{i.i.d.}{\sim} \begin{cases} N(1.5,1) & i = 1,\ldots,300 \\ 0 & i = 301,\ldots,1000 \end{cases} \\ \sigma_i^2 &\overset{i.i.d.}{\sim} \text{Inverse Gamma}(10,9) \end{aligned}$$

Then several simulated data sets are created based on independent samples from the noise distribution, u_{ij}. This simulation imitates the behavior of a repeated randomized gene expression study, such as a repeated drug response microarray study, where RNA is isolated from blood samples drawn from patients randomly assigned to a fixed drug dose or placebo.

For comparison a second set of slightly more complicated simulated microarray data are created. The data no longer follow model 4.2; a second unknown or unmodeled factor **z** also linearly affects the expression of a large number of genes. The more complicated model for expression is then:

$$\begin{aligned} x_{ij} &= \mu_i + \beta_i y_j + \gamma_i z_j + u_{ij} \\ &= \mu_i + \beta_i y_j + e_{ij} \\ u_{ij} &\overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_{\mathrm{i}}^2) \end{aligned} \tag{2.2}$$

In the hypothetical drug response microarray study, z_j could be an indicator representing the sex of patient j, e.g. $z_j = 1$ if patient j is female and $z_j = 0$ if patient

j is male. In this case, γ_i would be the average change in expression for gene i between males and females. The parameters γ_i for this study will be added to the other population level parameters and will be drawn from the distribution:

$$\gamma_i \overset{i.i.d.}{\sim} \begin{cases} N(1.5, 1) & i = 101, \ldots, 700 \\ 0 & i = 1, \ldots, 100 \ \& \ 701, \ldots, 1000 \end{cases}$$

For these simulations we are assuming that the unmodeled variable $\mathbf{z}$ affects a very large proportion of the genes. Although this type of effect would be unusual for any but the most influential variables in practice, it allows us to clearly demonstrate the behavior of rankings and error rate estimates under dependence. Since the samples are randomized to the two groups, the distribution of z_j will vary across the simulated studies. So after fixing the population level parameters, each simulated study is formed by drawing values of the independent noise u_{ij} according to the distribution described above and drawing z_j from the distribution:

$$z_j \overset{i.i.d.}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right)$$

The simulated data are then created according to model 2.2.

In each simulated study we will apply the same multiple testing procedure. We fit the linear model 2.2 assuming independence across features by least squares. We order the features according to the t-statistic comparing expression in the two groups defined by $\mathbf{y}$ and P-values are calculated based on the t_{19} null distribution.

2.2 Results

Figure 2.1 shows heatmaps for two of the simulated microarray data sets under independent and noise dependent distributions for the errors. Each row in a heatmap represents the expression values for a specific gene across arrays, and each column represents a separate array; blue indicates low expression values and yellow indicates high expression values. The solid block of yellow in the upper left hand corner of

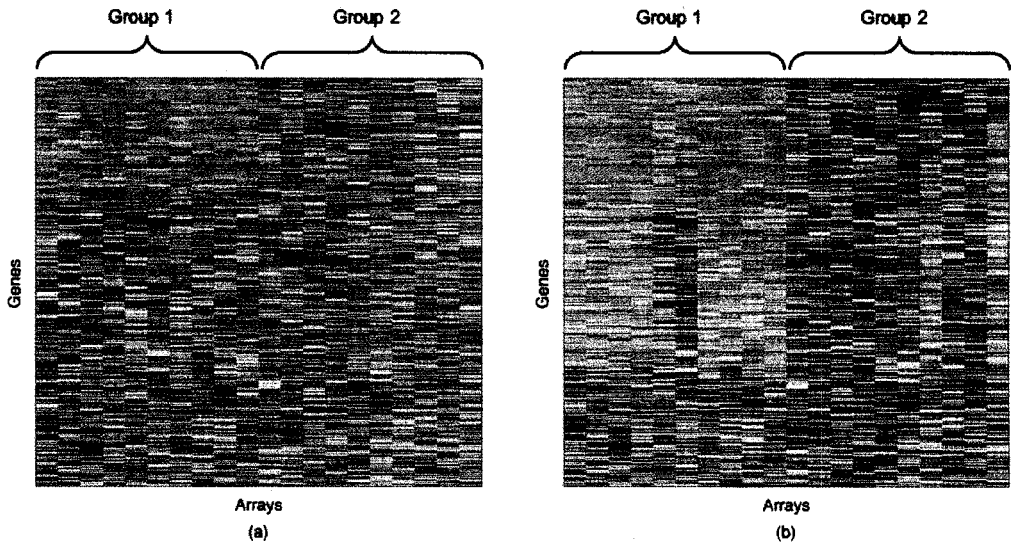

Figure 2.1: Heatmap images depicting the expression values for two simple simulated microarray studies, each with 1000 genes and 20 arrays for the (a) independent data and (b) noise dependent data.

Figures 2.1(a) and 2.1(b) is the differential expression signal with respect to the primary variable and the vertical yellow blocks of yellow in Figure 2.1(b) represents the differential expression signal due to the unmodeled factor. The rest of this section is devoted to demonstrating the effects of the variation in $\mathbf{z}$ on the ranking, error rate estimation, and global measures of significance for the genes in this simulated study.

In the first step of our simple multiple testing procedure, a t-statistic is calculated estimating the magnitude of the association between each feature and the primary variable. The t-statistic can be thought of as an estimator of the signal-to-noise ratio, $\frac{|\beta_i|}{\sigma_i}$. Ideally if we ranked the features from largest to smallest based on the magnitude of their t-statistics it would be the same as ranking the features by the true signal to noise ratio. Since the t-statistics are random variables, this is not true for any single simulated study, but we can look at the average ranking over all one hundred simulated studies. Figure 2.2(a) shows the average ranking from the t-statistics for the

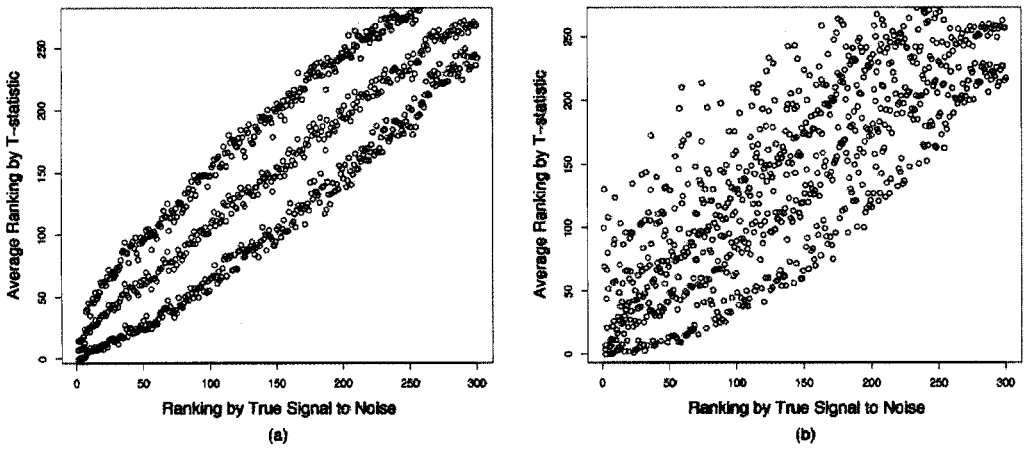

Figure 2.2: Plots of the average ranking from the t-statistics over 100 simulated studies (black dots) plus or minus one standard deviation (blue dots) for the (a) independent data and (b) noise dependent data.

differentially expressed genes under independent noise versus versus the true ranking defined by the signal to noise ratio. On average the true and estimated rankings show high correlation with reasonable variability. Figure 2.2(b) is the same plot for the noise dependent data. Under noise dependence, the estimated ranking shows much less correlation with the true ranking, and the variability is increased over simulated studies.

This variation in ranking is obviously troublesome in a real microarray study, since a lab or organization typically only has sufficient resources to follow up a small subset of features in subsequent studies. Errors in the ranking may result in wasted effort following up genes that appear to show high-differential expression across groups, but in fact are highly ranked only due to noise. Even ignoring the limitations imposed by finite resources, variability in the ranking of features changes the global biological picture derived from the microarray analysis. Bioinformaticists typically use the top ranked features to map the pathways and networks involved in the disease or drug

response being studied using tools like Ingenuity Pathways Analysis [44]. Variation in the ranking of genes for differential expression alters the set of significant genes and hence alters the pathways inferred to be involved in the etiology of the disease or response.

Noise dependence also has an impact on the second step of a multiple testing procedure: estimating error rates. This difficulty stems primarily from variations in the distribution of the null P-values from study to study, where the null P-values refer to the P-values for the genes whose coefficient $\beta_i = 0$. Most theoretical arguments regarding error rate estimation require the null P-values to be uniformly distributed both for any specific microarray study and across repeated studies.

Figure 2.3(a-d) shows histograms of the null P-values for four of the simulated studies under independent noise, which appear to be uniformly distributed. Figure 2.3(e-h) shows the null P-values for the same four studies simulated under noise dependence. Here the null P-values show strong variation, ranging from conservative to anti-consertive. The reason for this variation in null P-values is confounding between the primary variable $\mathbf{y}$ and the unmodeled factor $\mathbf{z}$. When $\mathbf{y}$ and $\mathbf{z}$ are highly correlated in a specific simulated study by chance, some of the signal due to $\mathbf{z}$ is captured in the differential expression statistics for $\mathbf{y}$. Conversely, when $\mathbf{z}$ and $\mathbf{y}$ are nearly orthogonal, a consistent pattern of noise is added to the data for many genes, which pushes the P-values away from zero.

Variation in the null P-values across studies also results in ill-behaved estimates of the false discovery rate (FDR) and the estimated proportion of null hypotheses, two common error measures considered for high-throughput studies [66, 48]. One estimator of the false discovery rate that is commonly used in practice is the Q-value [79, 84]. The Q-value can be estimated directly from the P-values or test statistics calculated from the first step of a MTP. If we call all genes significant with a Q-value less than α then we expect approximately $100 \times \alpha\%$ false discoveries among the significant genes.

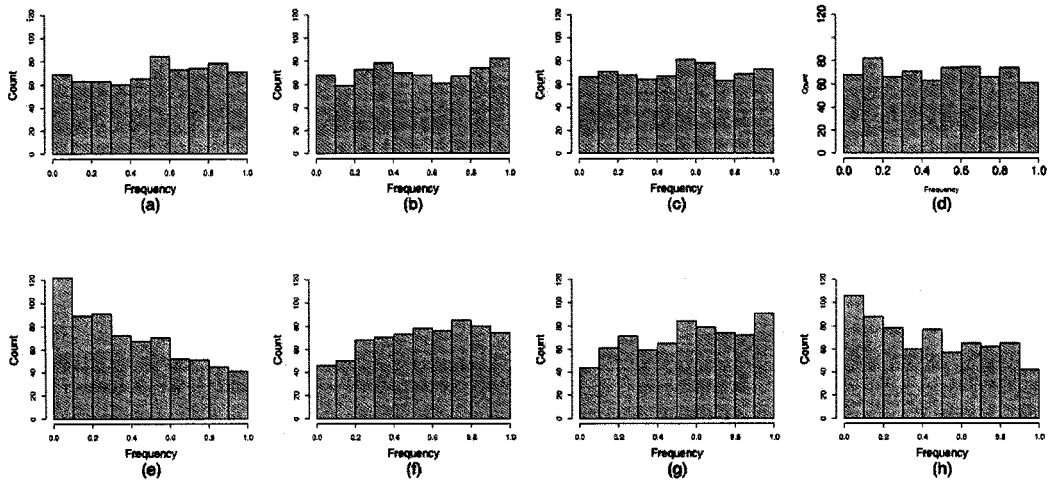

Figure 2.3: Histograms of the null P-values, corresponding to genes 300-1000 in the simulated study for four realizations of the (a-d) independent data and (e-h) noise dependent data, where the absolute value of the correlation between the primary variable and the unmodeled factor was 0.40, 0.10, 0.10, and 0.31 in histograms (e-h), respectively.

Figure 2.4 plots the estimated Q-value versus the true FDR for the one hundred simulated microarray studies. Each grey line represents the Q-value estimates from a single simulated study. In both the independent and noise dependent simulations, the Q-value estimates seem to be conservative on average (i.e., there are more lines above the line of identity than below it). However, there is substantially less variability among the Q-value estimates from the independent error simulation. This is significant, because in practice we only observe one set of Q-values, corresponding to one line in Figure 2.4 and we do not know the true false discovery rate. Thus, when there is noise dependence we may often greatly overestimate or underestimate the false discovery rate.

Another global measure of significance from a microarray study is the proportion of truly null hypotheses, π_0. For each of our simulated studies we have fixed $\pi_0 = 0.7$, since 300 of the 1,000 genes are differentially expressed across groups. There exists

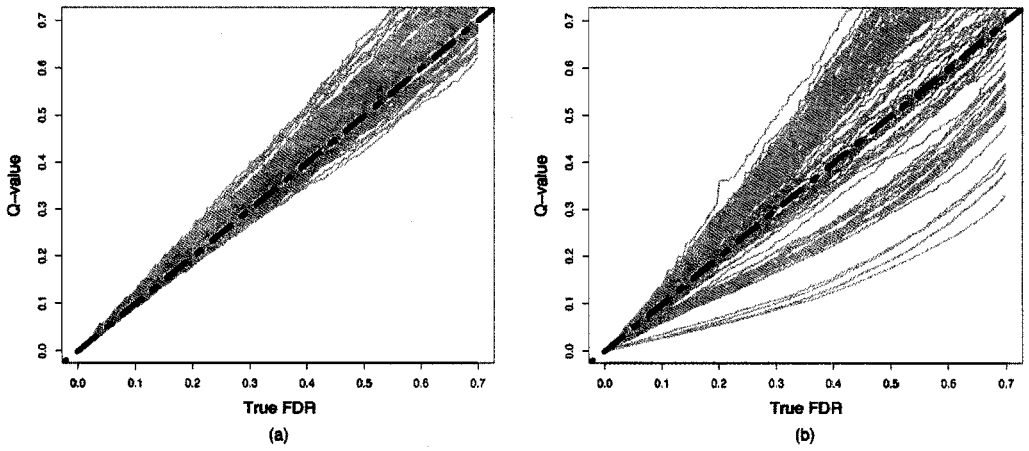

Figure 2.4: A plot of the estimated Q-value versus the true FDR for the (a) independent data and (b) noise dependent data. Each grey line represents the estimates from a single study, and the blue dotted line is the line of equality

a conservatively biased estimator, $\widehat{\pi}_0$ for π_0, based on the global distribution of P-values for a single study [79]. When the errors across features are independent $\widehat{\pi}_0$ should be approximately Normally distributed with mean slightly larger than π_0 and ideally with small variance. Figure 2.5(a) shows the distribution of this estimator for the simulated studies with independent errors. As expected, the distribution is centered on a value slightly larger than 0.7 and has relatively small variance. When noise dependence is added to the simulation in Figure 2.5(b), the estimates of π_0 are much more variable. The estimated proportion of null hypotheses ranges from 0.33 to 1.00 for the noise dependent data; these estimates would support completely different conclusions based on the same simulated experiment.

2.3 Interpretation and Problem Statement

We have used the results of 100 replicated studies drawn from the same hypothetical population to show that noise dependence, (1) affects the ranking of the differentially

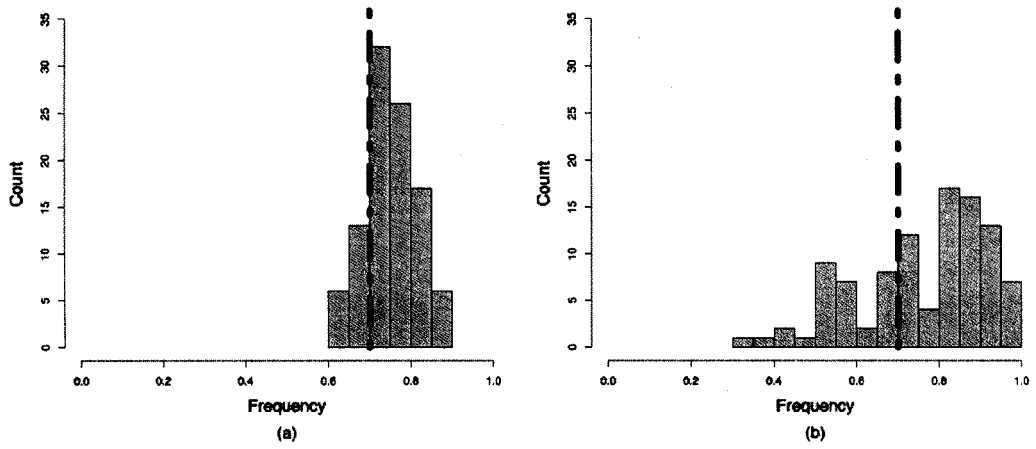

Figure 2.5: Histograms of the estimated proportion of truly null hypothesis tests $\widehat{\pi}_0$ over 100 simulated microarray studies for the (a) independent data and (b) noise dependent data. The vertical dotted blue line indicates the true value of π_0.

expressed genes and (2) alters the distribution of the P-values for the null hypothesis tests from study to study. In practice, a gene expression study is performed once and the true ranking and null distribution are not known. So what is considered variability across repeated studies in this experiment is in fact error in the results of any single study. In our simulated randomized microarray experiment, the FDR and π_0 estimates are conservatively biased when averaged over all 100 simulated studies for both independent and noise dependent errors. However, if we consider any particular study, this is not the case. For instance, one simulated study under noise dependence estimated the proportion of true nulls as 0.33. A histogram of the P-values for all of the genes from this study appears in Figure 2.6.

The shape of this P-value distribution is exactly what we would expect under independence of errors. A subset of the P-values is clustered around zero, and appear to be highly significant for differential expression. The remaining P-values are distributed fairly uniformly between zero and one and appear to be from the hypothe-

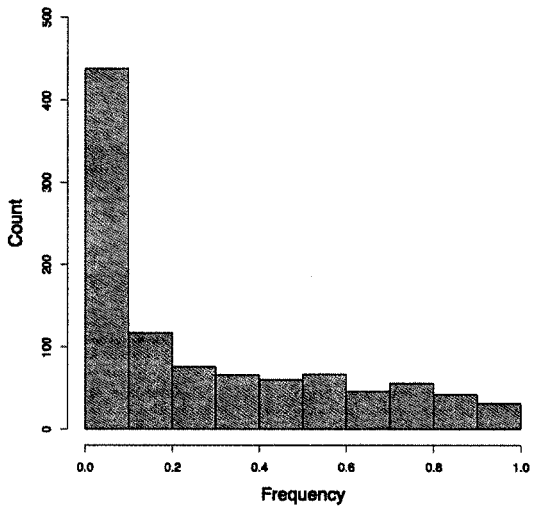

Figure 2.6: A histogram of the P-values from a single simulated microarray study under noise dependence. In this simulated study the absolute value of the correlation between the primary variable and the unmodeled factor is 0.52.

sized null distribution. Yet, in this case, the signal (the source of the small P-values) is due to the unmodeled factor, $\mathbf{z}$, which causes noise dependence between genes. The signal due to the unmodeled factor results in both biased ranking and error rate estimation for this specific microarray study. It is our goal in this dissertation to develop methods that account for sources of noise dependence in specific realizations of a microarray study. As a result this approach should both reduce the variability across independent realizations of a study and reduce bias in any specific study.

Chapter 3

A FRAMEWORK FOR NOISE DEPENDENCE

The goal of this chapter is to provide a general statistical framework for noise dependence in high-dimensional experiments. In subsequent chapters, we will apply this framework to create estimators that account for noise dependence. Before examining the properties of complicated data from a high-throughput experiment, consider the following simple motivating example: suppose we observe $X_1, \ldots, X_n$ such that:

$$\begin{aligned} \mathbf{E}[X_i] &= 0 \\ \mathbf{Var}[X_i] &= \sigma^2 + \tau^2 \\ \mathbf{Cov}[X_i, X_j] &\sim \tau^2 \end{aligned}$$

In this example we can decompose each X_i into an observation specific component e_i and a common component, Z. In other words we can always write:

$$\begin{aligned} X_i &= Y + e_i \\ Z &\sim \mathrm{N}(0, \tau^2) \\ e_i &\sim \mathrm{N}(0, \sigma^2) \end{aligned}$$

It is easy to see that $\mathbf{Var}[X_i] = \sigma^2 + \tau^2$ and $\mathbf{Cov}[X_i, X_j] = \mathbf{Var}[Z] = \sigma^2$, but $\mathbf{E}[X_i|Z] = Z$ and $\mathbf{Cov}[X_i, X_j|Z] = 0$. In other words, by conditioning on the common component, Z, the random variables $X_i|Z$ are mutually independent.

Now suppose that Z is not observed and we instead condition on the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. In this case, the covariance between any pair of X_i is $\mathbf{Cov}[X_i, X_j|\bar{X}] = -\frac{\tau^2}{n}$. Thus as $n \to \infty$ we have $\mathbf{Cov}[X_i, X_j|\bar{X}] \to 0$. Asymptotically the conditional random variables $X_i|\bar{X}$ are mutually independent. Even more generally, suppose we condition on $\hat{X}$, any asymptotically consistent estimator of Z. Then

as n goes to infinity, application of the dominated convergence theorem results in: $\mathbf{Cov}[X_i, X_j|\widehat{X}] \rightarrow \mathbf{Cov}[X_i, X_j|Z] = 0$. So the observations $X_i|\widehat{X}$ are asymptotically mutually independent as long as $\widehat{X}$ is consistent for the common factor Z.

Our framework for noise dependence in high-throughput experiments is an extension of this idea to more complicated data types. We seek to identify low dimensional variables that account for the dependence in high-dimensional data and estimates of those variables that show good properties as the number of features in a study grows large. In the next section we show that for Normally distributed high-dimensional data it is always possible to find a low dimensional set of parameters that completely capture the dependence structure across features. We then broaden our framework to dependence for high-dimensional data regardless of the underlying distribution of the common factors and the error terms.

3.1 Normally Distributed High-Throughput Data

Recall that for a general high-throughput experiment we observe feature level data $\mathbf{X}$ and primary variables $\mathbf{Y}$. For simplicity we will consider the case where only a single primary variable $\mathbf{y}$ is considered, but the extensions to multiple $\mathbf{Y}$ is trivial. We assume that each array represents an independent sample from the population, and thus the j^{th} column of the data matrix $\mathbf{x}_{\cdot j}$ is independent of the other columns. Suppose the data for feature i and sample j follow model 1.1. Since $f(\cdot)$ almost always parameterizes an additive or linear model in high-dimensional data analysis, we will write:

$$f(\mathbf{y}_j) = \sum_{k=1}^{d} \beta_{ik} s_k(\mathbf{y}_j)$$

Our model for the data point x_{ij} is then given by the following expression.

$$x_{ij} = \sum_{k=1}^{d} \beta_{ik} s_k(\mathbf{y}_j) + e_{ij}$$

Often it will be more convenient to work with the matrix form of this model.

$$\mathbf{X} = \boldsymbol{\beta}\mathbf{S}^T + \mathbf{E} \tag{3.1}$$

Here, $\boldsymbol{\beta}$ is an $m \times d$ matrix with element (i, j) equal to β_{ij}; $\mathbf{S}$ is an $n \times d$ matrix with element (j, k) equal to $s_k(\mathbf{y}_j)$; and $\mathbf{E}$ is an $m \times n$ matrix with element (i, j) equal to e_{ij}.

As a first step, suppose that the elements of $\mathbf{E}$ are Normally distributed, where each column of $\mathbf{E}$ is assumed to be independent and the covariance across features is given by the $m \times m$ matrix, $\boldsymbol{\Sigma}$. Then it is possible to calculate maximum likelihood estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ [58] and perform a multiple testing procedure based on the likelihood ratio test for the hypotheses:

$$H_{0i} : \boldsymbol{\beta}_{\cdot i} = \mathbf{0} \quad \text{vs.} \quad H_{1i} : \boldsymbol{\beta}_{\cdot i} \neq \mathbf{0}$$

Although this approach is direct and the corresponding estimates and null distributions have been derived [58], estimating $\boldsymbol{\Sigma}$ well is extremely difficult. For instance, the maximum likelihood estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \mathbf{X}\mathbf{S}\left(\mathbf{S}^T\mathbf{S}\right)^{-1} \\ \widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\left(\mathbf{X} - \widehat{\boldsymbol{\beta}}\mathbf{S}^T\right)\left(\mathbf{X} - \widehat{\boldsymbol{\beta}}\mathbf{S}^T\right)^T \end{aligned}$$

The matrix $\widehat{\boldsymbol{\Sigma}}$ has m^2 elements which are estimated on the basis of n observations. Since m is typically on the order of tens or hundreds of thousands, the number of parameters to estimate is on the order of hundreds of millions or billions. Meanwhile, the number of samples is usually no more than several hundred. This is an example of the so-called curse of dimensionality, where the number of parameters in high-dimensional models grows much faster than the number of samples. Although the task is daunting for large numbers of features, shrinkage estimates for $\boldsymbol{\Sigma}$ have been proposed that behave better than the maximum likelihood estimate, however these approaches still require specification of a restrictive target covariance matrix and heavy computational effort [72].

A second approach for addressing noise dependence is to decompose the variability in $\mathbf{E}$ into a dependent and an independent component. For multivariate Normal random variables a very general decomposition exists as shown in Lemma 1.

Lemma 1. *Let* $\mathbf{e} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$, *where* $\boldsymbol{\Sigma}$ *is a positive definite matrix and* $\mathbf{C} = \text{diag}\{\sigma_{11}, \ldots, \sigma_{\text{mm}}\}$. *Then there exists a matrix* $\mathbf{A}$ *of constants, a constant* λ_0 *and random vectors* $\mathbf{z} \sim \text{MVN}(\mathbf{0}, \mathbf{I})$ *and* $\mathbf{u} \sim \text{MVN}(\mathbf{0}, \lambda_0 \mathbf{C})$ *such that:*

$$\mathbf{e}^* = \mathbf{A}\mathbf{z} + \mathbf{u} \tag{3.2}$$

where $\mathbf{e}^*$ *and* $\mathbf{e}$ *have the same distribution and* $\mathbf{A}$ *has rank* k *with* $0 \leq k < m$

Proof. Since $\boldsymbol{\Sigma}$ is positive definite and symmetric, all of the diagonal elements of $\boldsymbol{\Sigma}$ must be positive, so $\mathbf{C}$ is positive definite and $\mathbf{C}^{-1}$ exists and is positive definite. But $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}^{-1}\boldsymbol{\Sigma} = \mathbf{C}\mathbf{S}$, where $\mathbf{S}$ is positive definite. Let $\lambda_0 > 0$ be the smallest eigenvalue of $\mathbf{S}$. Then $\mathbf{S} = \mathbf{S} - \lambda_0\mathbf{I} + \lambda_0\mathbf{I}$ and the matrix $\mathbf{S}^* = \mathbf{S} - \lambda_0\mathbf{I}$ is non-negative definite. Applying the spectral theorem we can write $\boldsymbol{\Sigma} = \mathbf{C}(\mathbf{S}^* + \lambda_0\mathbf{I}) = \mathbf{L}^T\mathbf{L} + \lambda_0\mathbf{C}$, where $0 \leq \text{rank}(\mathbf{L}) < \text{m}$. Setting $\mathbf{A} = \mathbf{L}^T$, we have $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T + \lambda_0\mathbf{C}$. Using properties of the Normal distribution $\mathbf{E}[\mathbf{e}^*] = 0$ and $\mathbf{Var}[\mathbf{e}^*] = \mathbf{A}\mathbf{A}^T + \lambda_0\mathbf{C}$, so $\mathbf{e}^* \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$ as required. □

We can apply Lemma 1 to model 3.1 to obtain the following.

$$\mathbf{x}_{\cdot j} = \boldsymbol{\beta}\mathbf{S}_{j\cdot}^T + \mathbf{A}\mathbf{z}_j + \mathbf{u}_j \tag{3.3}$$

In this model, $\mathbf{u}_j \perp \mathbf{z}_j$ and the elements of both vectors are mutually independent. The vector $\mathbf{u}_j$ captures the variance of the features, while the linear term $\mathbf{A}\mathbf{z}_j$ captures the covariance or in the case of Normal data, dependence, across features. If we let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$, then equation 3.3 can be written in matrix form.

$$\mathbf{X} = \boldsymbol{\beta}\mathbf{S}^T + \mathbf{A}\mathbf{Z} + \mathbf{U} \tag{3.4}$$

The advantage of this decomposition is that all of the noise dependence across features is quantified by the term $\mathbf{AZ}$. If $\mathbf{Z}$ were known it could be included in the model, the dependence across features would be eliminated as shown in Lemma 2.

Lemma 2. *Suppose the data from a high-throughput experiment are distributed according to model 3.1 and* $\mathbf{Z}$ *is known. Then the residuals from fitting model 3.3 are noise-independent.*

Proof. Define,

$$\mathbf{\Psi} = (\boldsymbol{\beta} \quad \mathbf{A}) \qquad \mathbf{W} = (\mathbf{S} \quad \mathbf{Z}^T)$$

so model 3.3 can be written as

$$\mathbf{X} = \mathbf{\Psi}\mathbf{W}^T + \mathbf{U}$$

with corresponding projection matrix:

$$\mathbf{P}_w = \mathbf{I} - \mathbf{W}\left(\mathbf{W}^T\mathbf{W}\right)^- \mathbf{W}^T$$

where $\left(\mathbf{W}^T\mathbf{W}\right)^-$ is a generalized inverse of $\mathbf{W}^T\mathbf{W}$. Then we can write the residuals:

$$\begin{aligned} \widehat{\mathbf{U}} &= \mathbf{X}\mathbf{P}_w \\ &= (\mathbf{\Psi}\mathbf{W}^T + \mathbf{U})\mathbf{P}_w \\ &= \mathbf{U}\mathbf{P}_w \end{aligned}$$

Thus, each row of $\widehat{\mathbf{U}}$ is a linear combination of elements of the corresponding rows of $\mathbf{U}$, which are independent by assumption. Therefore the rows of the residual matrix are noise-independent as required. □

Estimating the $m \times n$ $\mathbf{Z}$ is an improvement over estimating the $m \times m$ matrix $\mathbf{\Sigma}$. Yet, there is still one parameter to estimate for each data point in the data set, so standard estimation of $\mathbf{Z}$ is still not usually feasible. A key observation is that the $m \times n$ matrix, $\mathbf{Z}$, can be decomposed into a $m \times r$ matrix, $\mathbf{\Gamma}$, and a $r \times n$ matrix,

$\mathbf{G}^T$; in other words, $\mathbf{Z} = \mathbf{\Gamma}\mathbf{G}^T$. The dimension r is the rank of the matrix $\mathbf{Z}$, where $0 < r \leq n$ for any non-trivial $\mathbf{Z}$. Theorem 1 shows that if $\mathbf{G}^T$ is known and is included in model 3.1, then the residuals are still noise independent.

Theorem 1. *Suppose the data from a high-throughput experiment are distributed according to model 3.1 and* $\mathbf{G}^T$ *is known. Then the residuals from fitting the following model are noise independent.*

$$\mathbf{X} = \boldsymbol{\beta}\mathbf{S}^T + \mathbf{\Gamma}\mathbf{G}^T + \mathbf{U}^* \tag{3.5}$$

Proof. Define:

$$\mathbf{\Psi}_g = (\boldsymbol{\beta} \quad \mathbf{\Gamma}) \qquad \mathbf{W}_g = (\mathbf{S} \quad \mathbf{G})$$

so model 3.5 can be written as

$$\mathbf{X} = \mathbf{\Psi}_g \mathbf{W}_g^T + \mathbf{U}$$

with corresponding projection matrix:

$$\mathbf{P}_{w_g} = \mathbf{I} - \mathbf{W}_g \left(\mathbf{W}_g^T \mathbf{W}_g\right)^- \mathbf{W}_g^T$$

However $\mathbf{Z}$ and $\mathbf{G}$ have the same column space, so the corresponding projection matrices, $\mathbf{P}_{w_g}$ and $\mathbf{P}_w$ are equal. Following the argument of Lemma 2, the residuals are noise-independent. □

Surprisingly, knowing a specific $r \times n$ matrix where $r \leq n$ is sufficient to eliminate the noise dependence in Normally distributed high-dimensional data of dimension $m \times n$. We call the columns of $\mathbf{G}$ a set of surrogate variables for the data $\mathbf{X}$ since they act as a surrogate for the matrix $\mathbf{Z}$, which accounts for the dependence across features.

Definition 2. Ideal Surrogate Variables *The columns of any matrix* $\mathbf{G}$ *are a set of ideal surrogate variables for the high-dimensional data* $\mathbf{X}$*, if the following equality*

holds:

$$\begin{aligned} \mathbf{X} &= \boldsymbol{\beta}\mathbf{S}^T + \mathbf{E} \\ &= \boldsymbol{\beta}\mathbf{S}^T + \boldsymbol{\Gamma}\mathbf{G}^T + \mathbf{U} \end{aligned}$$

where the elements of $\mathbf{U}$ *are mutually independent.*

It is clear that the surrogate variables are not unique; any matrix $\mathbf{G}$ with the same column space as $\mathbf{Z}$ is a set of surrogate variables for the data from model 3.1. Since our goal is inference on $\boldsymbol{\beta}$, it is not important to identify the true underlying factors causing noise dependence, we can estimate any linear combination of those factors spanning the same column space. This is an important point, since it allows us to chose the form of the surrogate variables that is most consistent for estimation.

Regardless of the choice of $\mathbf{G}$, to estimate the surrogate variables we must estimate at most n^2 parameters, which represents a substantial improvement over estimating the m^2 parameters of the matrix $\boldsymbol{\Sigma}$. This is a rare example of the curse of dimensionality working in favor of parameter estimation. The fact that $n \ll m$ means that it is often possible to borrow information across features to get very accurate estimates of $\mathbf{G}$.

We have shown that for Normal data, dependence across features can always be modeled with a low-dimensional matrix. Obviously, it is not always the case that the feature level data from a high-throughput experiment are Normally distributed. Next we turn our attention to a more general result, analogous to Theorem 1 when different factors and different error distributions are considered. In the next section, it is shown that such an analog exists for quite general assumptions and conditions are proposed for accurate estimation of $\boldsymbol{\beta}$.

3.2 Extensions of the Framework

In the previous section, we developed a general model for dependence when the unmodeled factors and independent noise terms were Normally distributed. Here we

extend the model to the more general case where:

$$\mathbf{X} = \boldsymbol{\beta}\mathbf{S}^T + \mathbf{g}(\mathbf{K}) + \mathbf{U} \tag{3.6}$$

In other words, we no longer require the complete residuals $\mathbf{E} = \mathbf{g}(\mathbf{K}) + \mathbf{U}$ to be Normal, nor do we require $\mathbf{g}(\mathbf{K})$ or $\mathbf{U}$ to be Normal. All we require is that $\mathbf{E}$ can be partitioned into additive parts, $\mathbf{g}(\mathbf{K}) + \mathbf{U}$, where the rows of $\mathbf{U}$ are independent.

Model 3.5 can be motivated from a more biological perspective, regardless of the distributional assumptions for $\mathbf{U}$ and $\mathbf{g}$. Suppose that in addition to the primary variables $\mathbf{Y}$ there are p other factors that affect the feature level data $\mathbf{X}$. Let $\mathbf{K}$ be the $p \times n$ matrix where the data for factor j constitutes the j^{th} row of the matrix $\mathbf{K}$. Examples of these other factors in a human expression study could be environmental or genetic factors that influence the expression of thousands of genes. Then let $\mathbf{g}(\cdot) : \mathbb{R}^{p \times n} \rightarrow \mathbb{R}^{m \times n}$ be the possibly non-linear function of the random variables $\mathbf{K}$ describing their effect on the feature level data. The feature level data then take the form of equation 3.6, where the rows of the matrix $\mathbf{U}$ are assumed to be noise-independent, since the influence of any common variation is parameterized within $\mathbf{g}(\mathbf{K})$. Theorem 2 states that if model 3.6 holds, then again, knowing an $n \times n$ matrix $\mathbf{G}$ is sufficient to render the residuals noise-independent.

Theorem 2. *Suppose the data from a high-throughput experiment are distributed according to 3.6. Then there exists an $r \times n$ matrix $\mathbf{G}$ such that the residuals from fitting model 3.5 are noise-independent.*

Proof. Since $\mathbf{g}(\mathbf{K})$ is a matrix of dimension $m \times n$ with $n < m$, there exist a $m \times g$ matrix, $\boldsymbol{\Gamma}$, and a $r \times n$ matrix, $\mathbf{G}^T$ where r is the rank of $\mathbf{g}(\mathbf{K})$ such that:

$$\mathbf{g}(\mathbf{K}) = \boldsymbol{\Gamma}\mathbf{G}^T$$

Following the argument of Theorem 1, since the rows of $\mathbf{U}$ are noise-independent, including $\mathbf{G}^T$ in the model relating the features to the primary variables results in noise-independent residuals. □

The results of Theorem 2 are quite general, since it encompasses situation where the dependence term can be modeled additively with the primary variable. Furthermore, the matrix $\mathbf{G}$ may be estimable for quite flexible dependence structures, since the number of features is large. However, fitting model 3.5 and obtaining unique estimators requires more restrictive assumptions in practice. The basic idea of Lemma 3 is that unique estimates exist only when the total number of degrees of freedom used for fitting model 3.5 does not exceed the number of independent samples, n.

Lemma 3. *Suppose model 3.5 or 3.6 holds, where (1)* $\mathbf{S}$ *and* $\mathbf{G}$ *are known matrices of rank* d *and* r*, respectively, (2) the columns of* $\mathbf{W} = (\mathbf{S} \quad \mathbf{G})$ *are linearly independent and (3) the expectation of the feature specific noise is zero. Then* $\widehat{\boldsymbol{\beta}}$ *is unbiased if and only if* $(r + d) \leq n$.

Proof. Estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$ are given by the equation:

$$(\widehat{\boldsymbol{\beta}} \quad \widehat{\boldsymbol{\Gamma}}) = \mathbf{X}\mathbf{W}\left(\mathbf{W}^T\mathbf{W}\right)^-$$

Standard linear model theory guarantees the estimates are unbiased if and only if the generalized inverse $\left(\mathbf{W}^T\mathbf{W}\right)^-$ is unique, i.e. the matrix $\mathbf{W}^T\mathbf{W}$ is invertible [76]. But $\mathbf{W}^T\mathbf{W}$ has dimension $(r + d) \times (r + d)$ and $\mathbf{W}$ has dimension $n \times (r + d)$. Therefore the rank of $\mathbf{W}^T\mathbf{W}$ is equal to the minimum of $(r + d)$ and n. So, the matrix $\mathbf{W}^T\mathbf{W}$ is invertible if and only if $(r + d) \leq n$. □

Often, the statistics used for testing associations in models 3.5 and 3.6 are based on both estimates of the standard deviation, which are in turn functions of the residuals, and estimates for the parameters $\widehat{\boldsymbol{\beta}}$. The following lemma will be useful when we consider multiple testing dependence in Chapter 7.

Lemma 4. *Suppose model 3.5 or 3.6 holds, where (1)* $\mathbf{S}$ *and* $\mathbf{G}$ *are known matrices of rank* d *and* r*, respectively, (2) the columns of* $\mathbf{W} = (\mathbf{S} \quad \mathbf{G})$ *are linearly independent, (3) the expectation of the feature specific noise is zero and (4)* $(r + d) \leq n$*. Then the estimates of* $\beta_{i\cdot}$ *are independent across features where* $\beta_{i\cdot} = \mathbf{0}$.

Proof. The model for feature i is as follows:

$$\mathbf{x}_{i\cdot} = \boldsymbol{\beta}_{i\cdot}\mathbf{S}^T + \boldsymbol{\gamma}_{i\cdot}\mathbf{G}^T + \mathbf{u}_{i\cdot}$$

with corresponding estimate:

$$\widehat{\boldsymbol{\beta}_{i\cdot}} = (\mathbf{x}_{i\cdot} - \widehat{\boldsymbol{\gamma}}_{i\cdot}^{\perp}\mathbf{G}^T)\mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}$$

where:

$$\begin{aligned}\widehat{\boldsymbol{\gamma}}_{i\cdot}^{\perp} &= \mathbf{x}_{i\cdot}\mathbf{R}\mathbf{G}\left(\mathbf{G}^T\mathbf{R}\mathbf{G}\right) \\ \mathbf{R} &= \mathbf{I} - \mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\end{aligned}$$

So if $\boldsymbol{\beta}_{i\cdot} = \mathbf{0}$ then:

$$\begin{aligned}\widehat{\boldsymbol{\beta}_{i\cdot}} &= (\boldsymbol{\gamma}_{i\cdot}\mathbf{G}^T + \mathbf{u}_{i\cdot} - \widehat{\boldsymbol{\gamma}}_{i\cdot}^{\perp}\mathbf{G}^T)\mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1} \\ &= \mathbf{u}_{i\cdot}(\mathbf{I} - \mathbf{R})\mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}\end{aligned}$$

Since each $\widehat{\boldsymbol{\beta}_{i\cdot}}$ is only a function of the noise independent data $\mathbf{u}_{i\cdot}$, the estimates are independent. □

3.3 Examples

We have defined a very general framework for noise-dependence in high-throughput experiments. Before moving on to discuss algorithms for estimating surrogate variables, it may be helpful to consider some concrete examples of model 3.5. Here, we provide two examples from gene expression studies, where the feature level data x_{ij} represents the relative abundance of transcript i on array j. These examples represent the two most popular gene expression study designs: static and timecourse differential expression analysis. These examples also serve to demonstrate the flexibility of our model.

3.3.1 Static Differential Expression

The simplest static differential expression study measures gene expression for m transcripts on $\frac{n}{2}$ healthy samples and $\frac{n}{2}$ control samples. The primary variable data, $\mathbf{y}$ is simply an indicator of case or control status. In addition to case control status, suppose the age of the patients $\mathbf{g}_1$ and the gender of the patients $\mathbf{g}_2$ are linearly associated with expression. Then the model for expression is:

$$\begin{aligned} \mathbf{X} &= \boldsymbol{\mu} + \boldsymbol{\beta}\mathbf{y}^T + \boldsymbol{\gamma}_1\mathbf{g}_1^T + \boldsymbol{\gamma}_2\mathbf{g}_2^T + \mathbf{U} \\ &= \boldsymbol{\mu} + \boldsymbol{\beta}\mathbf{y}^T + \boldsymbol{\Gamma}\mathbf{G} + \mathbf{U} \end{aligned}$$

which conforms to the structure of equation 3.5. If age and gender were unknown or unmodeled we could also write this model in terms of more complicated surrogate variables, such as the sum and difference of the patients age and gender: $\mathbf{h}_1 = \mathbf{g}_1 + \mathbf{g}_2$ and $\mathbf{h}_1 = \mathbf{g}_1 - \mathbf{g}_2$. Then:

$$\begin{aligned} \mathbf{X} &= \boldsymbol{\mu} + \boldsymbol{\beta}\mathbf{y}^T + \boldsymbol{\gamma}_1\mathbf{g}_1^T + \boldsymbol{\gamma}_2\mathbf{g}_2^T + \mathbf{U} \\ &= \boldsymbol{\mu} + \boldsymbol{\beta}\mathbf{y}^T + \boldsymbol{\lambda}_1\mathbf{h}_1^T + \boldsymbol{\lambda}_2\mathbf{h}_2^T + \mathbf{U} \\ &= \boldsymbol{\mu} + \boldsymbol{\beta}\mathbf{y}^T + \boldsymbol{\Lambda}\mathbf{H} + \mathbf{U} \end{aligned}$$

for an appropriately chosen set of coefficients $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$. There are an infinite number of ways that we could choose $\mathbf{h}_1$ and $\mathbf{h}_2$ without changing $\mathbf{U}$ or the coefficients for case-control status $\boldsymbol{\beta}$. Each of these distinct sets of vectors corresponds to a distinct set of surrogate variables.

3.3.2 Timecourse Differential Expression

The simplest timecourse differential expression study measures gene expression for m transcripts from a single individual at n consecutive time points. The primary variable data, $\mathbf{y}$, is a vector of time points $(t_1, \ldots, t_n)^T$ where the t_i are arranged in order from smallest to largest. One flexible model for timecourse trends models the

expression for gene i at time point j as a linear combination of K basis functions $\mathbf{s}(t_j) = (s_1(t_j), \ldots, s_K(t_j))^T$. Examples of the basis functions could be polynomials (e.g., $s_k(t_j) = t_j^k$) or natural cubic splines. In this example, suppose that there is a second quantitative variable, $\mathbf{g}$, describing the level of a nutrient in the substrate for observation j that affects expression. Assume that the g_j are not arranged in increasing or decreasing order, and that gene expression is proportional to the square of the level of nutrient. Then we can define a new variable $\mathbf{g}_2$ where $g_{2j} = g_j^2$, and we can write a model for expression as:

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\beta}\mathbf{S}^T + \boldsymbol{\gamma}\mathbf{g}_2^T + \mathbf{U}$$

where $\mathbf{S} = (\mathbf{s}(t_1), \ldots, \mathbf{s}(t_n))$. Note that this again falls into the general form of equation 3.5. In this case, there is a natural ordering of the arrays with respect to time, but the nutrient level does not have a similar ordering. Although the values of $\mathbf{g}_2$ represent a smooth function of nutrient level, the manifestation of these variables in the expression data will appear to be a non-linear function of time. Even in this complicated case, it is still possible to write down a model for noise-dependence that fits the general form of equation 3.5.

3.4 Summary of Key Ideas

In this chapter we have developed a new framework for accounting for noise dependence in high throughput experiments. Our framework is based on the idea of decomposing the dependence between features into a set of common shared factors that can be estimated from the data directly. One important point is that these unmodeled factors are random variables across repeated samples. However, for any *fixed* sample, we can treat these random variables as fixed parameters that are associated with the expression of many features simultaneously. We showed that this framework is very general and can be used to model noise dependence within a large class of

flexible models used in practice.

The key observation of this Chapter is that it is not necessary to estimate high-dimensional population level parameters, such as the covariance among all features, $\mathbf{\Sigma}$. Instead, it is sufficient to estimate the fixed values of the random factors, $\mathbf{G}$, when modeling noise dependence. One major advantage of this approach is that estimating $\mathbf{G}$ requires estimating less than n^2 parameters, while estimating $\mathbf{\Sigma}$ requires estimating m^2 parameters. To put this reduction in context, consider a microarray experiment with 1,000 genes and 20 arrays. In this case, $\mathbf{\Sigma}$ has 1,000,000 parameters to estimate, while $\mathbf{G}$ has at most 400. Furthermore, given the increasing size of high-throughput data sets, where m is now commonly on the order of hundreds of thousands or millions, it may soon be impossible computationally to routinely estimate $\mathbf{\Sigma}$.

A second key advantage to estimating unmodeled factors rather than population parameters such as $\mathbf{\Sigma}$ is that we can correct the error in specific samples. Even if $\mathbf{\Sigma}$ were known, it would only be useful for estimating the variation in the distribution of test statistics or P-values across samples. The population average information would not be useful for correcting bias in the results of any specific sample. Put another way, by estimating and accounting for the unmodeled factors in any particular sample, we actually improve inference over the case where the population level parameters are known.

Chapter 4

SURROGATE VARIABLE ESTIMATION

In Chapter 3 we defined surrogate variables as any set of variables that eliminate noise dependence when included in the model relating the feature level data to the primary variable. In this chapter we propose algorithms for estimating surrogate variables and explore their properties through the use of simulations. The first algorithm is essentially regression on principal components, a procedure that is well known in the statistical literature and has recently been proposed in a slightly modified form for genetics data [65]. The second is a modified version of the first algorithm, where the principal components are calculated after regressing out the primary variable. The third algorithm represents joint work with John Storey and was recently published in Leek and Storey (2007)[51]. The basic idea is that rather than calculating principal components averaging over the whole data set, the components are calculated on the basis of relevant subsets of features only. The fourth algorithm is an extension of the algorithm proposed by Leek and Storey that further partitions the feature data based on those features that appear to show strong association with the primary variable and those that do not.

The algorithms presented in this chapter all use principal components analysis (PCA) for identifying low dimensional matrices that capture important structure in higher-dimensional matrices [58]. Principal components are estimated via the singular value decomposition (SVD), a matrix decomposition designed to partition the variation in rows and columns of the matrix $\mathbf{X}$ into linear components that represent maximal variation.

Definition 3. Singular Value Decomposition *Any real* $m \times n$ *matrix* $\mathbf{X}$ *has a*

decomposition $\mathbf{X} = \mathbf{UDV}^T$*, called the singular value decomposition. The matrices* $\mathbf{U}$ *and* $\mathbf{V}$ *are unitary with dimensions* $m \times n$ *and* $n \times n$*, respectively. The columns of* $\mathbf{U}$ *are called the left singular vectors of* $\mathbf{X}$ *and the columns of* $\mathbf{V}$ *are called the right singular vectors of* $\mathbf{X}$*. The matrix* $\mathbf{D}$ *is diagonal with non-negative elements. The number of non-zero diagonal elements of* $\mathbf{D}$ *is equal to the rank of* $\mathbf{X}$*.*

The classical singular value decomposition (or principal components) approach has been successfully applied in several areas of genomics. For example, Alter *et al.* applied the singular value decomposition to identify significant trends in gene expression studies [5]. They showed that the right singular vectors, or eigengenes, represent trends that account for a large proportion of the variation in the expression matrix.

The surrogate variable estimates presented here are based on the right singular vectors of carefully defined transformations or subsets of the matrix of high-throughput data. Throughout this chapter, and the rest of this dissertation, the right singular vectors will be referred to as the principal components of the matrix $\mathbf{X}$. We use the singular value decomposition, rather than some other matrix decomposition (e.g., independent components analysis) as the SVD is both easy to compute and is optimal in the sense of providing the least squares solution among all bilinear fits to the data $\mathbf{X}$. Before presenting our general algorithms for estimating surrogate variables we first turn our attention to the problem of estimating the number of surrogate variables to include in any given analysis.

4.1 Estimating the Number of Components

The problem of estimating the number of significant components in PCA has received a large and thorough treatment in the statistical and econometrics literature [58]. Estimates based on the rank of the matrix $\mathbf{X}$ [53], inflection points in scree plots [16], and, in the case of normal data, approximate distribution theory for the singular

values [47] have been proposed. In the algorithms presented here, we use variations of Buja and Eyuboglu permutation algorithm for calculating the number of significant principal components [14].

When estimating the number of significant right singular vectors for the entire matrix $\mathbf{X}$, ignoring the primary variables, it is still necessary to center each row of the data before calculating the singular value decomposition or performing Algorithm 1. This can easily be accomplished by defining $\mathbf{S}$ to be a $n \times 1$ matrix of all ones in Step 1 of the algorithm.

The intuition behind the Buja and Eyuboglu algorithm is that the statistic T_k represents the proportion of variation explained by the kth right singular value of the matrix $\widehat{\mathbf{E}}$. In a completely unstructured matrix, $T_k \approx \frac{1}{n}$, while the statistics for important right singular vectors will be much larger. The null distribution corresponds to all matrices with no structure across the rows, but identical row variances to the observed matrix. By randomly permuting each individual row of the matrix $\widehat{\mathbf{E}}$ any structure across features is eliminated, and the corresponding T_k^{0b} statistics should follow the null distribution. In Chapter 7, we show that for large values of m and n Algorithm 1 produce similar results to a recently proposed algorithm that has been shown to be asymptotically consistent as m and n grow large. We also propose a second consist estimator for fixed n and large m and show that the Buja and Eyuboglu algorithm gives comparable results as m grows large. We employ Algorithm 1, in the algorithms that follow, since simulation results seem to indicate better performance for small m and n.

4.2 Algorithms for Estimating Surrogate Variables

The four algorithms presented in this section are described in order according to their complexity. The intuition behind each of the algorithms is to directly estimate surrogate variables that explain a large percentage of the consistent variation across features from the data. Regardless of the choice of algorithm, after surrogate variables

Algorithm 1 Buja and Eyuboglu (1992) A permutation algorithm for estimating the number of significant principal components of a matrix $\mathbf{X}$ when including the primary variable $\boldsymbol{\beta}\mathbf{S}^T$.

1: Form estimates $\boldsymbol{\beta}$ by fitting the model $\mathbf{X} = \boldsymbol{\beta}\mathbf{S}^T + \mathbf{E}$ and calculate the residuals matrix $\widehat{\mathbf{E}} = \mathbf{X} - \widehat{\boldsymbol{\beta}}\mathbf{S}^T$.

2: Calculate the singular value decomposition of the residual matrix $\widehat{\mathbf{E}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$.

3: Let d_ℓ be the ℓth singular value, which is the ℓth diagonal element of $\mathbf{D}$, for $\ell = 1, \ldots, n$. If $\mathbf{S}^T$ has d linearly independent rows, then the last d singular values are exactly zero and we remove them from consideration. For right singular value $k = 1, \ldots, n-d$ set the observed statistic to be:

$$T_k = \frac{d_k^2}{\sum_{\ell=1}^{n-d} d_\ell^2}$$

which is the variance in the residual matrix explained by the kth right singular vector.

4: Form a matrix $\widehat{\mathbf{E}}^*$ by permuting each row of $\widehat{\mathbf{E}}$ independently to remove any structure across rows of the matrix.

5: Fit the model $\widehat{\mathbf{E}}^* = \boldsymbol{\beta}^*\mathbf{S}^T + \mathbf{E}_0$ and calculate the residuals $\widehat{\mathbf{E}}_0 = \widehat{\mathbf{E}}^* - \widehat{\boldsymbol{\beta}}^*\mathbf{S}^T$.

5: Calculate the singular value decomposition of the centered and permuted matrix $\widehat{\mathbf{E}}_0 = \mathbf{U}_0\mathbf{D}_0\mathbf{V}_0^T$.

6: For right singular value k form a null statistic:

$$T_k^0 = \frac{d_{0k}^2}{\sum_{\ell=1}^{n-d} d_{0\ell}^2}$$

as above, where $d_{0\ell}$ is the ℓth diagonal element of $\mathbf{D}_0$.

7: Repeat steps 4-7 a total of B times to obtain null statistics T_k^{0b} for $b = 1, \ldots, B$ and $k = 1, \ldots, n-d$.

8: Compute the P-value for right singular vector k as:

$$p_k = \frac{\#\{T_k^{0b} \geq T_k; b = 1, \ldots, B\}}{B}$$

9: Estimate the number of significant surrogate variables by $\widehat{r}(\alpha) = \sum_{k=1}^{n-d} \mathbf{1}(p_k \leq \alpha)$, for a pre-specified threshold α. For ease of exposition we will often drop the explicit dependence on α and write $\widehat{r}$.

Algorithm 2 PCA An algorithm for estimating surrogate variables based on PCA.

1: Estimate the number of significant principal components, $\widehat{r}$, of $\mathbf{X}$ using Algorithm 1 with $\mathbf{S}$ set to be an $n \times 1$ vector of ones.

2: Calculate the singular value decomposition of $\mathbf{X} = \mathbf{UDV}^T$.

3: Estimate surrogate variable k by the kth column of $\mathbf{V}$, i.e. $\widehat{\mathbf{G}}_k = \mathbf{V}_{\cdot k}$ for $k = 1, \ldots, \widehat{r}$.

are estimated, they are treated as fixed independent variables in the model relating the feature level data and the primary variable.

In some sense, the simplest surrogate variable estimates are the principal components of the centered data matrix of high-throughput data. Algorithm 2 takes the singular value decomposition of the entire matrix of centered feature level data and estimates the surrogate variable k by the kth principal component.

The advantages of Algorithm 2 are simplicity and ease of calculation. On the other hand, the principal components are calculated including the signal from the primary variable. Since principal components represent directions in the data that explain a large proportion of variation, when the signal due to the primary variable is strong, the estimated surrogate variables will be biased toward that signal. In the simulations that follow, Algorithm 2 performs the worst of the four proposed algorithms, largely due to the bias in the surrogate variable estimates from the primary variable signal.

Although a number of versions of Algorithm 2 have been proposed, significant attention has focused on one variation used in the association study literature to account for population stratification [65]. For applications in whole genome association studies, signal from the primary variable is very small in relation to the population structure since there are likely to be a relatively small number of SNPs that are truly associated with the quantitative trait or disease outcome. This low level signal has a negligible impact when principal components are calculated for a matrix with thousands of features, and principal components regression can be applied with minimal

Algorithm 3 Residual PCA An algorithm for estimating surrogate variables based on residual PCA.

1: Estimate the number of significant principal components, $\widehat{r}$, of $\mathbf{X}$ using Algorithm 1 with $\mathbf{S}$ set to be the matrix of primary variables.
2: Fit the model $\mathbf{X} = \boldsymbol{\beta}\mathbf{S}^T + \mathbf{E}$ and calculate the matrix of residuals $\widehat{\mathbf{E}} = \mathbf{X} - \widehat{\boldsymbol{\beta}}\mathbf{S}^T$.
2: Calculate the singular value decomposition of the matrix of residuals $\widehat{\mathbf{E}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$.
3: Estimate surrogate variable k by the kth column of $\mathbf{V}$, i.e. $\widehat{\mathbf{G}}_k = \mathbf{V}_{\cdot k}$ for $k = 1, \ldots, \widehat{r}$.

bias.

A second, slightly more complicated, version of principal components regression addresses the problem of bias due to the primary variable. After identifying the significant principal components, the first step in Algorithm 3 is to estimate and subtract the effect due to the primary variable. The surrogate variable estimates are then calculated as the principal components of the residual feature level data.

Surrogate variable estimates calculated using Algorithm 3 are orthogonal to the primary variable. When the factors causing noise-dependence across features are, in fact, truly orthogonal to the primary variable, the estimates from Algorithm 3 perform well. Indeed, if the orthogonality assumption holds, we will show in Chapter 7 that the surrogate variable estimates from Algorithm 3 are consistent for the true surrogate variables as the number of features grows large, even in finite samples.

An obvious question is, how likely is it that the unmodeled factors causing noise dependence are orthogonal to the primary variables in a high-dimentional study? Even for a well-designed randomized study, the number of potential factors causing noise dependence is large, and it is unlikely that they will all be orthogonally configured with respect to the primary variable in any given study. However, as the number of observations or arrays grows in a randomized study, the unmodeled factors will tend to be uncorrelated with the primary variables.

In certain randomized studies with a large number of observations Algorithm 3 may be sufficient for estimating surrogate variables. However, often there will be some level of correlation between the unmodeled factors and the primary variables. We therefore seek an approach that allows for correlation between the primary variable and the unmodeled factor, and at the same time reduces potential bias. One such approach is to reduce the data matrix to carefully selected subsets of features before performing the principal component analysis.

Algorithm 4 combines parts of the two previous algorithms to calculate surrogate variable estimates based on subsets of the feature level data. First, signatures of the surrogate variables are identified in the residuals of the feature data after subtracting the effect of the primary variable. The features most highly associated with each residual surrogate variable are identified. Then principal components analysis is performed on the original data for the subset of features most highly associated with each residual surrogate variable.

Algorithm 4 combines the advantages of the two previous algorithms. First, the signatures of the unmodeled factors are identified in the residuals, so the selection of the subset of features for each surrogate variable is not affected by signal from the primary variable. Each subset is enriched for features showing strong association with the corresponding surrogate variable. The maximal source of variation within that subset is then likely to be the surrogate variable of interest. By calculating the PCA on the original data for that subset of features, Algorithm 4 also allows for correlation with the primary variable.

In some cases, there will naturally be overlap between the features associated with the primary variable and associated with each unmodeled factor. If this overlap is strong, Algorithm 4 may, to a lesser extent, suffer from some of the same difficulties as standard regression on principal components. Namely, strong signal from the primary variable may alter the principal components of each subset and bias the corresponding surrogate variables.

Algorithm 4 Subset PCA An algorithm for estimating surrogate variables based on subset PCA.

1: Estimate the number of significant principal components, $\widehat{r}$, of $\mathbf{X}$ using Algorithm 1 with $\mathbf{S}$ set to be the matrix of primary variables.

2: Fit the model $\mathbf{X} = \boldsymbol{\beta}\mathbf{S}^T + \mathbf{E}$ and calculate the matrix of residuals $\widehat{\mathbf{E}} = \mathbf{X} - \widehat{\boldsymbol{\beta}}\mathbf{S}^T$.

3: Calculate the singular value decomposition of the matrix of residuals $\widehat{\mathbf{E}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Let $\mathbf{v}_k$ be the kth column of $\mathbf{V}$ (for $k = 1, \ldots, n$).

For each significant right singular variable $\mathbf{v}_k$ $k = 1, \ldots, \widehat{r}$

4: Regress $\mathbf{v}_k$ on each row, $\mathbf{x}_{i\cdot}$ of $\mathbf{X}$ and calculate a P-value testing for an association between the residual right singular vector and each feature's data.

5: Estimate of the number of features truly associated with the right singular vector by $\widehat{m}_1 = \lfloor (1 - \widehat{\pi}_0) \times m \rfloor$. Let $s_1, \ldots, s_{\widehat{m}_1}$ be the indices of the features with the $\widehat{m}_1$ smallest P-values from this test.

6: Form the $\widehat{m}_1 \times n$ reduced feature matrix $\mathbf{X}_r = (\mathbf{x}_{s_1}, \ldots, \mathbf{x}_{s_{\widehat{m}_1}})^T$. Calculate the right singular vectors of $\mathbf{X}_r$ and denote these by $\mathbf{v}_j^r$ for $j = 1, \ldots, n$.

7: Let $j^* = \arg\max_{1 \le j \le n} \mathbf{cor}(\mathbf{v}_k, \mathbf{v}_j^r)$ and set $\widehat{\mathbf{G}}_{\cdot k} = \mathbf{v}_{j^*}^r$.

Algorithm 5 Reduced Subset PCA An algorithm for estimating surrogate variables based on reduced subset PCA.

1: Estimate the number of significant principal components, $\widehat{r}$, of $\mathbf{X}$ using Algorithm 1 with $\mathbf{S}$ set to be the matrix of primary variables.
2: Regress $\mathbf{S}^T$ on each row $\mathbf{x}_i$ of $\mathbf{X}$ and calculate a P-value testing for association between the primary variable and the original feature level data.
3: Let $c \in (0,1)$ be a given fixed constant, calculate $\widehat{m}_0^c = \lfloor c \times m \rfloor$ and let $s_1, \ldots, s_{\widehat{m}_0^c}$ be the indices of the features with the $\widehat{m}_0^c$ largest P-values from this test.
4: Form the $\widehat{m}_0^c \times n$ reduced feature matrix $\mathbf{X}_r^c = (\mathbf{x}_{s_1}, \ldots, \mathbf{x}_{s_{\widehat{m}_0^c}})^T$.
5: Perform Algorithm 4 on the reduced data matrix $\mathbf{X}_r^c$ to obtain surrogate variable estimates.

One possible solution to this problem is to down weight those features that show strong association with the primary variable before computing the subset principal components. A number of different weighting schemes exist, but perhaps the simplest is to calculate a P-value measuring the linear association between the original data for each feature and the primary variable. Those features with small P-values are excluded from the data set. Then subset PCA is performed on the reduced feature matrix.

Algorithm 5 removes feature data that shows a very strong association with the primary variable before calculating surrogate variables. The choice of the parameter c dictates how many features are removed. When $c = 1$ Algorithm 5 reduces to Algorithm 4. As c shrinks to zero, more and more of the highly associated features are eliminated from the calculation. A tradeoff exists between removing signal due to the primary variable and retaining enough features to calculate principal components. In practice when $\widehat{m}_0^c$ is too small, the surrogate variables can not be accurately estimated.

4.3 Simulations

It is clear from the description of the four surrogate variable algorithms that two key components determine the quality of surrogate variable estimates, (1) correlation between the primary variable and the unmodeled factors causing the noise dependence and (2) the degree of overlap in the features associated with the primary variable and the unmodeled factors. To assess the relative performance of the four algorithms we performed four simulated experiments consisting of 100 simulated microarray studies each. Just as in Chapter 2, each simulated study has 1,000 genes and 20 arrays divided into two equal groups. The experiments vary in the level of correlation and overlap they allow between the primary variable and the unmodeled factor. Each of the four estimation algorithms is applied to analyze each of the simulated microarray studies. For each experiment, the algorithms are compared with respect to ranking of the truly significant genes, null distribution behavior, conservative estimation of the proportion of true nulls, and accurate and conservative false discovery rate estimation.

4.3.1 Simulation Design

Just as in Chapter 2, the data for each feature in each simulated study follow the model:

$$\begin{aligned} x_{ij} &= \mu_i + \beta_i y_j + \gamma_i z_j + u_{ij} \\ u_{ij} &\sim \mathrm{N}(0, \sigma_{\mathrm{i}}^2) \end{aligned} \tag{4.1}$$

where the number of arrays and genes are held fixed at 20 and 1,000, respectively. The primary variable is configured as:

$$y_j = \begin{cases} 1 & j = 1, \ldots, 10 \\ 0 & j = 11, \ldots, 20 \end{cases}$$

and the population parameters μ_i, β_i, and σ_i^2 are simulated initially and then held fixed, in order to mimimic a hypothetical population. The population level parameters

are drawn from the following distributions.

$$\begin{aligned}
\mu_i &\overset{i.i.d.}{\sim} N(0,1) \\
\beta_i &\overset{i.i.d.}{\sim} \begin{cases} N(1.5,1) & i = 1,\ldots,300 \\ 0 & i = 301,\ldots,1000 \end{cases} \\
\sigma_i^2 &\overset{i.i.d.}{\sim} \text{Inverse Gamma}(10,9)
\end{aligned}$$

Each simulated data set is then created based on independent draws from the noise distribution u_{ij}. The parameters γ_i for each study are added to the other population level parameters and the distribution will vary across the four simulated experiments. We will also allow the distribution of z_j to vary from experiment to experiment, to mimic a randomized or observational study. The distributions are defined as follows for the four experiments.

Experiment 1

$$\begin{aligned}
\gamma_i &\overset{i.i.d.}{\sim} \begin{cases} N(1.5,1) & i = 201,\ldots,1000 \\ 0 & i = 1,\ldots,201 \end{cases} \\
z_j &\overset{i.i.d.}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right), \quad \text{j} = 1,\ldots,20
\end{aligned}$$

Experiment 2

$$\begin{aligned}
\gamma_i &\overset{i.i.d.}{\sim} \begin{cases} N(1.5,1) & i = 101,\ldots,600 \\ 0 & i = 1,\ldots,101 \;\&\; 601,\ldots,1000 \end{cases} \\
z_j &\overset{i.i.d.}{\sim} \text{Bernoulli}\left(\frac{1}{2}\right), \quad \text{j} = 1,\ldots,20
\end{aligned}$$

Experiment 3

$$\begin{aligned}
\gamma_i &\overset{i.i.d.}{\sim} \begin{cases} N(1.5,1) & i = 201,\ldots,1000 \\ 0 & i = 1,\ldots,201 \end{cases} \\
z_j &\overset{i.i.d.}{\sim} \begin{cases} y_j & j = 1,\ldots,7 \;\&\; 11,\ldots 17 \\ \text{Bernoulli}\left(\frac{1}{2}\right) & j = 8,\ldots 10 \;\&\; 18\ldots 20 \end{cases}
\end{aligned}$$

Experiment 4

$$\gamma_i \overset{i.i.d.}{\sim} \begin{cases} N(1.5,1) & i = 101,\ldots,600 \\ 0 & i = 1,\ldots,101\ \&\ 601,\ldots,1000 \end{cases}$$

$$z_j \overset{i.i.d.}{\sim} \begin{cases} y_j & j = 1,\ldots,7\ \&\ 11,\ldots 17 \\ \text{Bernoulli}\left(\frac{1}{2}\right) & j = 8,\ldots 10\ \&\ 18\ldots 20 \end{cases}$$

Wald statistics and P-values are calculated parametrically using linear model 3.5 where the surrogate variables are estimated with one of the four algorithms presented in the previous section. The significance threshold α for identifying surrogate variables in Algorithm 1 is fixed at 0.05 and the parameter c from Algorithm 5 is fixed at 0.80.

4.3.2 Experiment One: Randomized Factor/Low Overlap

In experiment one, we allow almost no overlap between the genes affected by the primary variable and those affected by the unmodeled factor. Furthermore, the unmodeled factor is randomized with respect to the primary variable. This is the best possible case for estimating surrogate variables, as the signal from the primary variable and the unmodeled factor are easily distinguished. Experiment one represents a carefully designed and randomized microarray experiment with the additional (and slightly unrealistic) assumption that each variable affects a distinct subset of genes.

Table 4.1 summarizes the correlation between the unmodeled factor and the estimated surrogate variable. All four algorithms perform very well, which is not surprising in light of the strong signal from the unmodeled factor. In terms of relative performance, the subset algorithms are more accurate and less variable than the whole data set algorithms. Figure 4.1 plots the average ranking by t-statistics after each of the four algorithms have been applied versus the true ranking defined by the signal to noise ratio. Each of the algorithms performs reasonably well in ranking the truly significant genes for experiment one.

Another goal of the surrogate variable estimation algorithm is to correct the dis-

Table 4.1: A table of the average (standard deviation) correlation between the unmodeled factor and the estimates from the four surrogate variable estimation algorithms across the four simulated experiments.

	Algorithm			
Experiment	PCA	Residual PCA	Subset PCA	Reduced Subset PCA
One	98.6 (1.2)	96.8 (4.4)	99.6 (0.1)	99.6 (0.1)
Two	94.5 (3.5)	96.8 (3.7)	97.7 (0.6)	99.3 (0.6)
Three	97.7 (0.8)	93.2 (6.8)	99.7 (0.1)	99.9 (0.04)
Four	92.3 (1.4)	93.8 (19.0)	93.6 (19.0)	99.8 (0.1)

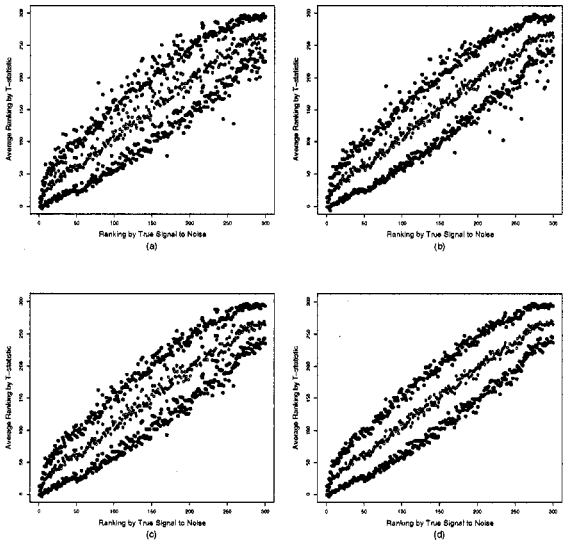

Figure 4.1: Plots of the average ranking from the adjusted t-statistics over 100 simulated studies (black dots) plus or minus one standard deviation (blue dots) for the (a) PCA, (b) Residual PCA, (c) Subset PCA and (d) Reduced Subset PCA algorithms applied to the noise-dependent data from experiment one.

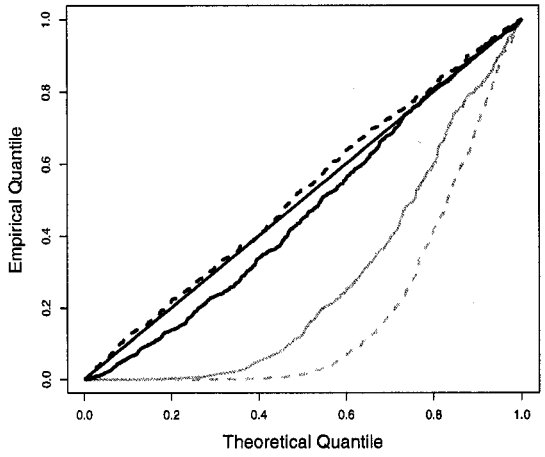

Figure 4.2: Plots of the quantiles of the KS-test P-values for each of the 100 simulated studies from experiment one versus the $\mathcal{U}(0,1)$ quantiles. The P-values are corrected by PCA (grey), Residual PCA (dashed grey), Subset PCA (solid blue), and Reduced Subset PCA (dashed blue)

tribution of the P-values under noise dependence. For each of the 100 simulated microarray studies, we applied a nested KS-test, described below, to determine if the corrected null P-values do follow the null distribution. Figure 4.2 shows a quantile-quantile plot of the KS-test P-values comparing the null P-values for each algorithm from each individual study to the uniform distribution. The subset PCA algorithms on average do a better job of correcting the null distribution than the complete data PCA algorithms.

A related goal is to correct the variation in the Q-value estimates due to noise dependence as described in Figure 2.4(b). Figure 4.3 plots the estimated Q-values versus the true false discovery rate after applying each of the surrogate variable estimation algorithms. Q-values estimated after applying the complete data PCA algorithms appear to be anti-conservatively biased consistently across studies. This is not surprising, as surrogate variable estimates from the standard PCA approach are biased

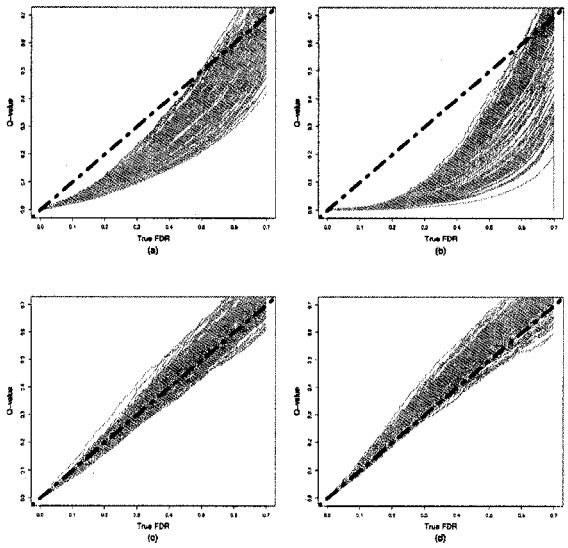

Figure 4.3: A plot of the estimated Q-value versus the true FDR after adjusting for the (a) PCA, (b) Residual PCA, (c) Subset PCA and (d) Reduced Subset PCA algorithms in experiment one. Each grey line represents the estimates from a single study, and the blue dotted line is the line of equality

by the strong signal from the primary variable, and the estimates from the residual PCA algorithm are biased whenever the unmodeled factor is correlated with the primary variable. On the other hand, the subset PCA algorithms appear to reduce variation in Q-value estimation across studies, without inducing any consistent pattern of bias. Again, this is not surprising, given that there is very little overlap in the genes affected by the primary variable and the unmodeled factor.

4.3.3 Experiment Two: Randomized Factor/High Overlap

In experiment two, the unmodeled factor is still randomized with respect to the primary variable. However, we now consider the more realistic scenario where there is more overlap in the features affected by the primary and unmodeled variable. Experiment two represents a carefully designed and randomized microarray experiment where there is strong overlap between the features affected by the primary variable

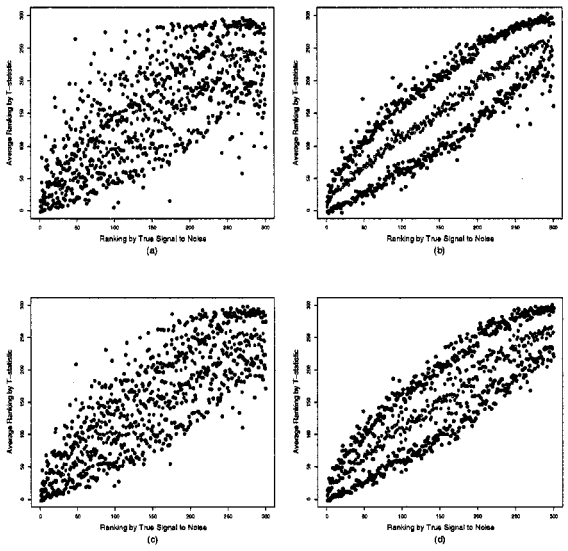

Figure 4.4: Plots of the average ranking from the adjusted t-statistics over 100 simulated studies (black dots) plus or minus one standard deviation (blue dots) for the (a) PCA, (b) Residual PCA, (c) Subset PCA and (d) Reduced Subset PCA algorithms applied to the noise-dependent data from experiment two.

and those affected by the unmodeled factor.

From Table 4.1 each algorithm again performs well in estimating the surrogate variables. However, the stronger overlap leads to increased variability in the surrogate variable estimates, particularly for the subset algorithms. Figure 4.4 plots the average ranking by t-statistics after each of the four algorithms have been applied versus the true ranking defined by the signal to noise ratio. Each algorithm performs slightly worse than for experiment one, however, the biggest difference is that the whole data set PCA and the subset PCA algorithm show more variable rankings, since they do not take into account the overlap between the primary variable and the unmodeled factor.

Although the rankings for the significant features are more variable, it is still important to correctly estimate the null distribution and the FDR even when there is strong overlap in the genes affected by the primary variable and the unmodeled factor.

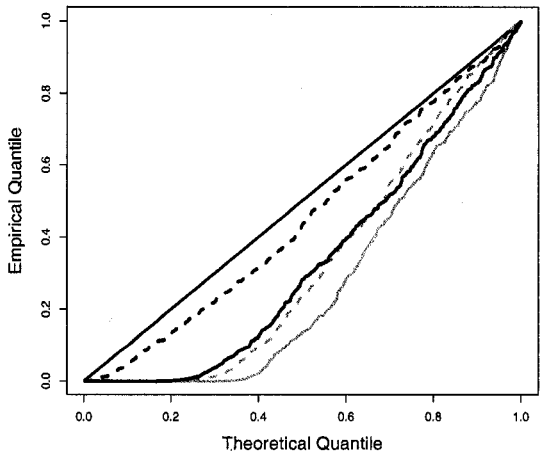

Figure 4.5: Plots of the quantiles of the KS-test P-values for each of the 100 simulated studies from experiment two versus the $\mathcal{U}(0,1)$ quantiles. The P-values are corrected by PCA (grey), Residual PCA (dashed grey), Subset PCA (solid blue), and Reduced Subset PCA (dashed blue)

Figure 4.5 shows a quantile-quantile plot of the KS-test P-values comparing the null P-values to the uniform distribution after applying each algorithm. The reduced subset PCA algorithm appears to mostly correct the bias in the null distribution, while the other approaches do not. Subset PCA, even in the case of strong overlap, still outperforms the whole data set PCA algorithms.

Figure 4.6 plots the estimated Q-values versus the true false discovery rate for each of the surrogate variable estimation algorithms. Q-values estimated after applying the complete data PCA algorithms are strongly anti-conservatively biased across all of the simulated studies. The subset PCA algorithms appear to reduce variation in Q-value estimation across studies, and the reduced subset PCA algorithm also seems to eliminate bias in the Q-value estimates. The residual bias in the subset PCA algorithm is likely due to the high overlap between the genes affected by the primary variable and the surrogate variable. Even when a subset of the genes affected by the

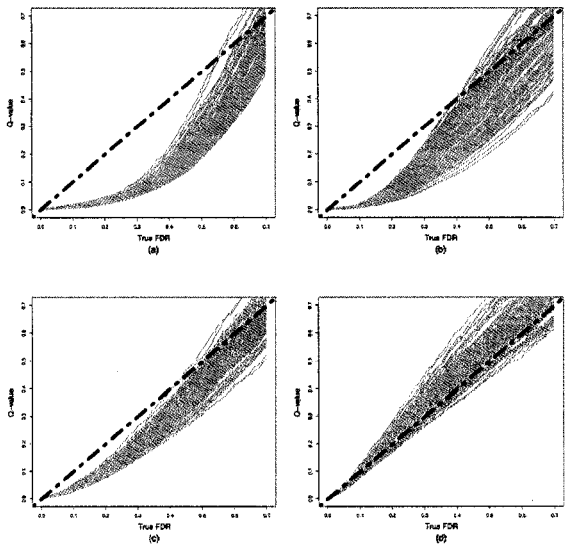

Figure 4.6: A plot of the estimated Q-value versus the true FDR after adjusting for the (a) PCA, (b) Residual PCA, (c) Subset PCA and (d) Reduced Subset PCA algorithms in experiment two. Each grey line represents the estimates from a single study, and the blue dotted line is the line of equality

surrogate variable are selected, those genes are also affected by the primary variable, and so the principal components estimates from the subset are biased.

4.3.4 Experiment Three: Non-Random Factor/LowOverlap

In experiment three, the unmodeled factor is now correlated with the primary variable across studies and we revert to the case of low overlap in the features affected by the primary and unmodeled factor. Experiment three represents an observational microarray experiment, where there are likely to be a number of factors that are correlated with the primary variable. In a traditional study, it would be impossible to elicit an estimate of the surrogate variable under these assumptions, because there is confounding between the primary variable and the unmodeled factor. However, as we will see, borrowing strength across features allows for estimation even in this type of biased study.

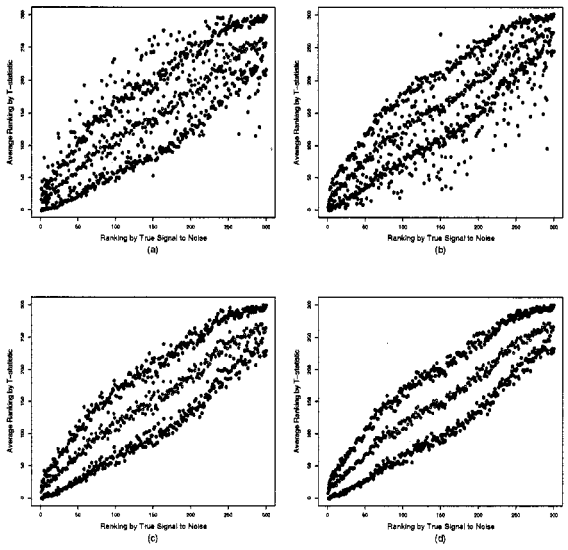

Figure 4.7: Plots of the average ranking from the adjusted t-statistics over 100 simulated studies (black dots) plus or minus one standard deviation (blue dots) for the (a) PCA, (b) Residual PCA, (c) Subset PCA and (d) Reduced Subset PCA algorithms applied to the noise-dependent data from experiment three.

The results from Table 4.1 show that confounding only seems to reduce the accuracy for the residual PCA algorithm. Since the unmodeled factor is never orthogonal to the primary variable in these simulations, this drop in performance is expected. Figure 4.7 plots the average ranking by t-statistics after each of the four algorithms have been applied versus the true ranking defined by the signal to noise ratio. All four algorithms perform reasonably well in ranking the significant statistics, particularly the PCA algorithms based on subsets. In fact, the ranking from the subset PCA algorithm for the biased experiment with low-overlap (Figure 4.7(c)) is better than the ranking from the same algorithm for the randomized study with high-overlap (Figure 4.4(c)).

Figure 4.8 shows a quantile-quantile plot of the KS-test P-values comparing the null P-values to the uniform distribution after applying each algorithm. Both subset PCA algorithms correct the null distribution entirely, even for a biased experiment.

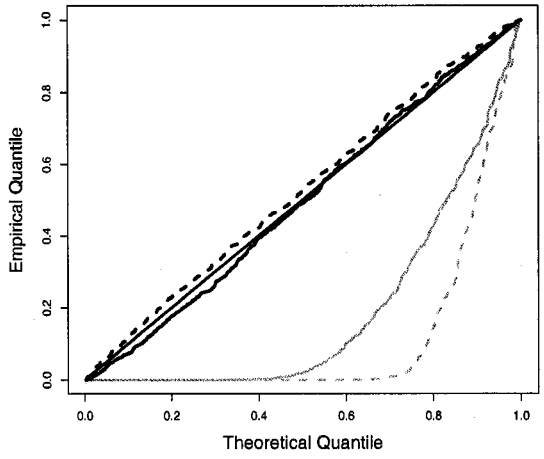

Figure 4.8: Plots of the quantiles of the KS-test P-values for each of the 100 simulated studies from experiment three versus the $\mathcal{U}(0,1)$ quantiles. The P-values are corrected by PCA (grey), Residual PCA (dashed grey), Subset PCA (solid blue), and Reduced Subset PCA (dashed blue)

Not surprisingly, the corresponding null distributions under the complete data PCA algorithms are strongly biased, especially the residual PCA algorithm that estimates surrogate variables as orthogonal to the primary variable.

Figure 4.9 plots the estimated Q-values versus the true false discovery rate for each of the surrogate variable estimation algorithms. The subset PCA algorithms reduce Q-value estimation variation across studies, and in fact result in conservatively biased Q-value estimates for nearly every individual study. Thus, Q-values estimated with subset PCA behave better, in the sense of being more conservative, than even Q-values estimated for the randomized study with the same degree of overlap. The bias in the study translates directly into strong anti-conservative bias in Q-value estimation for complete data set PCA algorithms, particularly for the residual PCA algorithm on the entire data set.

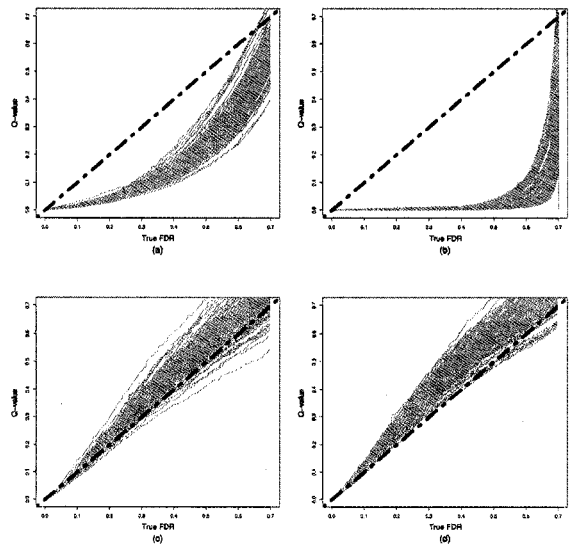

Figure 4.9: A plot of the estimated Q-value versus the true FDR after adjusting for the (a) PCA, (b) Residual PCA, (c) Subset PCA and (d) Reduced Subset PCA algorithms in experiment three. Each grey line represents the estimates from a single study, and the blue dotted line is the line of equality

4.3.5 Experiment Four: Non-Random Factor/High Overlap

Experiment four is the worst case scenario for surrogate variable estimation. The unmodeled factor is now correlated with the primary variable across studies and there is high overlap in the features affected by the primary and unmodeled factor. Experiment four again mimics an observational microarray experiment, where there are likely to be a number of factors that are correlated with the primary variable. However, in this situation the added difficulty is that there is only a small subset of genes where only the unmodeled factor acts. None of the four algorithms is designed to appropriately estimate the surrogate variables in this situation, but we include the results here to give some indication of how robust the estimates are to strong confounding and overlap.

When there is high overlap and correlation between the primary variable and the

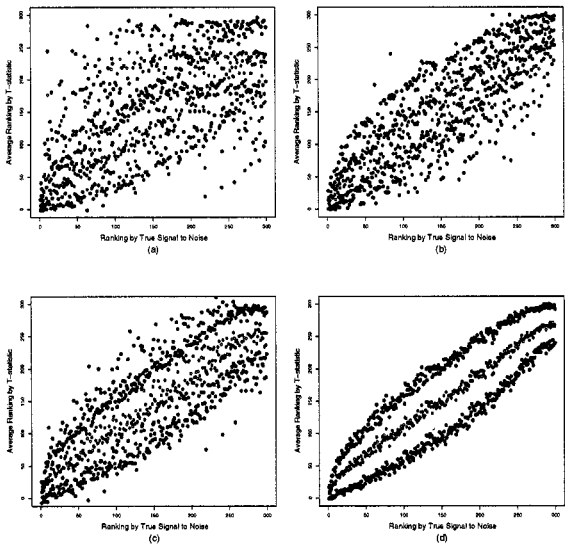

Figure 4.10: Plots of the average ranking from the adjusted t-statistics over 100 simulated studies (black dots) plus or minus one standard deviation (blue dots) for the (a) PCA, (b) Residual PCA, (c) Subset PCA and (d) Reduced Subset PCA algorithms applied to the noise-dependent data from experiment four.

unmodeled factor, surrogate variable estimation is very difficult. The results shown in table 4.5 reflect this difficulty, as the relative performance the first three algorithms is much poorer than in the previous simulations. The reduced subset PCA algorithm still estimates the the surrogate variables reasonably well, even in the face of strong bias. Figure 4.10 plots the average ranking by t-statistics after each of the four algorithms have been applied versus the true ranking defined by the signal to noise ratio. The subset PCA approaches still rank the significant features remarkably well, even in this worst case scenario.

Figure 4.11 shows a quantile-quantile plot of the KS-test P-values comparing the null P-values to the uniform distribution after applying each algorithm. With the exception of the reduced subset PCA algorithm, the null distribution, even after correcting for surrogate variables is biased.

Figure 4.12 plots the estimated Q-values versus the true false discovery rate for

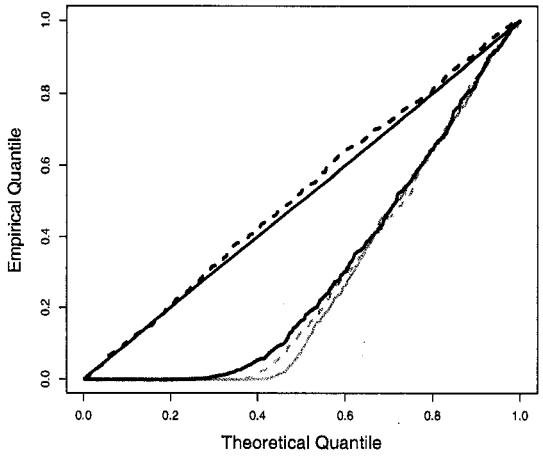

Figure 4.11: Plots of the quantiles of the KS-test P-values for each of the 100 simulated studies from experiment fiyr versus the $\mathcal{U}(0,1)$ quantiles. The P-values are corrected by PCA (grey), Residual PCA (dashed grey), Subset PCA (solid blue), and Reduced Subset PCA (dashed blue)

each of the surrogate variable estimation algorithms. There is now strong bias in the Q-value estimates for all of the algorithms with the exception of the reduced subset PCA. The subset PCA algorithm shows better behavior than the two whole data set algorithms, however the P-values from this algorithm still show bias at low values of the true FDR, which are the values most often considered in practice.

4.3.6 Simulation Experiment Summary

The simulation experiments described in this chapter indicate that including estimates of surrogate variables in the model relating expression to primary variables can substantially mitigate the effect of noise dependence. It also appears that the subset PCA algorithms proposed in this dissertation provide substantial practical improvement over the whole data set PCA Algorithms that have been proposed in practice. Surprisingly, these simple simulation results indicate that the subset PCA algorithms

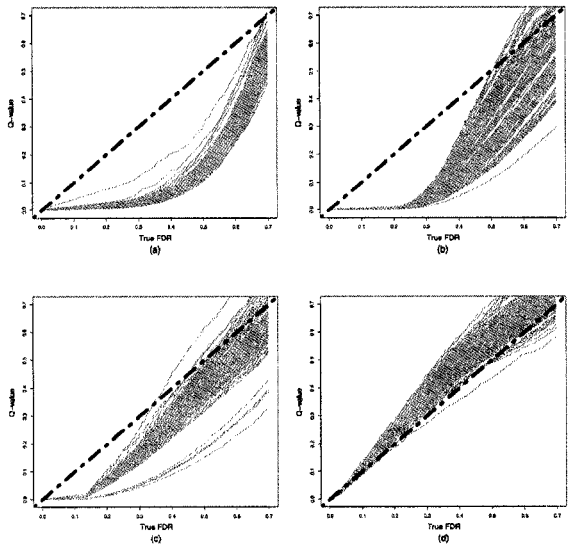

Figure 4.12: A plot of the estimated Q-value versus the true FDR after adjusting for the (a) PCA, (b) Residual PCA, (c) Subset PCA and (d) Reduced Subset PCA algorithms in experiment four. Each grey line represents the estimates from a single study, and the blue dotted line is the line of equality

can correct for noise-dependence even under biased study designs often encountered in observational studies. Put another way, by accounting for noise dependence in high-throughput data analysis, it is possible to get gene rankings and error rate estimates that are nearly as stable and unbiased as in a similar randomized study. The reduced subset PCA algorithm in particular shows promising behavior under each type of simulated experiment performed here.

For ease of exposition, our simulations have focused on the case of a single unmodeled factor causing noise dependence. However, even when multiple unmodeled factors act on the same study, each factor is influenced by the same two characteristics, (1) the degree of overlap in the subsets the factor affects and the subsets the primary variable affects and (2) the correlation with the primary variable. The results and conclusions from our simulation results can thus be easily extended to the case of multiple unmodeled factors.

4.3.7 The Nested KS-Test

The false discovery rate (FDR) has been discussed extensively and it has been pointed out that the distribution of the null P-values must be correct or conservative for FDR estimation or any other standard statistical significance measure to behave properly. What is meant for distribution of the null P-values to be correct is that they are Uniformly distributed in the interval $(0, 1)$. The null P-values have a conservative distribution if they are pushed toward one relative to the $\mathcal{U}(0, 1)$ distribution. In other words, the null P-values are conservative if their distribution is stochastically larger than the $\mathcal{U}(0, 1)$ distribution. P-values are constructed to have the Uniform distribution property under the null hypothesis, and if this cannot be done exactly the conservative version is calculated. In a simulation study where the right answer is known, there is no off-the-shelf approach to test whether the null P-values have a proper distribution.

In this dissertation, we use a Kolmogorov-Smirnov (KS) test on the set of null P-values from each study for deviation from the Uniform. However, we want to test whether this is true over many repeated simulations to avoid getting lucky on one particular simulated data set. If the set of null P-values are Uniform, then the P-value resulting from the KS test should also follow the Uniform distribution. Therefore, by examining the KS test P-values over all simulations, we can again apply a KS test to verify that these are Uniformly distributed. Here we have employed this nested KS test to compare the relative behavior of each multiple testing procedure discussed. If the quantiles of the KS test P-values follow the diagonal line in a quantile-quantile plot against the quantiles of the Uniform distribution, then this is very strong evidence that the P-values resulting from the procedure are correct.

Chapter 5

PROOF OF CONCEPT ANALYSES

We have developed a statistical framework for noise dependence in high-throughput data and proposed novel algorithms for estimating surrogate variables to account for dependence. Here we perform several proof of concept analyses in real gene expression data to illustrate the power of the surrogate variable approach in practice. We estimate and account for surrogate variables in three distinct gene expression microarray studies, where each study contains clear patterns of noise dependence across features, Figure 5.1. Recall that in a heatmap, a row represents the expression values for one transcript across the arrays, so noise dependence appears as consistent blocks of color in a column. These studies represent the major classes of microarray studies performed in practice: genetic dissection of expression variation [12, 11], differential expression analysis between disease classes [38], and differential expression over time [85].

Subset PCA is able to accurately identify and estimate the impact of unmodeled factors in each type of study, using only the expression data itself. We further show that subset PCA improves accuracy and consistency in detecting differential expression for each type of study. Accounting for surrogate variables reorders the significant gene lists to more accurately and reproducibly reflect the ordering of the genes with respect to their true differential expression signal. The surrogate variable approach is particularly useful in obtaining reproducible results in microarray studies, because adjusting for surrogate variables reduces differential expression due to sources other than the primary variables as we will see in Chapter 6. These results indicate that noise dependence is prevalent across a range of studies. Surrogate variables can be

used to capture and account for these patterns to improve the characterization of biological signal in expression analyses. The results in this chapter have partially been published in Leek and Storey (2007) [51].

5.1 *Proof of Concept: Genetics of Gene Expression in Yeast*

Several recent studies have carried out the genetic dissection of expression variation at the genome-wide level[12, 71, 60]. Brem *et al.* [12, 11] measured expression genome wide in 112 segregants of a cross between two isogenic strains of yeast. They also obtained genotypes for each segregant at markers covering 99% of the genome. The data set consists of gene expression measurements for 6,216 genes in the 112 segregants of a cross between two isogenic strains of yeast, as well as genotypes across 3,312 markers. It was shown that many gene expression traits are *cis*-linking, i.e., the quantitative trait locus (QTL) linkage peak coincided with the physical location of the open reading frame for the expression trait [98]. At the same time, it was also shown that a number of gene expression traits are *trans*-linking, with linkage peaks at loci distant from the physical location of their open reading frames. In particular, several pivotal loci each appear to influence the expression of hundreds or even thousands of gene expression traits. Similar highly influential loci have been observed in other organisms [60, 71]. These pivotal loci act as a major source of noise dependence, regardless of whether genotypes have been measured in an expression study.

As proof of concept, the Brem *et al.* dataset was used to show that well-defined noise-dependence exists in actual studies and that surrogate variables can properly capture and incorporate this structure into the statistical analysis of measured variables of interest. First, we analyzed the full dataset to identify the expression traits under the influence of these pivotal transacting loci, as well as the patterns of noise dependence induced by these loci. Linkage P-values were calculated from an F-test comparing an additive genetic model to the null model of no genetic association. Then we applied subset PCA to identify surrogate variables from only the expression

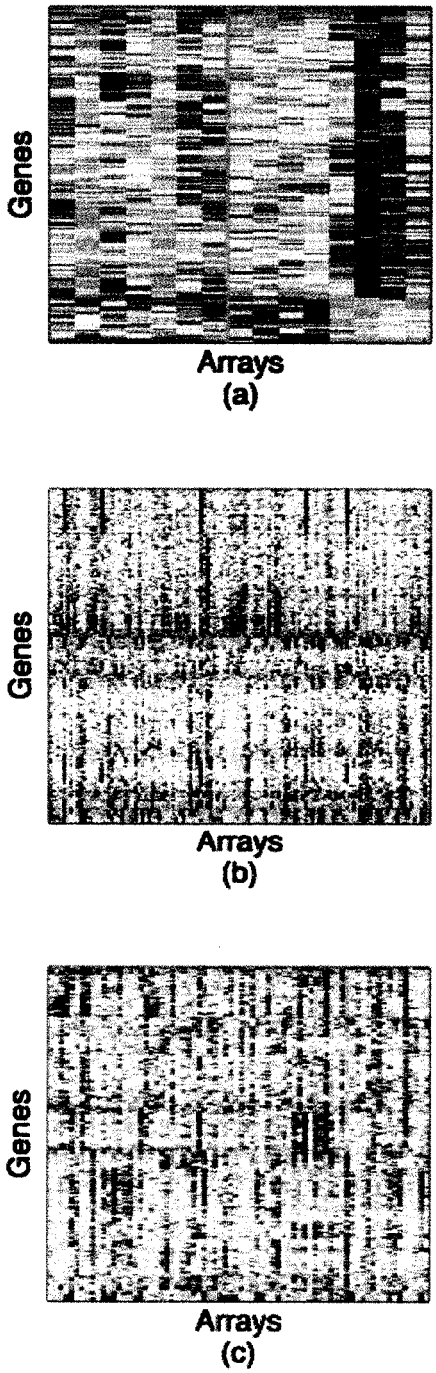

Figure 5.1: Heatmaps of hierarchically clustered gene expression data for a random subset of 1,000 genes from three studies are shown. (a) Hedenfalk *et al.* compared gene expression across tumor subtypes defined by germline *BRCA* mutations (yellow divides *BRCA* tumor subtypes), (b) Brem *et al.* measured expression in naturally recombining yeast populations, and (c) Rodwell *et al.* measured gene expression in kidney samples for patients ranging in age from 27-92 y.

data, ignoring the genotype data. Linkage analysis was performed again including the surrogate variables as covariates, showing that the effects from the pivotal loci are now negligible. In other words, the surrogate variables were able to capture and remove the effects of these few pivotal loci without the need for genotypes.

A number of expression traits have signicant *trans*-linking eQTL mapping to pivotal loci on Chromosomes II, III, VIII, XII, XIV, and XV (Figure 5.2(a)). In the surrogate variable-adjusted analysis, the majority of the *trans*-linkages to the pivotal loci have been eliminated (Figure 5.2(b)). The pervasive *trans*-linkage signal mapping to the pivotal loci can be viewed as noise dependence, or what has been called global expression heterogeneity. The reduction in *trans*-linkage to these loci in the surrogate variable-adjusted signicance analysis indicates that subset PCA estimates effectively capture genetic noise-dependence.

Pivotal *trans*-linkage signals indicate large-scale effects of a few loci. However, subtle and potentially more interesting *cis*-linkage may be lost in the presence of substantial genetic heterogeneity. To assess the impact of including surrogate variables on power to detect *cis*-linkage, we calculated linkage P-values only for markers located within three centimorgans of the open reading frame of each trait. On chromosomes without a pivotal locus (Chromosomes I, IV, V, VI, VII, IX, X, XI, and XIII) the surrogate variable-adjusted analysis finds substantially more *cis*-linkage signal. At a Q-value cutoff of 0.05, the adjusted analysis finds 1,894 significant *cis*-linkages, compared with 1,604 for the unadjusted analysis. This increase is consistent across a range of FDR cutoffs (Table 5.1) and illustrates the potential increase in power obtained from applying surrogate variables.

5.2 Proof of Concept: Human Expression Studies

We applied the SVA approach to two human studies [38, 69], representing the two common human study designs: disease class and timecourse.

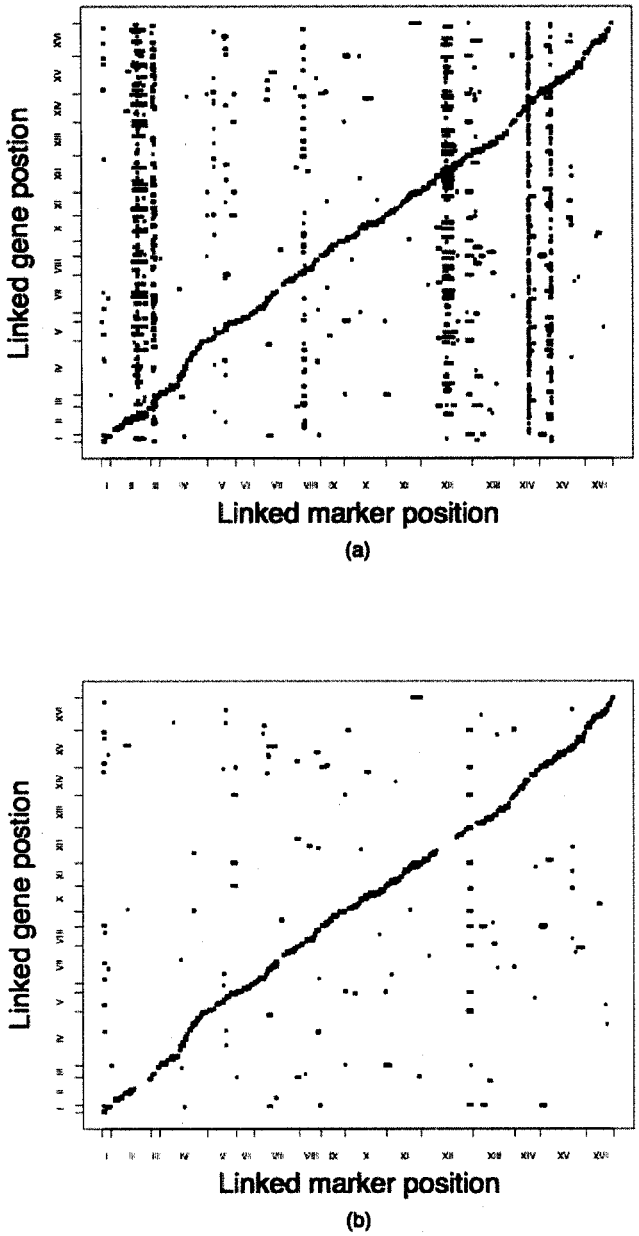

Figure 5.2: (a) A plot of significant linkage peaks (P-value $< 1e-7$) for expression QTL in the Brem *et al.* [12, 11] study by marker location (x-axis) and expression trait location (y-axis). (b) Significant linkage peaks (P-value $< 1e-7$) after adjusting for surrogate variables. Large *trans*-linkage peaks on Chromosomes II, III, VII, XII, XIV, and XV have been eliminated without reducing *cis*-linkage peaks.

Table 5.1: The results of the significance analysis in the three gene expression studies. The results of the genetics of gene expression study include the number of significant *cis*-linkages before and after adjusting for surrogate variables [12, 11]. The disease class results report the number of genes differentially expressed between *BRCA1* and *BRCA2* before and after adjusting for surrogate variables [38]. For the time-course study, the number of genes differentially expressed with respect to age are shown for an unadjusted analysis, an analysis adjusted for tissue type, and an analysis adjusted for surrogate variables [85]. The surrogate variable-adjusted analysis may result in an increase or decrease in the number of significant results depending on the direction and degree to which the unmodeled factors (now captured by surrogate variables) were confounded with the primary variables.

		Q-value Threshold			
Study	Analysis Type	0.01	0.025	0.05	0.10
Genetics of Gene Expression	Unadjusted	1,063	1,343	1,604	1,951
	Adjusted	1,421	1,650	1,875	2,292
Disease Class	Unadjusted	1	19	96	275
	SV Adjusted	1	1	11	59
Timecourse	Unadjusted	161	273	422	823
	Tissue Adjusted	270	482	795	1548
	SV Adjusted	195	367	563	991

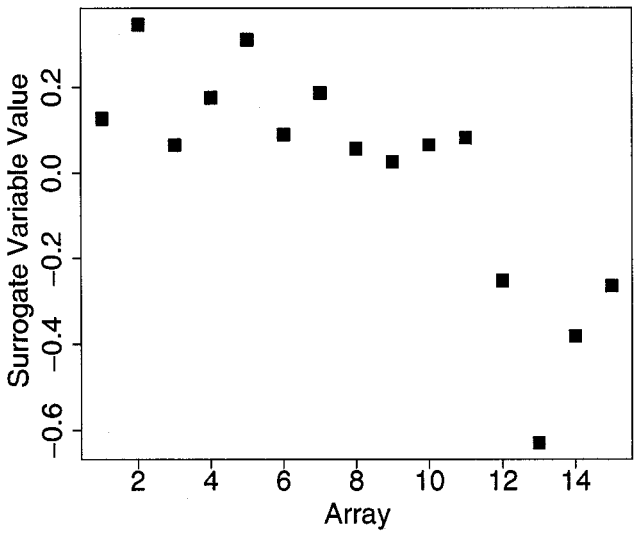

Figure 5.3: A plot of the top surrogate variable estimated from the breast cancer data. The *BRCA1* group is relatively homogeneous (black squares), but the *BRCA2* group shows substantial heterogeneity (blue squares).

5.2.1 Disease Class Study

Hedenfalk *et al.* [38] measured expression in seven *BRCA1* and eight *BRCA2* mutation-positive tumor samples. The goal of the study was to identify genes that showed differential expression across breast cancer tumor subtypes defined by these germline mutations. The dataset consists of gene expression for 3,226 genes in seven *BRCA1* and eight *BRCA2* mutation-positive tumor samples; several genes with apparent outliers were removed as previously described [84] for a total of 3,170 genes. Hierarchical clustering [32] of the data reveals notable substructure within the *BRCA2* samples [39]. We applied SVA and identied a single surrogate variable that appears to capture this trend (Figure 5.3).

We included this surrogate variable in a signicance analysis comparing *BRCA1*

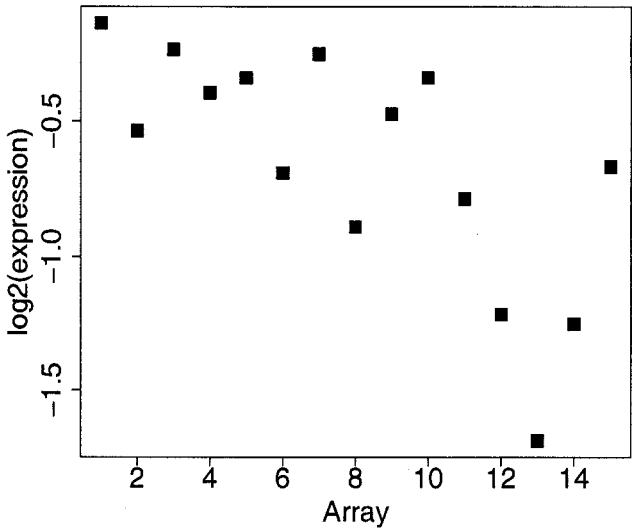

Figure 5.4: A plot of the expression for eukaryotic translation initiation factor 2, *EIF2S2*, which follows a similar pattern to the top surrogate variable from the *BRCA* data.

and *BRCA2* tumors. Differential expression was calculated using a t-test based on standard linear regression for the disease class data. The adjusted analysis finds fewer signicant genes at standard Q-value cutoffs (Table 5.1). This can be understood in the context of substructure within the *BRCA2* group. Many of the genes declared differentially expressed at the most extreme levels of significance are highly associated with the top surrogate variable. Thus, differential expression for a number of genes is driven primarily by noise-dependence. Adjusting for the top surrogate variable can eliminate spurious differential expression.

As an example, eukaryotic translation initiation factor 2 (*EIF2S2*) is declared differentially expressed with a Q-value of 0.01 in the unadjusted analysis. However, Figure 5.4 shows the first four *BRCA2* samples have nearly identical expression values to the *BRCA1* samples for this gene, and the expression values are highly correlated

with the surrogate variable plotted in Figure 5.3. Thus, it is unlikely that differential expression is being driven by the difference in *BRCA* genotypes, but rather by some other confounding factor due to the observational nature and small sample size of the study. The Q-value for this gene increases to 0.58 after adjusting for the top surrogate variable.

As shown in the simulated expression studies from the previous chapter, adjusting for surrogate variables also increases the accuracy and stability of the ordering of the signicant gene lists. Since it is standard practice to examine only the most signicant genes for further study, a surrogate variable adjusted analysis may result in completely distinct biological conclusions. Genes such as *EIF2S2* have substantially different positions in the adjusted versus the unadjusted analysis. These genes may represent spurious signal due to the confounding that would reduce the quality of the gene list.

5.2.2 Timecourse Study

Rodwell *et al.* [69] measured genomewide expression in kidney tissue samples from 133 patients. The dataset consists of gene expression measurements in kidney samples from normal kidney tissue obtained at nephrectomy from 133 patients; the 34,061 genes analyzed in Storey *et al.* [85] were also analyzed here. Seventy-four of the tissue samples were obtained from the cortex and 59 from the medulla. The goal of the study was to identify genes whose expression changed with age. We applied a recently developed procedure for time-course significance analysis to identify differential expression with respect to age [85]. In these data, tissue type had a strong impact on the expression of thousands of genes. We first performed a timecourse differential expression analysis with tissue type included as a covariate. We also performed a second differential expression analysis ignoring tissue type.

We then applied SVA to the expression data ignoring the tissue information. The top surrogate variable identied by SVA had a correlation of 0.86 with tissue type (Figure 5.5). The SVA algorithm identied 84% of the genes as likely to be associated

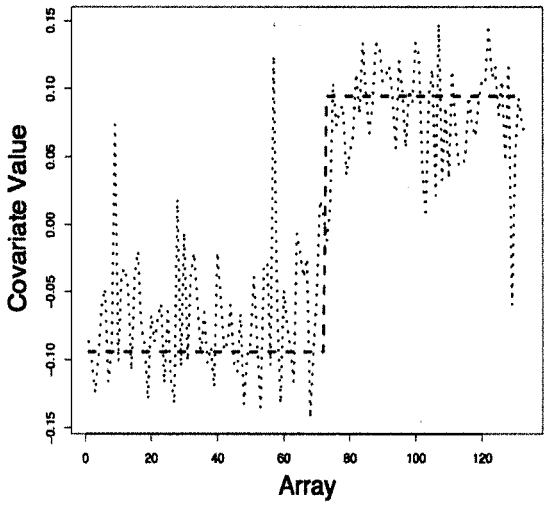

Figure 5.5: A plot of tissue type versus array for the Rodwell *et al.* study (dotted line) and the top surrogate variable estimated from the expression data when tissue was ignored (dashed line). There is strong correlation between the top surrogate variable and the tissue type variable.

with the top surrogate variable, indicating pervasive signal due to tissue type, as can be directly verified. To determine if this surrogate variable captured the overall effect of tissue type, we performed a third differential expression analysis ignoring tissue type and including the top surrogate variable as a covariate.

At standard Q-value cutoffs, the results of the analysis adjusted for the top surrogate variable appeared to be very similar to the results when the true tissue type was included (Table 5.1). At a Q-value cutoff of 0.05, 100% of the 422 genes declared significant by the unadjusted analysis were declared significant by the tissue-adjusted analysis. At the same cutoff, 96% of the 563 genes declared signicant in the surrogate variable-adjusted analysis were also declared significant in the tissue-adjusted analysis. That is, 141 genes were significant in the surogate variable-adjusted analysis that were also signicant in the tissue-adjusted analysis, but were not signicant in the unadjusted analysis. These genes represent an increase in power to detect dif-

ferential expression after adjusting for a surrogate variable in place of an unmodeled confounding factor.

Chapter 6

IMPROVING REPRODUCIBILITY IN HUMAN CLINICAL GENOMIC STUDIES

An important feature of the surrogate variable approach to significance analysis is that the ranking and false discovery estimates are more stable across repeated experiments. Increased stability in expression analysis has important consequences, particularly for clinical genomics, where any technology or predictive tool applied to patients must generate results that are reproducible and accurate. Although the maturation of high throughput technology allows investigators to associate expression variation with clinical outcomes [13, 3, 4, 36, 55], genomics has suffered from a lack of reproducible results in the clinical setting [23, 91]. This has most thoroughly been analyzed in the cancer field, where early prognostic studies were small, poorly designed, and lacked proper validation [61]. A major source of variability that has not been adequately addressed is noise dependence, sometimes called molecular heterogeneity. Noise dependence plays a particularly important role in human studies, where it is virtually impossible to control all of the genetic [71, 86] and environmental [46, 93] confounders that may affect the relationship between gene or protein expression signatures and clinical outcomes.

6.1 Trauma Glue Grant Analysis

In this chapter we analyze data from the Inflammation and the Host Response to Injury Program [89, 17], as a targeted case study of the general phenomena of molecular heterogeneity and irreproducibility in clinical genomics. The results presented in this chapter represent joint work with J. Perren Cobb, Ronald Tompkins, John

Storey, and the Inflammation and the Host Response to Injury Program Investigators; the results will appear in a forthcoming publication. We show that widespread noise-dependence in this study obscures the relationship between genomic signatures and clinical outcomes. We then apply our newly developed technique for surrogate variable estimation to carefully model and account for heterogeneity in data generated by this Program, resulting in major improvements in reproducibility of both global signal and specific functional outcomes.

The Trauma Glue Grant is a collaborative effort among physicians, experimental biologists, and bioinformaticians designed to characterize genomic responses to blunt trauma and burn injury, with the goal of identifying novel molecular markers for complications and adverse outcomes. For a subset of these patients (n = 168), genome-wide gene expression was measured for leukocyte fractions isolated from peripheral blood samples using validated protocols and Affymetrix HU133 Plus 2.0 oligonucleotide arrays, as previously described [17]. The collection of data on the168 patients may be separated arbitrarily into four phases defined by independent and non-overlapping periods of time, each corresponding to approximately half a year (Table 6.1).

To assess reproducibility within this study, we separated data collection into four independent phases. Each phase is subject to the same sampling scheme, so we can consider Phases II-IV as replication studies for Phase I. For each of the phases, we performed an identical analysis: the association between gene expression at the time of hospital admission and subsequent Marshall score for multiple organ failure (MOF1, excluding central nervous system), using a standard linear model. To reduce potential bias due to technical factors, we excluded genes with greater than 6% missing values and those with low RNA quality scores. Using Ingenuity Pathway Analysis [44], a database tool for exploring functional and network relationships among genes, we identified the major functional groups overrepresented among the informational genes for each phase. A comparison of these major functional groups across phases gives

Table 6.1: Characteristics of patients in the Trauma Glue Grant gene expression study of inflammation and the host response to injury.

Characteristic	Phase I	Phase II	Phase III	Phase IV
Age, mean (SD),y	32.3(9.5)	33.5(11.8)	32.7(11.3)	39.0 (11.2)
Male sex, No. (%)	24 (57)	24 (65)	40 (71)	19 (61)
Racial/Ethnic Group, No. (%)				
White, non-Hispanic	33 (79)	30 (81)	48 (86)	23 (75)
Black, non-Hispanic	4 (10)	1 (3)	2 (4)	2 (6)
Hispanic	3 (7)	2 (5)	2 (4)	4 (13)
Other or missing	2 (5)	4 (11)	4 (5)	2 (6)
Date of Microarray, First, Last	5/25/04-9/02/04	9/27.04-3/24/05	8/24/05-2/14/06	5/25/06-7/01/06
Modified Marshall Score (excluding CNS) mean (SD)	4.2 (2.3)	3.0 (1.8)	3.9 (2.0)	4.4 (2.0)

some indication of the functional reproducibility of the inferred differential expression signal.

For the four phases of the study, the estimates of the proportion of truly null hypotheses obtained without accounting for surrogate variables are 14%, 0%, 23%, and 4%, respectively. This type of variability is not surprising in light of the variation in π_0 estimates due to noise dependence illustrated by Figure 2.5 . However, it is obviously not satisfactory that the estimated number of differentially expressed genes ranges from zero to nearly one quarter of all the genes in the four replicated study phases.

Figure 6.1 clearly shows the global distribution of signal changes from phase to phase. The variation in global signal directly translates into poorly replicated functional conclusions across the four phases of the study. For example, among the differentially expressed genes, those involved in immune response are overrepresented at the $P < 0.05$ level for only Phases I, II, and IV; and genes involved in inflammatory disease are only overrepresented in Phases I and III (Figure 6.2(a)). From these results and those below, we hypothesize that the primary reason for the irreproducible results in this Glue Grant study is the molecular heterogeneity within each phase.

Molecular heterogeneity, or noise dependence, has been defined in this dissertation as a consistent pattern of variation across gene expression measurements within a patients genomic profile. Figure 6.3 is a heatmap of expression data from 3000 genes for phases I-IV of the study. Since each separate row of the heatmap includes the expression values for a distinct gene across patients, molecular heterogeneity can be seen in the heatmap as consistent blocks of vertical blocks of color.

As a simple hypothetical example of the variance that could produce such a patttern, consider a clinical expression study comparing patients with low and high MOF1 scores. If both male and female patients are sampled, then some genes will be differentially expressed with respect to MOF1 score, some with respect to sex, and some with respect to both. When analyzing the data with respect to MOF, if unaccounted

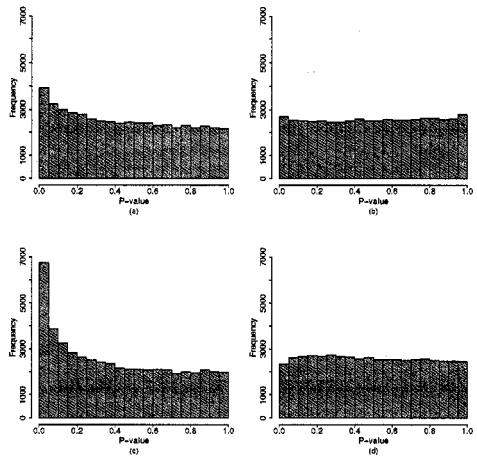

Figure 6.1: Histograms of P-values calculated from the four temporally defined phases, demonstrating molecular heterogeneity manifested as global irreproducibility in signal.

for, the consistent sex related expression patterns would be considered molecular heterogeneity.

A specific example of the motivation for surrogate variables can be seen in Figure 6.4, where separate subsets of genes are associated with the primary clinical outcome and a second distinct pattern of variation. To a greater or lesser extent, the heatmaps for each of the four phases (Figure 6.1), show similar patterns of signal and noise-dependence due to some other unmodeled variable. The data for each heatmap is arranged so that MOF1 is increasing from left to right, and it is clear that the primary variable and unmodeled factor have different relative configurations across the four phases.

Estimating and accounting for surrogate variables using the subset PCA algorithm reduced variation in the global signal structure across the four phases. The estimates for the proportion of differentially expressed genes for the four phases after subset PCA are: 15%, 11%, 10%, and 7%. Figure 6.5 shows that the global distribution of expression signal is also stabilized compared with the unadjusted analysis. Function-

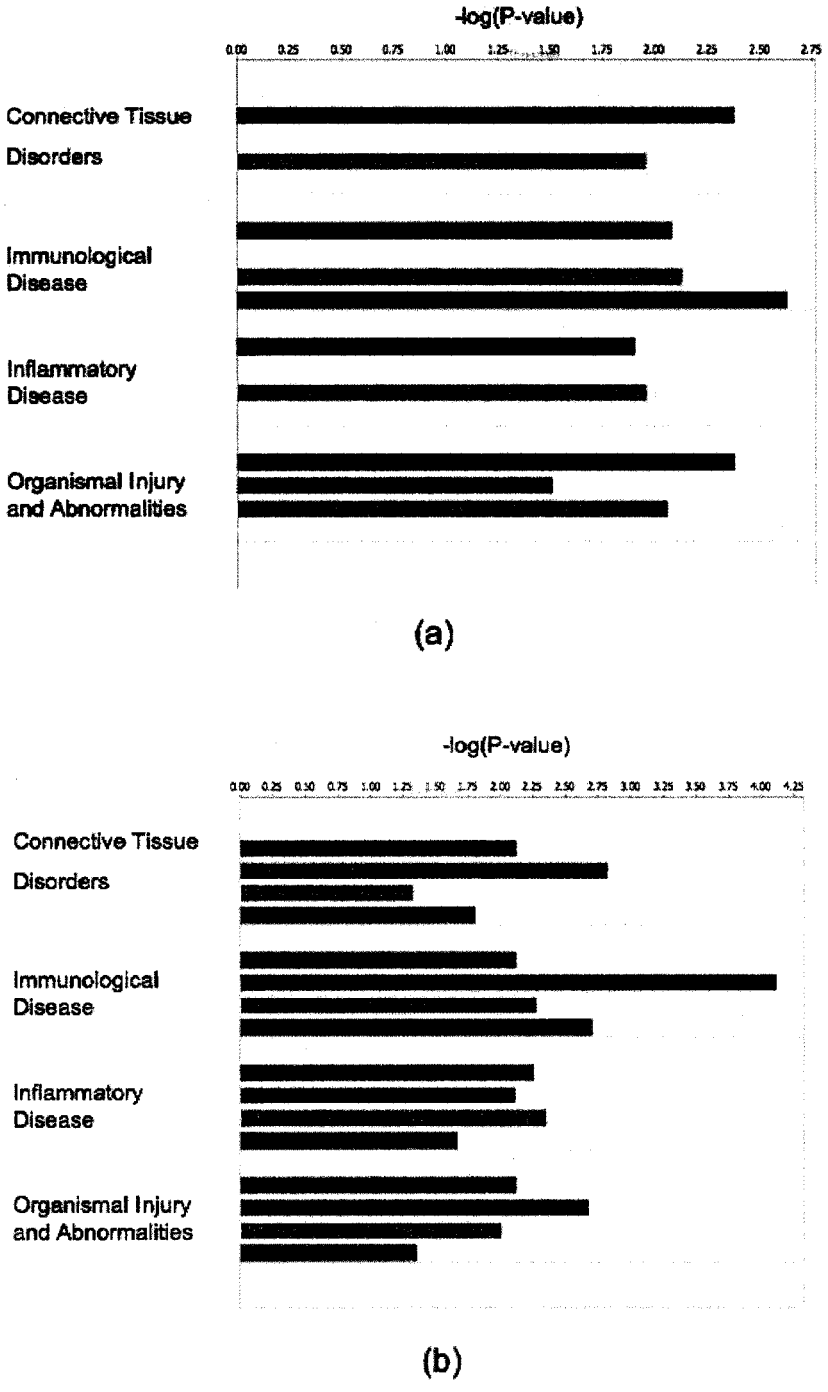

Figure 6.2: Categories with significant functional enrichment for the phase one (navy), phase two (light blue), phase three (azure), and phase four (black). (a) Functional enrichment results for the unadjusted analysis. (b) Functional enrichment results for the surrogate variable adjusted analysis.

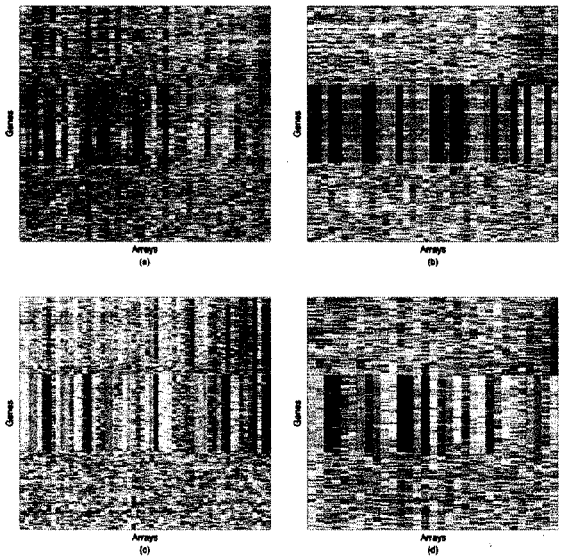

Figure 6.3: Heatmaps of microarray expression data from the Trauma Glue Grant study arranged according to common trends from (a) Phase I, (b) Phase II, (c) Phase III and (d) Phase IV.

ally, the results from the four phases show greater consistency and reproducibility (Figure 6.2(b)). Among the differentially expressed genes, both immune response genes and inflammatory disease genes are overrepresented in the SVA adjusted analysis for all four phases ($P < 0.05$) according to an Ingenuity Pathway Analysis.

These increases in functional annotation stability are not surprising in light of the variable correlation between the estimated surrogate variables and MOF1 across the four phases. For example, the correlation coefficient between the top surrogate variable and MOF1 across the four phases is -0.15, 0.01, -0.31, and -0.11. This indicates that there is a fluctuating source of molecular heterogeneity confounding the relationship between expression and clinical outcome. Besides these global measures of reproducibility, we demonstrated through simulation that the stability of the ranking of the genes for differential expression when applying surrogate variables is nearly equivalent to the stability obtained under the ideal scenario of no heterogeneity (e.g.,

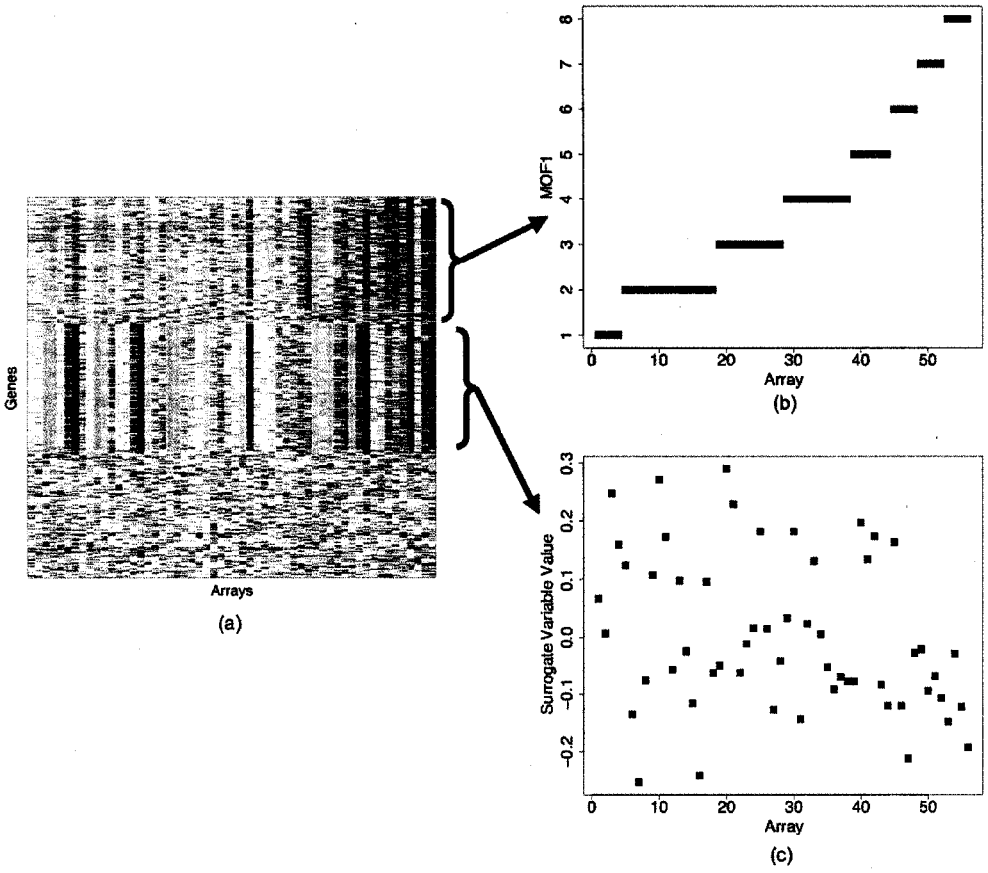

Figure 6.4: Motivation for surrogate variables in the Trauma expression data. (a) A heatmap of microarray expression data arranged according to common trends, derived from phase III of the Trauma Glue Grant study. (b) Genes 1-1000 are strongly associated with (modified) Marshall score for multiple organ dysfunction syndrome (excluding central nervous system data) at the time of the patient admission. (c) Genes 1001-2000 show a common pattern of heterogeneity present in phase III.

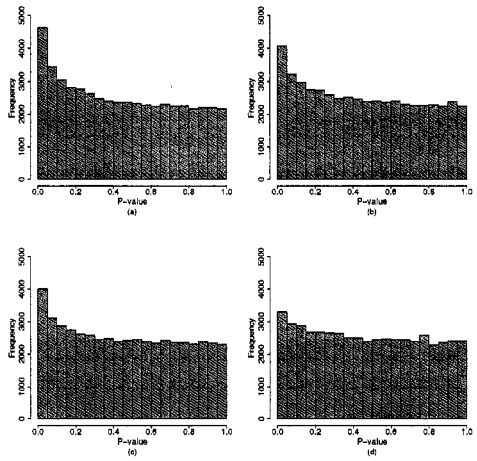

Figure 6.5: Histograms of P-values calculated after adjusting for surrogate variable estimates that account for noise dependence. The global distribution of signal now appears to be quite similar across phases, with slight differences due to the varying sample size, and hence the power, across the phases.

Figure 4.1).

Surrogate variable analysis is not meant to replace the inclusion of variables that are measured and known to be relevant for modeling expression changes. If a reasonable number of relevant variables are available, then they should be included among the primary variables used to build the surrogate variables and subsequently analyze the expression data. However, when a large number of clinical variables have been measured, as is the case for the Trauma Glue Grant, it is often not known what subset of these variables to include in the model for expression. Given the large number of possible models one has to consider, the degrees of freedom are not available to search over all possible models, in which case SVA can be employed to avoid this intractable problem. SVA is also capable of capturing a complex model of several variables using a smaller number of degrees of freedom.

We have shown that accounting for molecular heterogeneity in the trauma expression data set substantially improves reproducibility. However, this is not an isolated

example of molecular heterogeneity. Noise-dependence is a ubiquitous phenomena in genomics studies; it has been recognized as a key hurdle in genomic analysis, from the first expression profiles of cancer [38, 64],to recent comprehensive databases such as the Connectivity Map[50] or the Onconomics Project [68]. This raises the question: why is heterogeneity so common in genomic studies, and particularly common in clinical genomics studies?

However, heterogeneity can arise from a variety of sources often times too difficult to measure or model. Genetic variation across patients [71, 86], varying environmental exposures [46], geographically defined ancestry [82] and demographic variables such as age [69] may all contribute heterogeneity to clinical genomics studies. In most cases, it is impossible to measure all the variables that could potentially confound the relationship between expression profiles and clinical outcomes. Even in studies where hundreds of relevant clinical, environmental, demographic, and genetic markers are measured, it is difficult to determine the appropriate combination of covariates that appropriately account for even the major sources of heterogeneity. The dynamic, complex behavior of gene and protein expression makes heterogeneity unavoidably present in clinical genomics, which motivates development and application of new approaches that specifically measure and account for it.

6.2 Summary

It is clear from the data presented in this chapter that molecular heterogeneity can affect both global genomic signatures of disease and the functional relationships of genes inferred from clinical studies. By explicitly modeling and accounting for noise-dependence with new tools such as surrogate variables, it is possible to substantially improve the accuracy of of biomarkers and the reproducibility of genomic signatures.

Chapter 7

THEORETICAL RESULTS

The two steps of any multiple testing procedure, significance ranking and error rate estimation, may be biased and highly variable in the presence of noise-dependence. The big picture goal of the surrogate variable framework and estimation algorithms proposed in this dissertation is to reduce or eliminate the bias and variability in multiple testing procedures. In this chapter we specialize this goal to two specific points. First we describe a theoretical framework that connects the results of Chapter 3 with previous work in error rate estimation for multiple testing to show that knowing surrogate variables can eliminate multiple testing dependence. Second, we focus on theoretical results that form a backbone for a general theoretical treatment of surrogate variable estimation. We conclude by drawing connections between the Algorithms proposed in this dissertation and maximum likelihood estimates under simplifying assumptions.

7.1 Multiple Testing Dependence

The first step in a multiple testing procedure is to calculate a statistic for each feature measuring its association with the primary variable. There is a growing interest in developing new statistics specifically for high-dimensional data that borrow strength across features, such as variance shrinkage statistics [22], empirical Bayes statistics [31], and the recently proposed optimal discovery procedure statistics [80, 81]. These statistics can improve power and possibly improve the ranking of features for association with the primary variable. However, much of the statistical literature dedicated to inference for high-dimensional data has focused on appropriately controlling error

rates.

The error rate measure that has received much of this focus is the false discovery rate (FDR) [9]. One of the reasons is that the FDR is a more liberal error rate than other potential measures such as the family-wise error rate [9, 79]. A second reason is that the FDR naturally quantifies the level of noise in the significant feature lists produced from a high-throughput study. If the FDR is controlled at 5%, the percent of false associations among the significant features should be, on average, less than 5%. In high-throughput molecular biology experiments, such as gene expression experiments, the significant features are often used to build a network based on curated ontologies of biological function. Controlling the level of noise in the significant feature list is one way of controlling the level of noise in the inferred functional networks from downstream tools.

Estimates for the FDR have been developed based on P-values from the m hypothesis tests, one for each feature in the high-throughput study. The first FDR estimates were based on step-up algorithms that sequentially considered the P-values from smallest to largest. A simple plug in was proposed by Storey (2002) [79],

$$\widehat{\text{FDR}}_\lambda(t) = \frac{\widehat{\pi}_0(\lambda)t}{(\sum_{i=1}^m \mathbf{1}(P_i \leq t) \vee 1)/m}$$

where:

$$\widehat{\pi}_0(\lambda) = \frac{\sum_{i=1}^m \mathbf{1}(P_i > \lambda)}{(1-\lambda)m}$$

is an estimate of the proportion of features whose data truly follows the distribution under the null hypothesis. Storey, Taylor and Siegmund showed that $\widehat{FDR}_\lambda(t)$ defines a class of conservatively biased point estimates for the false discovery rate under the assumption that the P-values from truly null hypothesis tests are independent and uniformly distributed [83].

Theorem 3. Storey, Taylor and Siegmund (2004) *Suppose the P-values from the true null hypothesis tests are independent and uniformly distributed. Then for a fixed $\lambda \in [0, 1)$, $\mathbf{E}[\widehat{\text{FDR}}_\lambda(t)] \geq \text{FDR}(t)$.*

The assumptions of Theorem 3 are the same as those that Benjamini and Hochberg (1995) [9] used to show strong control of the FDR with their step-up procedure. While this is a satisfying result, in practice, it is not always the case that the P-values for the null statistics are independent or uniformly distributed.

A more powerful result is possible when the number of hypothesis tests grows to infinity, an assumption that makes sense for high-throughput data sets where the number of features ranges from several thousand to millions. Asymptotically in m, Storey, Taylor and Siegmund showed that for certain types of dependence between the P-values, the FDR estimate $\widehat{\text{FDR}}_\lambda(t)$ is also simultaneously conservatively consistent for $\text{FDR}(t)$ [83].

Theorem 4. Storey, Taylor and Siegmund (2004) *Suppose the following four conditions hold.*

1. $\frac{1}{m_0}\sum_{i=1}^{m_0} \mathbf{1}(p_i \leq t) \rightarrow_{a.s.} G_0(t)$.

2. $\frac{1}{m_1}\sum_{i=m_0+1}^{m} \mathbf{1}(p_i \leq t) \rightarrow_{a.s.} G_1(t)$.

3. $0 < G_0(t) \leq t$ *for each* $t \in (0, 1]$.

4. $\lim_{m\rightarrow\infty} \frac{m_0}{m} \equiv \pi_0$ *exists.*

Where G_0 *and* G_1 *are continuous. Then for each* $\delta > 0$,

$$\lim_{m\rightarrow\infty} \inf_{t\geq\delta} \left\{ \widehat{\text{FDR}}_\lambda(t) - \text{FDR}(t) \right\} \geq 0$$

Here the P-values do not have to be independent, as in Theorem 3; the assumption is merely that distribution functions exist for the collection of null and alternative P-values. The distribution of the null P-values must also be stochastically greater than or equal to the $\mathcal{U}(0, 1)$ distribution. The types of dependence allowed by Theorem 4

were called "weak dependence" by Storey, Taylor and Siegmund. Examples of this type of dependence include ergodic dependence or dependence in finite blocks [83].

Considerable effort has been applied toward understanding the behavior of false discovery rate estimates when the assumption of "weak dependence" does not hold [48, 66, 62]. In the statistical literature, multiple testing dependence is defined as dependence across P-values from multiple hypothesis tests [62, 30]. In some cases, standard FDR estimates are conservative, even under dependence. For example, Benjamini and Yekutelli show that if the P-values are positive regression dependent, then the Benjamini-Hochberg estimate controls the FDR [10]. However, assumptions such as positive regression dependence are not intuitive, and are unlikely to hold in practice. Empirical corrections to the null distribution of the P-values have also been proposed, but suffer from unidentifiability of the null and alternative distributions of P-values under dependence [30]

We take a distinct approach to addressing multiple testing dependence based on estimating sources of noise-dependence in the feature level data. When noise-dependence exists across features, there will also be dependence between the P-values calculated for those features, so the first assumption of Theorem 3 is violated. Moreover, the common assumption of Theorems 3 and 4 is that the null P-values come from a distribution that is uniform or stochastically greater than the uniform. However, in Chapter 2, we showed that under noise-dependence the null P-values may have a distribution that is stochastically smaller than the uniform, e.g. Figure 2.3(e) and (h). This occurs when the unmodeled factors causing noise dependence are correlated with the primary variable.

A question of considerable interest is if the assumptions of Theorems 3 and 4 are satisfied for noise-dependent data when the true surrogate variables are included in the model between the feature level data and the primary variable. To answer this question, we begin by assuming model 3.6 holds, so the feature level data can be

written,

$$\mathbf{X} = \boldsymbol{\beta}\mathbf{S}^T + \mathbf{g}(\mathbf{K}) + \mathbf{U}$$

where the elements of $\mathbf{U}$ are mutually independent. Theorem 5 gives a special case where knowing the surrogate variables is sufficient for the assumptions of Theorem 3 to hold.

Theorem 5. *Suppose that a set of surrogate variables,* $\mathbf{G}$*, for model 3.6 are known where* $\text{rank}(\mathbf{G}) + \text{rank}(\mathbf{S}) < n$ *and that* $u_{ij} \overset{i.i.d.}{\sim} \text{N}(0, \sigma_i^2)$ *where* $0 < \sigma_i^2 < B$ *for* $i = 1, \ldots, m$*. Then if the hypotheses:*

$$H_{0i} : \boldsymbol{\beta}_{i\cdot} = 0 \qquad \text{vs.} \qquad \text{H}_{1\text{i}} : \boldsymbol{\beta}_{\text{i}\cdot} \neq 0$$

are tested using the standard F-statistic, the conditions of Theorem 3 are satisfied.

Proof. For each feature $\mathbf{x}_{i\cdot}$ we fit the alternative model:

$$\mathbf{x}_{i\cdot} = \boldsymbol{\beta}_{i\cdot}\mathbf{S}^T + \boldsymbol{\gamma}_i\mathbf{G}^T + \mathbf{u}_{i\cdot}$$

and calculate the residual sum of squares, RSS_i. We also fit the null model:

$$\mathbf{x}_{i\cdot} = \boldsymbol{\gamma}_i\mathbf{G}^T + \mathbf{u}_{i\cdot}$$

and calculate the corresponding residual sum of squares, RSS_i^0. Then for each feature we compute the F-statistic:

$$F_i = \frac{(\text{RSS}_i^0 - \text{RSS}_i)/(d)}{\text{RSS}_i/(n-d-r)}$$

and compare it to the $F_{d,n-d-r}$ distribution. The u_{ij} are normally distributed, so when $\boldsymbol{\beta}_{i\cdot} = \mathbf{0}$, F_i is distributed $F_{d,n-d-r}$ and the P-values for the null statistics are uniform. Further, the residuals are independent across rows under the alternative model fit by Theorem 2 on page 34. The residuals from the null model fit, when the null is true, can also be easily shown to be independent following the same argument as in Theorem 2. Since the F-statistics are based on functions of independent data they are independent, and the assumptions of Theorem 3 are met. □

Theorem 5 relies on knowing the distributional form of the feature specific noise. But a more general intuition can be drawn about the problem of multiple testing dependence from the proof of this theorem. If the surrogate variables are known, any test statistic based on a function of the residuals from the null and alternative model fits is independent across null features. A number of statistics, including the F-statistic and generalized likelihood ratio statistic based on the normal likelihood can be reduced to functions of the residuals [52]. We showed in Lemma 1 on page 29 that the estimated coefficients $\widehat{\beta}_{i\cdot}$ are also independent across features when the null hypothesis true. So statistics that are a function of both the estimated coefficient and the residuals such as the Wald statistic and score statistic are also independent across features.

The key point is that the independence assumption of Theorem 3 is exactly met under noise-dependence when the surrogate variables are included in the model. While the assumption of uniform P-values may be only approximately true with finite sample sizes, this is no better than the case where the hypothesis tests are independent. So as a point estimate, $\widehat{\text{FDR}}_\lambda(t)$, is as nicely behaved under noise-dependence where the surrogate variables are known as in the case when there is no multiple testing dependence.

Theorem 3 of Storey, Taylor and Seigmund [83] also shows that the above FDR estimate can be used to control FDR at a given level by adaptively choosing the threshold. This theorem also requires only the independence of the null P-values across features. Therefore, under the framework proposed here, we have extended both our Theorem 3 and Theorem 3 of Storey, Taylor and Seigmund to hold under arbitrarily strong dependence, as long as SVA is applied as described.

Theorem 3 shows that the estimate, $\widehat{\text{FDR}}_\lambda(t)$, strongly controls the FDR point-wise for every t. However, the result of Theorem 4 is in some sense more powerful, since it shows that the estimate $\widehat{\text{FDR}}_\lambda(t)$ is simultaneously conservative consistent across all significance thresholds as the number of tests grows large. This result can

be more useful as many high-dimensional experiments are exploratory, and it may not make sense to fix a cutoff t before performing the analysis. Theorem 6 shows that knowing the surrogate variables is also sufficient for the results of Theorem 4 to hold.

Theorem 6. *Suppose that a set of surrogate variables,* $\mathbf{G}$*, for model 3.6 are known where* $\text{rank}(\mathbf{G}) + \text{rank}(\mathbf{S}) < n$ *and that* $u_{ij} \overset{i.i.d.}{\sim} \text{N}(0, \sigma_i^2)$ *where* $0 < \sigma_i^2 \leq B$ *and* $-\infty < \beta_{ij} < \infty$ *for* $i = 1, \ldots, m$*. If the hypotheses:*

$$H_{0i} : \boldsymbol{\beta}_{i\cdot} = 0 \qquad \text{vs.} \qquad \text{H}_{1\text{i}} : \boldsymbol{\beta}_{\text{i}\cdot} \neq 0$$

are tested using a standard F*-statistic then assumptions (1-3) of Theorem 4 are satisfied.*

Proof. From the proof of Theorem 5 it is clear that the statistics from the null hypothesis tests are independent and identically distributed $F_{d,n-d-r}$, so assumptions (1) and (3) hold by the Glivenko-Cantelli theorem. When $\beta_{i\cdot}$ is not equal to zero, the statistic for feature i has a non-central F distribution with non-centrality parameter,

$$\eta(\boldsymbol{\beta}_{i,\cdot}, \sigma_i) = (\boldsymbol{\beta}_{i\cdot}\mathbf{S}^T(\mathbf{I} - \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T)^T(\boldsymbol{\beta}_{i\cdot}\mathbf{S}^T(\mathbf{I} - \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T)/\sigma_i^2$$

So the P-values are independent draws from a countable mixture of continuous distributions on $\mathbb{R}^+$ indexed by the parameters $\boldsymbol{\beta}_{i\cdot}$ and σ_i^2, and the empirical distribution functions for countable mixtures that assign probability zero to boundaries of convex sets converge [29] so condition (2) is satisfied. □

Theorem 6 shows that for normally distributed noise-dependent feature data and a specific test-statistic, FDR estimates are simultaneously conservatively consistent. As with the finite sample result, the intuition is considerably more general. Including surrogate variables eliminates the dependence across features, and from Lemma 3 also eliminates the bias in the estimates $\widehat{\beta}_{i\cdot}$. The only consideration that remains is the distributional form of the feature specific noise.

7.2 Convergence of Principal Components

In the previous section we established that conditioning on the true surrogate variables when performing inference is sufficient to reduce or eliminate multiple testing dependence. However, an obstacle to using the surrogate variable approach in practice is to estimate the surrogate variables. In this dissertation we have focused on four algorithms for estimating surrogate variables. Each of these algorithms is based on principal components analysis of large data matrices. In this section we present a theorem about the convergence of principal components as the number of features m grows to infinity and the number of samples stays fixed. This type of result is appropriate in high-dimensional testing where $m \gg n$.

Results regarding the consistency of principal components as the number of independent samples grows large already exist in the statistical [6] and econometric [18] literature. More recently, consistent estimators of the principal components have also been developed when both the number of samples and the number of features is large [7]. However, for high-dimensional data analysis problems, the number of independent samples is usually much smaller than the number of features. Thus, asymptotic results are needed for the case where the number of features grows large, while the number of samples remains fixed.

Estimating the principal components across features is a non-standard statistical problem, since the data for each feature may have a different distribution and the data across features are dependent. However, under certain assumptions, by using information across features the principal components can surprisingly be consistently estimated as the number of features grows large.

Theorem 7. *Suppose* $\mathbf{X}$ *is an* $m \times n$ *matrix such that:*

$$\mathbf{X} = \mathbf{\Gamma}\mathbf{G}^T + \mathbf{U} \tag{7.1}$$

where $\mathbf{\Gamma}$ *is a* $m \times r$ *matrix of coefficients,* $\mathbf{G}$ *is a* $n \times r$ *matrix of factors of rank* r

and $u_{ij} \sim (0, \sigma_i^2)$. *Further suppose that the following assumptions hold.*

1. $0 < \sigma_i^2 \leq B_1$

2. $0 < \mathbf{E}[u_{ij}^4] \leq B_2$

3. $\lim_{m\to\infty} \frac{1}{m}\mathbf{\Gamma}^T\mathbf{\Gamma} = \Delta$, *where* Δ *is positive definite.*

4. $\mathbf{G}\mathbf{\Gamma}^T\mathbf{\Gamma}\mathbf{G}^T$ *has eigenvalues* $\lambda_1, \ldots, \lambda_n$ *where the top* r *eigenvalues are unique and positive and the remaining eigenvalues are zero, i.e.,* $\lambda_1 > \lambda_2 > \cdots > \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$.

5. $\sum_{i=1}^n g_{ij} = 0$

If $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_n)$ *and* $\widehat{\boldsymbol{\lambda}} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_n)$ *are the eigenvalues and eigenvectors of:*

$$\begin{aligned} \mathbf{W}_m &= \frac{1}{m}\mathbf{X}^T\mathbf{X} - \widehat{\bar{\sigma}}^2\mathbf{I} \\ \widehat{\bar{\sigma}}^2 &= \frac{1}{m}\sum_{i=1}^m \widehat{\sigma}_i^2 = \frac{1}{m}\sum_{i=1}^m \left[\sum_{j=1}^n x_{ij}^2 - (x_{ij} - \bar{x}_{i\cdot})^2\right] \end{aligned}$$

Then:

1. $\widehat{\mathbf{V}} \to_{a.s.} \mathbf{V}$

2. $\widehat{\boldsymbol{\lambda}} \to_{a.s.} \boldsymbol{\lambda}$

where $\boldsymbol{\lambda}$ *and* $\mathbf{V}$ *are the eigenvalues and eigenvectors of* $\mathbf{G}\Delta\mathbf{G}^T$.

Proof.

$$\begin{aligned} \mathbf{W}_m &= \frac{1}{m}\mathbf{X}^T\mathbf{X} - \widehat{\bar{\sigma}}^2\mathbf{I} \\ &= \underbrace{\frac{1}{m}\mathbf{G}\mathbf{\Gamma}^T\mathbf{\Gamma}\mathbf{G}}_{i} + \underbrace{\frac{1}{m}\mathbf{G}\mathbf{\Gamma}^T\mathbf{U}}_{ii} + \underbrace{\frac{1}{m}\mathbf{U}^T\mathbf{\Gamma}\mathbf{G}^T}_{iii} + \underbrace{\frac{1}{m}\mathbf{U}^T\mathbf{U}}_{iv} - \underbrace{\widehat{\bar{\sigma}}^2\mathbf{I}}_{v} \end{aligned}$$

We consider each of these terms individually.

i: This term converges to $\mathbf{G}\Delta\mathbf{G}^T$ by assumption 4.

ii: Let $\mathbf{M} = \frac{1}{m}\mathbf{G}\mathbf{\Gamma}^T\mathbf{U} = \frac{1}{m}\mathbf{B}\mathbf{U}$, then $m_{ij} = \frac{1}{m}\sum_{\ell=1}^{m} b_{i\ell}u_{\ell j}$ where $\mathbf{E}[b_{i\ell}u_{\ell j}] = 0$ and $\mathbf{Var}[b_{i\ell}u_{\ell j}] = b_{i\ell}^2\sigma_\ell^2$. So by the Kolmogorov Strong Law of Large Numbers (KSLLN) [33] $m_{ij} \rightarrow_{a.s.} 0$ for all i, j.

iii : By symmetry, this term also converges almost surely to zero.

iv : Let $\mathbf{S} = \frac{1}{m}\mathbf{U}^T\mathbf{U}$, and consider the off-diagonal element $s_{ij} = \frac{1}{m}\sum_{\ell=1}^{m} u_{\ell i}u_{\ell j}$, where $\mathbf{E}[u_{\ell i}u_{\ell j}] = 0$ and $\mathbf{Var}[u_{\ell i}u_{\ell j}] = \mathbf{E}[u_{\ell i}^2u_{\ell j}^2] - \mathbf{E}[u_{\ell i}u_{\ell j}]^2 = (\sigma_\ell^2)^2$. So again by KSLLN $s_{ij} \rightarrow_{a.s.} 0$. Now consider the diagonal elements $s_{ii} = \frac{1}{m}\sum_{\ell=1}^{m} u_{\ell i}^2$, where $\mathbf{E}[u_{\ell i}^2] = \sigma_\ell^2$ and $\mathbf{Var}[u_{\ell i}^2] = \mathbf{E}[u_{\ell i}^4] - \mathbf{E}[u_{\ell i}^2]^2$. By assumptions 1 and 2, the variances are bounded, so by KSLLN $s_{ii} \rightarrow_{a.s.} \bar{\sigma}^2 = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} \sigma_i^2$, which exists because σ_i^2 is bounded for all i.

v : $\mathbf{E}[\widehat{\sigma}_i^2] = \mathbf{E}\left[\sum_{j=1}^{n} x_{ij}^2 - (x_{ij} - \bar{x}_{i\cdot})^2\right] = n\sigma_i^2 + \sum_{k=1}^{r}\gamma_{ik}^2\sum_{\ell=1}^{n} f_{\ell k}^2 - (n-1)\sigma_i^2 - \sum_{k=1}^{r}\gamma_{ik}^2\sum_{\ell=1}^{n} f_{\ell k}^2 = \sigma_i^2$, where the second equality requires assumption 6. $\mathbf{Var}[\widehat{\sigma}_i^2] < B_3$ so by KSLLN, $\widehat{\bar{\sigma}}^2 \rightarrow_{a.s.} \bar{\sigma}^2$.

Combining terms (i-v) and applying Slutsky's theorem yields: $\mathbf{W}_m \rightarrow_{a.s.} \mathbf{G}\Delta\mathbf{G}^T$. Since the eigenvalues of a matrix are defined as roots of a determinant depending on the elements of that matrix, and since the roots of a polynomial equation are a continuous multi-valued function of the coefficients [40], the eigenvalues function is continuous. Thus, by the continuous mapping theorem the eigenvalues of $\mathbf{W}_n$ converge almost surely to the eigenvalues of $\mathbf{G}\Delta\mathbf{G}^T$. Further, since both the matrix $\mathbf{W}_n$ and the eigenvalues converge almost surely, and the eigenvectors can be obtained from a linear operation of these two elements, the eigenvectors also converge almost surely for the unique eigenvalues. □

Assumptions (1) and (2) of Theorem 7 will nearly always hold in practice, as the data for most high-throughput experiments will have bounded variation due to either technological or biological constraints on the system being measured. Assumption (3)

essentially means that any unmodeled factors that are to be consistently estimated must affect a non-negligible proportion of features. This makes sense, since we are considering asymptotic results in the number of features; if an unmodeled factor affected the data for only a fixed number of features, as the number of features grows to infinity, the signal from that factor will be drowned out by the large sample size. Assumption (4) merely says that no unmodeled factor has exactly the same influence across all the features. Since the magnitude of the effect for any two unmodeled factors is unlikely to be exactly the same in any realized study, this assumption may also be reasonable in most studies performed in practice.

7.3 Asymptotically Estimating the Number of Factors

One of the key steps in any principal component analysis is to select the appropriate number of factors to include. This problem has been addressed in a variety of ways since the invention of factor analysis. In the social sciences, graphical methods such as those based on scree plots and heuristic cutoffs based on the percent of variation explained by each factor have been the most popular means of selecting the number of factors.

Bai and Ng (2002) [7] showed that under the assumptions of Theorem 7 and the following additional assumptions,

1. $\mathbf{E}[u_{ij}^8] \leq B_1$

2. $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{G}_{\cdot i}\mathbf{G}_{\cdot,i}^T = \Delta_G$ where Δ_G is positive definite.

3. $\mathbf{E}\left[\frac{1}{n}\sum_{j=1}^{n} \|\frac{1}{\sqrt{m}}\sum_{i=1}^{m} \mathbf{G}_{\cdot i}u_{ij}\|\right] \leq B_2$

the number of factors in a principal component analysis can be consistently estimated as the number of features and the number of samples goes to infinity using Algorithm

Algorithm 6 Bai and Ng (2002) The following algorithm consistently estimates the true number of components r as $\min\{m,n\} \to \infty$.

1: Form estimates $\boldsymbol{\beta}$ by fitting the model $\mathbf{X} = \boldsymbol{\beta}\mathbf{S}^T + \mathbf{E}$ and calculate the residuals matrix $\widehat{\mathbf{E}} = \mathbf{X} - \widehat{\boldsymbol{\beta}}\mathbf{S}^T$.

2: Calculate the singular value decomposition of the residual matrix $\widehat{\mathbf{E}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$.

3: For $k = 1, \ldots, d$, calculate:

$$PC(k) = \log\left(\left|\left|\widehat{\mathbf{E}} - \widehat{\Gamma}_k \mathbf{G}_k^T\right|\right|_F\right) + k\left(\frac{m+n}{mn}\right)\log\left(\frac{mn}{m+n}\right)$$

where $\widehat{\Gamma}_k$ is the least squares estimate of Γ_k given $\mathbf{G}_k$.

4: Set $\hat{r} = \arg\max_k PC(k)$

Using random matrix theory it is also possible to consistently estimate the number of factors as only the number of features grows large, since the rate of convergence of the eigenvalues is known.

Theorem 8. *Suppose the assumptions* $1-6$ *of Theorem 7 hold where* n *is fixed. Then* $\mathbf{1}(\widehat{\lambda}_k \geq m^{-\eta}) \to_P 1$ *for* $k = 1, \ldots, r$ *and* $\mathbf{1}(\widehat{\lambda}_k \geq m^{-\eta}) \to_P 0$ *for* $k = r+1, \ldots, n$ *and* $0 < \eta < \frac{1}{2}$. *Thus* $\widehat{r} = \sum_{k=1}^{n} \mathbf{1}(\widehat{\lambda}_k \geq m^{-\eta})$ *consistently estimates the number of nonzero eigenvalues,* r, *as the number of features goes to infinity.*

Proof. To prove this theorem, we need to find the rate of convergence of the eigenvalue estimates $\hat{\lambda}_k$ to their true values λ_k. We begin by considering the distribution of the random matrix defined by,

$$\mathbf{W}_m = \underbrace{\frac{1}{m}\mathbf{G}\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}\mathbf{G}}_{i} + \underbrace{\frac{1}{m}\mathbf{G}\boldsymbol{\Gamma}^T\mathbf{U}}_{ii} + \underbrace{\frac{1}{m}\mathbf{U}^T\boldsymbol{\Gamma}\mathbf{G}^T}_{iii} + \underbrace{\frac{1}{m}\mathbf{U}^T\mathbf{U}}_{iv} - \underbrace{\widehat{\sigma}^2\mathbf{I}}_{v}$$

Let $\mathbf{K} = \boldsymbol{\Gamma}\mathbf{G}^T$ and first consider the term $\hat{\sigma}^2$:

$$\begin{aligned}
\hat{\bar{\sigma}}^2 &= \frac{1}{m}\sum_{i=1}^{m}\left[\sum_{j=1}^{n} x_{ij}^2 - (x_{ij} - \bar{x}_{i.})^2\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\left[\sum_{j=1}^{n}(k_{ij}+u_{ij})^2 - \left(k_{ij}+u_{ij} - (\bar{k}_{i.}+\bar{u}_{i.})\right)\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\left[\sum_{j=1}^{n} 2(k_{ij}+u_{ij})(\bar{k}_{i.}+\bar{u}_{i.}) - (\bar{k}_{i.}+\bar{u}_{i.})^2\right] \\
&= \frac{n}{m}\sum_{i=1}^{m}\bar{u}_{i.}^2 \\
&= \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n} u_{ij}u_{ik}
\end{aligned}$$

where the fourth step follows from assumption v. Rather than considering the joint distribution of $\mathbf{W}_m$ directly, we consider the set of components of terms $ii-v$ and use the Lindeberg-Feller Central Limit Theorem to demonstrate that they are asymptotically normal.

$$\mathbf{y}_i = \frac{1}{m}\left(k_{i1}\mathbf{u}_{i.}^T, \ldots, k_{in}\mathbf{u}_{i.}^T, u_{i1}\mathbf{u}_{i.}^T, \ldots, u_{in}\mathbf{u}_{i.}^T\right)^T$$

Then $\mathbf{E}[\mathbf{y}_i] = (0,\ldots,0,\sigma_i^2,0,\ldots,0,\sigma_i^2,0,\ldots,0,\sigma_i^2)^T$. Define $\mathbf{y}_i^* = \sqrt{m}(\mathbf{y}_i - \mathbf{E}[\mathbf{y}_i])$. We have $\mathbf{Cov}[\mathbf{y}_i^*] = \frac{1}{m}\mathbf{\Sigma}_i$. From assumptions 1-4, namely that the 2nd and 4th moments are positive and bounded, $\frac{1}{m}\sum_{i=1}^{m}\mathbf{\Sigma}_i \rightarrow \mathbf{\Sigma}$. Then from the Lindeberg-Feller CLT, $\sum_{i=1}^{m}\mathbf{y}_i^*$ is asymptotically normally distributed if the Lindeberg condition holds for every $\epsilon > 0$. To verify the Lindeberg condition consider:

$$\| \mathbf{y}_i^* \|^2 = \frac{1}{m}\left\{\sum_{j=1}^{n}(u_{ij}^2 - \sigma_i^2)^2 + \sum_{j=1}^{n}\sum_{k=1}^{n} k_{ij}^2 u_{ik}^2 + \sum_{k\neq j} u_{ij}^2 u_{ik}^2\right\}$$

Let $Z_i = \left\{\sum_{j=1}^{n}(u_{ij}^2-\sigma_i^2)^2 + \sum_{j=1}^{n}\sum_{k=1}^{n} k_{ij}^2u_{ik}^2 + \sum_{k\neq j} u_{ij}^2u_{ik}^2\right\}\mathbf{1}(\| \mathbf{y}_i^* \|^2 > \epsilon)$; we need to show that $\mathbf{E}[Z_i] \rightarrow 0$ for every i. Note that Z_i is only nonzero when

$$\frac{1}{m}\left\{\sum_{j=1}^{n}(u_{ij}^2 - \sigma_i^2)^2 + \sum_{j=1}^{n}\sum_{k=1}^{n} k_{ij}^2 u_{ik}^2 + \sum_{k\neq j} u_{ij}^2 u_{ik}^2\right\} > \epsilon$$

, an event that has probability zero as $m \to \infty$, so $Z_i \to_P 0$. It is also clear that $|Z_i| \leq m \parallel \mathbf{y}_i^* \parallel^2$ and $\mathbf{E}\{m \parallel \mathbf{y}_i^* \parallel^2\} < \infty$ by assumptions 1-3. Thus by the dominated convergence theorem $\mathbf{E}[Z_i] \to_P 0$ for each i and hence for every $\epsilon > 0$,

$$\sum_{i=1}^{m} \mathbf{E}\{\parallel \mathbf{y}_i^* \parallel \mathbf{1}(\parallel \mathbf{y}_i^* \parallel > \epsilon)\} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{E}\{Z_i\} \to_P 0$$

Since the Lindeberg condition is satisfied $\sum_{i=1}^{m} \mathbf{y}_i^*$ is asymptotically normally distributed. Thus, since $\mathrm{vec}(\mathbf{W}_\mathrm{m}) = \mathrm{g}(\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathbf{y}_\mathrm{i}) + \mathrm{vec}\left(\frac{1}{\mathrm{m}}\mathbf{G}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{G}\right)$, where the $\mathrm{vec}(\cdot)$ concatenates the columns of a matrix and g is a continuous function, we have by the Delta method:

$$\sqrt{m}\left(\mathrm{vec}(\mathbf{W}_\mathrm{m}) - \mathrm{vec}\left(\frac{1}{\mathrm{m}}\mathbf{G}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{G}\right)\right) \to \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_\mathrm{w})$$

so $\sqrt{m}(\mathbf{W}_m - \mathbf{S}_m) = O_P(1)$. Since $\lambda_r - \lambda_{r+1} = c > 0$ and $\mathbf{W}_n$ is symmetric and real, by Theorem 4.2 of Eaton and Tyler [28],

$$\begin{aligned} \sqrt{m}\left((\hat{\lambda}_1, \ldots, \hat{\lambda}_n)^T - (\lambda_1, \ldots, \lambda_n)^T\right) &= O_P(1) \\ \implies \sqrt{m}\hat{\lambda}_k &= \sqrt{m}\lambda_k + O_P(1) \quad \forall k \end{aligned} \tag{7.2}$$

Now consider:

$$\mathbf{1}(\hat{\lambda}_k \geq m^{-\eta}) = \mathbf{1}(m^\eta \hat{\lambda}_k \geq 1)$$

when $\lambda_k = 0$, $m^\eta \hat{\lambda}_k = m^{\eta - \frac{1}{2}}\left(m^{\frac{1}{2}}\hat{\lambda}_k\right) = m^{\eta-\frac{1}{2}} O_P(1) = o_P(1)$, where the last step follows from $\eta < \frac{1}{2}$. Thus for $\lambda_k = 0$, $\mathbf{1}(m^\eta \hat{\lambda}_k \geq 1) \to_P 0$. When $\lambda_k > 0$, $m^\eta \hat{\lambda}_k = m^{\eta-\frac{1}{2}}\left(m^{\frac{1}{2}}\hat{\lambda}_k\right) = m^\eta \lambda_k + o_P(1) \to \infty$. So when $\lambda_k > 0$, $\mathbf{1}(m^\eta \hat{\lambda}_k \geq 1) \to_P 1$, as required.

□

To evaluate the relative performance of the three approaches for estimating surrogate variables, we performed a simple simulation study. The data for each feature

in each study follow the model:

$$\begin{aligned} x_{ij} &= \beta_{1i} f_{1j} + \beta_{2i} f_{2i} + u_{ij} \\ u_{ij} &\sim \mathrm{N}(0, \sigma_{\mathrm{i}}^2) \end{aligned} \tag{7.3}$$

where we allowed the number of genes and arrays to vary as described in Table 7.1. For each value of n the two factors were fixed as:

$$f_{1j} = \begin{cases} -\frac{1}{2} & j = 1, \ldots, \frac{n}{2} \\ \frac{1}{2} & j = \frac{n}{2} + 1, \ldots, n \end{cases}$$

and

$$f_{2j} = \begin{cases} -\frac{1}{2} & j = 1, 3, \ldots, n-1 \\ \frac{1}{2} & j = 2, 4, \ldots, n \end{cases}$$

and the population parameters $\mu_i, \beta_{1i}, \beta_{2i}$, and σ_i^2 are simulated initially and then held fixed, in order to mimimic a hypothetical population. The population level parameters are drawn from the following distributions.

$$\begin{aligned} \mu_i &= 0 \\ \beta_{ki} &\overset{i.i.d.}{\sim} \begin{cases} N(0,1) & i = 1, \ldots, 1000 \end{cases} \\ \sigma_i^2 &\overset{i.i.d.}{\sim} \mathrm{Gamma}(1,2) \end{aligned}$$

The parameter α for the Buja and Eyuboglu Algorithm was fixed a 0.05 and the parameter η from Theorem 8 was fixed at $\frac{1}{3}$. Each algorithm was applied to 100 simulated expression data sets, simulated according to the sampling scheme outline above. Table 7.1 presents the results of this study. The number of samples, m, and the number of features, n, varied across simulations, but the number of unmodeled factors was fixed at two.

The Bai and Ng algorithm performs well when both the number of samples and the number of features is large, as one would expect. However, for small sample sizes, the number of signficant factors is severely over estimated. The estimator defined in

	Algorithm		
(m,n)	Bai and Ng	Buja and Eyuboglu ($\alpha = 0.05$)	Theorem 8 ($\eta = \frac{1}{3}$)
(100,10)	9.0 (0.0)	2.0 (0.0)	3.54 (0.8)
(1,000,10)	9.0 (0.0)	2.0 (0.0)	2.84 (0.72)
(10,000,10)	9 (0.0)	2.0 (0.0)	2.1 (0.3)
(100,20)	10.5 (8.5)	2.0 (0.0)	6.2 (1.3)
(500,200)	2.0 (0.0)	2.0 (0.0)	61.9 (4.8)
(1000,500)	2.0 (0.0)	2.0 (0.0)	153.67 (8.6)

Table 7.1: Results from 100 simulated gene expression studies with two unmodeled factors for different numbers of samples, n and features, m. The average estimated number of factors is shown for the three estimation approaches. Monte-Carlo standard errors are shown in parentheses.

Theorem 8 behaves well as the number of features grows for a fixed sample size, but does not behave well when both the number of features and the number of samples gets large. The permutation algorithm of Buja and Eyuboglu is highly accurate regardless of the number of features and the number of samples obtained. In fact, among the 100 simulated studies for each (m,n), the permutation algorithm always selected the correct number of factors. We chose to use the Buja and Eyuboglu factor selection criteria in the Algorithms proposed in this dissertation because it seems to perform as well as the asymptotic estimators for large sample sizes, but appears to have much better finite sample performance.

7.4 Residual PCA Consistently Estimates Orthogonal Factors

In general, it will not always be possible to obtain consistent estimates of the surrogate variables in any fixed microarray study. This can easily be seen from model 3.6, if the surrogate variables $\mathbf{G}$ have the exact column space of the primary variables $\mathbf{S}$, then it

is impossible to distinguish the signal due to the primary variable and the signal due to the unmodeled factors. Here we investigate one special case where it is possible to consistently estimate the surrogate variables as the number of features grows large and the number of samples stays fixed.

Theorem 9. *Suppose that the data from a high-throughput experiment can be modeled as in equation 3.6 and that the primary variables and the unmodeled factors are orthogonal. Then if the assumptions of Theorem 7 hold, the column space of the surrogate variables estimated from the residual PCA algorithm based on Theorems 7 and 8 converges to the column space of the true surrogate variables as* $m \to \infty$.

Proof. The residual PCA algorithm estimates surrogate variables on the basis of the residual matrix,

$$\begin{aligned} \widehat{\mathbf{E}} &= \mathbf{X} - \widehat{\boldsymbol{\beta}}\mathbf{S}^T \\ &= (\boldsymbol{\beta}\mathbf{S}^T + \boldsymbol{\Gamma}\mathbf{G}^T + \mathbf{U})(\mathbf{I} - \mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}) \\ &= \boldsymbol{\Gamma}\mathbf{G}^T + \mathbf{U}\mathbf{P}_f \end{aligned} \tag{7.4}$$

where the last step follows from the orthogonality of $\mathbf{S}$ and $\mathbf{G}$ and $\mathbf{P}_f$ is the projection matrix onto the linear space perpendicular to $\mathbf{S}^T$. The residuals $\mathbf{U}\mathbf{P}_f$ are still independent across rows and still have expectation zero. By assumption, model 7.4 satisfies the requirements for Theorems 8 and 7 to hold. Thus, if the number of factors, $\widehat{r}$, is selected by the algorithm in model 8 and the factors are estimated according to the algorithm of Theorem 7, the surrogate variables estimates consistently estimate the column space of the true surrogate variables. □

Of course the residual PCA algorithm we proposed in Chapter 4 is slightly different than the PCA algorithm in Theorem 9. In Theorem 9 we estimate the number of significant surrogate variables using the result of Theorem 8, rather than the permutation algorithm of Buja and Eyuboglu. However, the results presented in Table

7.1 indicate that the Buja and Eyuboglu algorithm is very similar to the behavior of the estimate from Theorem 8 as the number of features grows. The estimates of the surrogate variables are also based on the algorithm from Theorem 7, rather than the estimates of the right singular vectors from the principal components analysis. However, the estimates from the two approaches are mathematically identical.

7.5 Conjectures Regarding a General Estimate for Surrogate Variables

In the previous section we showed that it is possible to consistently estimate surrogate variables when the unmodeled factors are strictly orthogonal to the primary variables. The assumption of orthogonality may be appropriate in a randomized study with a large sample size, where the configuration of the unmodeled factors should be randomized with respect to the primary variables. However, many high-throughput studies have small sample sizes, are observational, or both.

In Chapter 4 we showed that the subset PCA algorithm and particularly the reduced subset PCA algorithm showed surprisingly good behavior even when the unmodeled factor affected an overlapping subset of features and was correlated with the primary variable. As the overlap and the correlation increase, the surrogate variable estimates increasingly are biased in the direction of the signal from the primary variable.

In general, it is not possible to perfectly distinguish subsets of features that are affected by the primary variable and the unmodeled factor. However, the power to test for associations with either variable increases with increasing sample size. We postulate that in studies with both a large sample size and a large number of features, it may be possible to theoretically disentangle even highly correlated primary variables and unmodeled factors.

7.6 Maximum Likelihood

The algorithms presented in this dissertation for estimating surrogate variables are closely related to maximum likelihood estimates under simplifying assumptions. For clarity, we will assume that r, the true number of surrogate variables to included in the model is known.

7.6.1 Residual PCA Estimates are Maximum Likelihood Estimates

Suppose that the data from a high-throughput experiment follow model 3.5, where $u_{ij} \overset{i.i.d.}{\sim} N(0,1)$ and $\mu, \boldsymbol{\beta}, \Gamma$ and $\mathbf{G}$ are assumed to be unknown parameters. If the dimension r, of the matrix $\mathbf{G}$ is assumed to be known, then the likelihood based on the feature level data is as follows.

$$L(\boldsymbol{\beta}, \Gamma, \mathbf{G}|\mathbf{X}, \mathbf{S}) = \prod_{i=1}^{m} \frac{1}{(2\pi)^{n/2}} \exp\left\{ -\frac{\left(\mathbf{x}_{i\cdot} - \mathbf{S}\boldsymbol{\beta}_{i\cdot}^T - \mathbf{G}\boldsymbol{\Gamma}_{i\cdot}^T\right)^T \left(\mathbf{x}_{i\cdot} - \mathbf{S}\boldsymbol{\beta}_{i\cdot}^T - \mathbf{G}\boldsymbol{\Gamma}_{i\cdot}^T\right)}{2} \right\}$$

So the log likelihood is:

$$\begin{aligned} \ell(\boldsymbol{\beta}, \Gamma, \mathbf{G}|\mathbf{X}, \mathbf{S}) &= \sum_{i=1}^{m} -\frac{n}{2}\log(2\pi) - \frac{\left(\mathbf{x}_{i\cdot} - \mathbf{S}\boldsymbol{\beta}_{i\cdot}^T - \mathbf{G}\boldsymbol{\Gamma}_{i\cdot}^T\right)^T \left(\mathbf{x}_{i\cdot} - \mathbf{S}\boldsymbol{\beta}_{i\cdot}^T - \mathbf{G}\boldsymbol{\Gamma}_{i\cdot}^T\right)}{2} \\ &\propto \sum_{i=1}^{m} -\left(\mathbf{x}_{i\cdot} - \mathbf{S}\boldsymbol{\beta}_{i\cdot}^T - \mathbf{G}\boldsymbol{\Gamma}_{i\cdot}^T\right)^T \left(\mathbf{x}_{i\cdot} - \mathbf{S}\boldsymbol{\beta}_{i\cdot}^T - \mathbf{G}\boldsymbol{\Gamma}_{i\cdot}^T\right) \\ &= -\left|\left|\mathbf{X} - \boldsymbol{\beta}\mathbf{S}^T - \Gamma\mathbf{G}^T\right|\right| \end{aligned} \tag{7.5}$$

Maximizing the log likelihood is accomplished by minimizing the distance between $\mathbf{X}$ and $\boldsymbol{\beta}\mathbf{S}^T + \Gamma\mathbf{G}^T$ where distance is measured in the Frobenius norm $||\cdot||$. Since $\boldsymbol{\beta}, \Gamma$, and $\mathbf{G}$ are assumed to be fixed and unknown, the maximization must occur simultaneously over the linear term $\boldsymbol{\beta}\mathbf{S}^T$ and the bilinear term $\Gamma\mathbf{G}^T$. Gabriel showed that the maximum of 7.5 can be calculated in two steps [34].

1. Calculate the least squares estimator $\widehat{\boldsymbol{\beta}}$, minimizing $\left|\left|\mathbf{X} - \boldsymbol{\beta}\mathbf{S}^T\right|\right|$.

2. Calculate the residual matrix $\widehat{\mathbf{E}} = \mathbf{X} - \widehat{\boldsymbol{\beta}}\mathbf{S}^T$ and take the singular value decomposition $\widehat{\mathbf{E}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. The maximum likelihood estimator of $\mathbf{G}$ is $\mathbf{V}_{\cdot 1:r}$, the first r columns of $\mathbf{V}$.

For a fixed r, the maximum likelihood surrogate variable estimates are exactly the residual PCA surrogate variable estimates. Although we have not assumed that the primary variables and the unmodeled factors are orthogonal, the maximum likelihood estimates for these factors turn out to be orthogonal regardless of our assumptions. In Chapter 4 we showed that including the surrogate variables estimated from the residual PCA algorithm often produced biased null distributions and error rate estimates, and the same difficulties are obviously associated with the maximum likelihood estimates.

7.6.2 Connections between Subset PCA and An Expectation Maximization Approach

The true model for the data in each row of $\mathbf{X}$ will include only a subset of the columns of $\mathbf{G}$. It is also true that only a subset of the features show true association with the primary variable $\mathbf{S}$. Suppose that we augment the observed data $\mathbf{X}, \mathbf{S}$ with random variables a_i and b_{ik} where:

$$a_i = \begin{cases} 1 & \text{if } ||\boldsymbol{\beta}_{\mathrm{i}\cdot}|| > 0 \\ 0 & \text{else} \end{cases}$$

$$b_{ik} = \begin{cases} 1 & \text{if } |\gamma_{\mathrm{ik}}| > 0 \\ 0 & \text{else} \end{cases}$$

a_i is an indicator that the primary variable is associated with feature i and b_{ik} is an indicator that unmodeled factor k is associated with feature i. We will assume in the development that follows that each feature is affected by at most one unmodeled factor. Coupled with the assumption that $u_{ij} \overset{i.i.d.}{\sim} N(0,1)$, the complete data likelihood

can then be written as follows

$$
\begin{aligned}
& L(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \mathbf{G} | \mathbf{X}, \mathbf{Y}, \mathbf{a}, \mathbf{b}) \\
= & \prod_{i=1}^{m} \prod_{k=1}^{r} \left[\frac{1}{(2\pi)^{n/2}} \exp \left\{ - \frac{\left(\mathbf{x}_{i\cdot} - \mathbf{S}\boldsymbol{\beta}_{i\cdot}^T - \mathbf{G}_{\cdot k} \Gamma_{ik} \right)^T \left(\mathbf{x}_{i\cdot} - \mathbf{S}\boldsymbol{\beta}_{i\cdot}^T - \mathbf{G}_{\cdot k} \Gamma_{ik} \right)}{2} \right\} \pi_{11}(k) \right]^{a_i b_{ik}} \\
\times & \left[\frac{1}{(2\pi)^{n/2}} \exp \left\{ - \frac{\left(\mathbf{x}_{i\cdot} - \mathbf{G}_{\cdot k} \Gamma_{ik} \right)^T \left(\mathbf{x}_{i\cdot} - \mathbf{G}_{\cdot k} \Gamma_{ik} \right)}{2} \right\} \pi_{01}(k) \right]^{(1-a_i) b_{ik}} \\
\times & \left[\frac{1}{(2\pi)^{n/2}} \exp \left\{ - \frac{\left(\mathbf{x}_{i\cdot} - \mathbf{S}\boldsymbol{\beta}_{i\cdot}^T \right)^T \left(\mathbf{x}_{i\cdot} - \mathbf{S}\boldsymbol{\beta}_{i\cdot}^T \right)}{2} \right\} \pi_{10}(k) \right]^{a_i (1-b_{ik})} \\
\times & \left[\frac{1}{(2\pi)^{n/2}} \exp \left\{ - \frac{\left(\mathbf{x}_{i\cdot} \right)^T \left(\mathbf{x}_{i\cdot} \right)}{2} \right\} \pi_{00}(k) \right]^{(1-a_i)(1-b_{ik})}
\end{aligned}
\tag{7.6}
$$

where $\pi_{11}(k) = \mathbf{E}[a_i b_{ik}]$, $\pi_{01}(k) = \mathbf{E}[(1 - a_i) b_{ik}]$,$\pi_{10}(k) = \mathbf{E}[a_i (1 - b_{ik})]$, and $\pi_{00}(k) = \mathbf{E}[(1 - a_i)(1 - b_{ik})]$. Obviously, assuming each feature is affected by at most one unmodeled factor is somewhat unrealistic, nevertheless a connection exists between this simplified model and the subset PCA approaches proposed in Chapter 4. The results we provide here can easily be extended to the case where multiple unmodeled factors may affect each feature, by taking the interior product in equation 7.6 over all possible subsets of the r columns of $\mathbf{G}$.

The complete data log-likelihood for model 7.6

$$
\begin{aligned}
& \ell(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \mathbf{G} | \mathbf{X}, \mathbf{Y}, \mathbf{a}, \mathbf{b}) \\
= & \sum_{i=1}^{m} \sum_{k=1}^{r} a_i b_{ik} \log(\pi_{11}(k)) + (1 - a_i) b_{ik} \log(\pi_{01}(k)) \\
& + a_i (1 - b_{ik}) \left[\log(\pi_{10}(k)) + (1 - a_i)(1 - b_{ik}) \log(\pi_{00}(k)) \right] \\
& - a_i b_{ik} \left[\frac{\left(\mathbf{x}_{i\cdot} - \mathbf{S}\boldsymbol{\beta}_{i\cdot}^T - \mathbf{G}_{\cdot k} \Gamma_{ik} \right)^T \left(\mathbf{x}_{i\cdot} - \mathbf{S}\boldsymbol{\beta}_{i\cdot}^T - \mathbf{G}_{\cdot k} \Gamma_{ik} \right)}{2} \right] \\
& - (1 - a_i) b_{ik} \left[\frac{\left(\mathbf{x}_{i\cdot} - \mathbf{G}_{\cdot k} \Gamma_{ik} \right)^T \left(\mathbf{x}_{i\cdot} - \mathbf{G}_{\cdot k} \Gamma_{ik} \right)}{2} \right] \\
& - a_i (1 - b_{ik}) \left[\frac{\left(\mathbf{x}_{i\cdot} - \mathbf{S}\boldsymbol{\beta}_{i\cdot}^T \right)^T \left(\mathbf{x}_{i\cdot} - \mathbf{S}\boldsymbol{\beta}_{i\cdot}^T \right)}{2} \right] \\
& - (1 - a_i)(1 - b_{ik}) \left[\frac{\left(\mathbf{x}_{i\cdot} \right)^T \left(\mathbf{x}_{i\cdot} \right)}{2} \right]
\end{aligned}
$$

To estimate the parameters of this model we can apply the expectation-maximization algorithm [24]. The log-likelihood is linear in terms of the products $a_i b_{ik}$,$(1-a_i)b_{ik}$,$a_i(1-b_{ik})$, and $(1 - a_i)b_{ik}$, so the E-step of the EM algorithm proceeds by calculating the conditional probabilities of these terms given $\mathbf{X}$ and $\mathbf{S}$. For example:

$$
p_{11}(k, i) = \mathbf{E}[a_i b_{ik} | \mathbf{X}] = \frac{\mathbf{Pr}_{11}(\mathbf{x}_i | \boldsymbol{\beta}_{i\cdot}, \Gamma_{ik}, \mathbf{G}_{\cdot k}) \pi_{11}(k)}{\sum_{t=0}^{1} \sum_{s=0}^{1} \mathbf{Pr}_{ts}(\mathbf{x}_i | \boldsymbol{\beta}_{i\cdot}, \Gamma_{ik}, \mathbf{G}_{\cdot k}) \pi_{ts}(k)}
$$

where the conditional probabilities $\mathbf{Pr}_{11}(\mathbf{x}_i | \boldsymbol{\beta}_{i\cdot}, \Gamma_{ik}, \mathbf{G}_{\cdot k}) = \mathbf{Pr}(\mathbf{x}_i | \boldsymbol{\beta}_{i\cdot}, \Gamma_{ik}, \mathbf{G}_{\cdot k}, a_i, b_{ik})$, etc., are defined in terms of simple normal densities. The rest of the conditional expectations can be calculated analogously.

The M-step of the EM algorithm calculates maximum likelihood estimators for the marginal probabilities π_{ts} and the parameters $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$ and $\mathbf{G}$. The maximum likelihood estimators for the marginal probabilities are again straightforward to calculate.

$$
\widehat{\pi}_{11} = \frac{\sum_{i=1}^{m} \mathbf{E}[a_i b_{ik} | \mathbf{x}_i]}{m}
$$

Where the rest of the probabilities can be calculated analogously. The maximum likelihood estimators for the coefficients $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$ and $\mathbf{G}$ can then be calculated by minimizing

the following sum of matrix norms.

$$\sum_{k=1}^{r} ||\mathbf{P}_{11}(k)(\mathbf{X} - \boldsymbol{\beta}\mathbf{S}^T - \boldsymbol{\Gamma}_{\cdot k}\mathbf{G}_{\cdot k}^T)|| + ||\mathbf{P}_{01}(k)(\mathbf{X} - \boldsymbol{\Gamma}_{\cdot k}\mathbf{G}_{\cdot k}^T)|| + ||\mathbf{P}_{10}(k)(\mathbf{X} - \boldsymbol{\beta}\mathbf{S}^T)|| + ||\mathbf{P}_{00}(k)\mathbf{X}|| \quad (7.7)$$

where $P_{11}(k)$ is a diagonal matrix with diagonal elements $p_{11}(k,i) = \mathbf{E}[a_i b_{ik}|\mathbf{X}]$, and so forth. Minimizing equation 7.7 is a convex optimization with $4 \times r$ terms. The last term for each factor, $||\mathbf{P}_{00}(k)\mathbf{X}||$, does not affect the optimization and can be ignored. There is no analytic optimum for the remaining terms, although some computationally intensive iterative algorithms have been proposed for minimizing the first and the second term for each factor. However these algorithms are not guaranteed to converge to a global maximum [35].

Given the computational complexity and uncertainty involved in calculating the maximum likelihood estimates, it is infeasible in practice to apply this EM algorithm. However, the form of the optimization problem 7.7 provides some intuition into the subset PCA algorithms proposed in Chapter 4. Consider the case where $\boldsymbol{\beta}_{i\cdot} = \mathbf{0}$ for all i. In that case, minimizing equation 7.7 is equivalent to minimizing.

$$||\mathbf{P}(k)(\mathbf{X} - \boldsymbol{\Gamma}_{\cdot k}\mathbf{G}_{\cdot k}^T)||$$

for some weights $\mathbf{P}(k)$ defined in terms of conditional probabilities that the kth factor is included in the model for each feature. The solution to this problem is nearly equivalent to the subset PCA solution when the weight $p(k,i)$ equals one when feature i is selected for surrogate variable k in Algorithm 4. In the case where we make no assumption about the value of $\boldsymbol{\beta}_{i\cdot}$, the minimization problem is more closely connected to the subset PCA approach. If we let the weights $\mathbf{P}_{ts}(k)$ be zero or one based on the inclusion of features in the estimation in Algorithm 5, we get nearly equivalent results from the expectation maximization approach and the subset PCA algorithm.

Chapter 8

CONCLUDING REMARKS

8.1 Summary of Present Work

This dissertation has focused on the problem of noise-dependence in high-dimensional data. Through simulations we showed that noise dependence can cause both bias and variability in (1) the ranking of features according to their association with a variable of interest and (2) estimating error rates for lists of features that are called significant. These simulation results indicate that, if unaccounted for, noise dependence can jeopardize the inference drawn from high-throughput experiments in molecular biology and may lead to wasted resources pursuing inconsequential features in follow-up studies.

We proposed a completely general framework for noise-dependence when the data from a high-throughput experiment are normally distributed and extended the framework to more general forms of noise-dependence. Surprisingly, in a high-dimensional experiment with $m \times n$ observations, where $m \gg n$ we showed that conditioning on an appropriate $n \times n$ matrix, is sufficient to eliminate all dependence across features. We defined estimators of the column space of this matrix to be surrogate variables for the high-dimensional problem. Surrogate variable estimates were proposed and explored through the use of simulation.

As a proof of concept, we applied our surrogate variable approach to the analysis of three example gene expression studies. These studies cover the range of the most popular expression studies: the genetic dissection of gene expression variation, disease class differential expression analysis, and timecourse differential expression analysis. Using these studies, we showed that surrogate variables accurately capture known

sources of noise dependence using only the expression data. We also applied surrogate variables to the analysis of data from the Trauma Glue Grant and showed that surrogate variable adjusted analyses produced results that were more stable both in the global significance and biological conclusions.

The theoretical results in this dissertation draw a connection between surrogate variables and multiple testing dependence. Namely, we showed that under certain model assumptions, multiple testing dependence is eliminated by condition on appropriately estimated surrogate variables. Combined with the results of Storey, Taylor, and Seigmund, the results of this dissertation provide quite general conditions for conservative estimation of false discovery rates. We also explored the theoretical properties of principal components estimates asymptotically in the number of features. We showed that for orthogonal primary variables and unmodeled factors it is possible to consistently estimate surrogate variables as the number of features grows large.

8.2 Future Work

We have focused on the problem of defining a framework for noise-dependence which also explains the well-studied phenomena of multiple testing dependence. We have shown that by accurately estimating surrogate variables, this dependence can be eliminated. The algorithms and theory proposed in this dissertation form the basis for creating new and more accurate estimates of the surrogate variables. An obvious extension of this work is to explore the theoretical properties of subset PCA algorithms under more relaxed assumptions about the relative configuration of the primary variable and the unmodeled factor. In addition to these obvious theoretical extensions, three substantial areas for future research are as follows.

8.2.1 Generalized Additive Models

The surrogate variable ideas proposed in this dissertation have been motivated by the linear model:

$$\mathbf{X} = \boldsymbol{\beta}\mathbf{S}^T + \boldsymbol{\Gamma}\mathbf{G}^T + \mathbf{U}$$

which can also be written in the following way.

$$\mathbf{E}[\mathbf{X}|\mathbf{S}, \mathbf{G}] = \boldsymbol{\beta}\mathbf{S}^T + \boldsymbol{\Gamma}\mathbf{G}^T$$

This type of simple linear model is an obvious choice for feature data on a quantitative scale, and is the natural parameterization for normal data. However, the feature data from a high-throughput study is not always quantitative; it is not uncommon to collect binary or ordered categorical feature data. If we assume that the feature data are drawn from an exponential family distribution, then we can write a generalized linear model for the data [59].

$$g(\mathbf{E}[\mathbf{X}|\mathbf{S}, \mathbf{G}]) = \boldsymbol{\beta}\mathbf{S}^T + \boldsymbol{\Gamma}\mathbf{G}^T \tag{8.1}$$

There are a number of standard choices for $g(\cdot)$, depending on the distribution of the feature level data. Since neither $\boldsymbol{\Gamma}$ or $\mathbf{G}$ is known in practice, model 8.1 is an example of a generalized linear-bilinear model [35].

A natural extension of the ideas proposed in this dissertation is to extend surrogate variable estimation to the whole class of generalized linear models. However, substantial challenges remain in calculating estimates from model 8.1. Iterative algorithms have been proposed for calculating estimates for generalized bilinear models, of the form:

$$g(\mathbf{E}[\mathbf{X}|\mathbf{S}, \mathbf{G}]) = \boldsymbol{\Gamma}\mathbf{G}^T$$

For example, a generalization of the so-called criss-cross regression for normally distributed data has been proposed for estimating the parameters of 8.1 [35]. The algorithm iteratively assumes the current value of $\boldsymbol{\Gamma}$ or $\mathbf{G}$ is known and fits a generalized

linear model for the other by iteratively re-weighted least squares. For the generalized bilinear model, this algorithm is guaranteed to converge, but not necessarily to the global maximum likelihood estimates. One goal is to identify estimators for model 8.1 that are computationally efficient, eliminate or reduce noise dependence across features and reduce or eliminate multiple testing dependence.

8.2.2 Prediction

One important goal in high-throughput studies is to identify relatively small subsets of the features that distinguish between groups or classes. For example, a number of microarray studies have focused on identifying gene expression signatures that distinguish tumor subtypes [38, 64]. In order for predictors to be useful in clinical practice, the feature level data must consistently distinguish the classes across arrays.

As we have shown, noise-dependence confounds the relationship between the primary variable, in this case the class variable, and the feature level data. We have also shown that in general, the confounding is mediated through unmodeled factors that act on the feature data and are correlated with the primary variable. By identifying and eliminating those factors subject to the effects of unmodeled factors before creating class predictors, it is likely that accuracy and consistency can be approved.

As an example, consider the case of tumor class prediction with microarray data. One possible algorithm for identifying predictors is as follows.

1. Estimate the surrogate variables for the expression data.

2. Identify the features associated with each surrogate variable and remove them from consideration as predictors.

3. Select a subset of the remaining features that consistently predict the class variable.

The idea is that features highly associated with a surrogate variable and, by proxy, an unmodeled factor will not show consistent expression patterns across repeated samples. Furthermore, for a single patient, it will not be possible to estimate the effect of the unmodeled factor, since it is unknown. Therefore, it makes sense to only predict the class variable from expression values that are not affected by common sources of noise-dependence. These features will show consistent expression patterns across individuals, and will produce more accurate predictors in general.

8.2.3 Identifying Clinical Variables that Explain Surrogate Variables

Surrogate variable analysis is arguably most useful in large clinical microarray studies in humans, where it is impossible to control all of the genetic and environmental factors that influence expression. These large studies typically measure hundreds or thousands of clinical covariates that may in some way be related to expression. Out of these hundreds of covariates a handful is likely to have a global impact on gene expression.

Rather than test each feature for association with each gene's expression, a more efficient approach may be to identify the global expression trends and attempt to correlate only the important trends with each of the measured covariates. Suppose we have an estimate of the surrogate variables $\widehat{\mathbf{G}}$ for the high dimensional data set $\mathbf{X}$, then one way to do this would be to identify the subset of clinical variables $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_k)$ that minimizes the matrix norm $\|\widehat{\mathbf{G}} - \mathbf{\Lambda H}\|$. In other words, identify the linear combination of a subset of clinical predictors that best fits the surrogate variables. This type of approach is much more efficient than testing all covariate feature associations. The resulting set of covariates also represent a joint clinical model for the major sources of variation in the gene expression data set.

BIBLIOGRAPHY

[1] A. A. Ahmed, M. Vias, N. G. Iyer, C. Caldas, and J. D. Brenton. Microarray segmentation methods significantly influence data precision. *Nucleic Acids Res*, 32(5):e50, 2004.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell.* Garland Science Textbooks, 2002.

[3] A. Alizadeh, M. Eisen, D. Botstein, P. O. Brown, and L. M. Staudt. Probing lymphocyte biology by genomic-scale gene expression analysis. *J Clin Immunol*, 18:373–9, 1998.

[4] A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, and *et al.* Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–11, 2000.

[5] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U.S.A.*, 97:10101–6, 2000.

[6] T. W. Anderson and Y. Amemiya. The asymptotic normal distribution of estimators in factor analysis under general conditions. *Ann Stat*, 16:759–71, 1988.

[7] J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70:191–221, 2002.

[8] R. Bellman. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, 1961.

[9] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate-a practical and powerful approach to multiple testing. *J Roy Stat Soc B*, 57:289–300, 1995.

[10] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*, 29:1165–88, 2001.

[11] R. B. Brem, J. D. Storey, J. Whittle, and L. Kruglyak. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature*, 436:701–703, 2005.

[12] R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296:752–755, 2002.

[13] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nat Genet*, 21(1 Suppl):33–7, 1999.

[14] A. Buja and N. Eyuboglu. Remarks on parallel analysis. *Multivariate Behav*, 27:509–40, 1992.

[15] C. S. Carlson, M. A. Eberle, L. Kruglyak, and D. A. Nickerson. Mapping complex disease loci in whole-genome association studies. *Nature*, 429(6990):446–52, 2004.

[16] R. B. Cattell. The scree test for the number of factors. *Multivariate Behav Res*, 1:245–76, 1966.

[17] J. P. Cobb, M. Mindrinos, C. Miller-Graziano, S. E. Calvano, H. V. Baker, and *et al.* Application of genome-wide expression analysis to human health and disease. *Proc Natl Acad Sci, U.S.A.*, 102:4801–6, 2005.

[18] G. Connor and R. A. Korajczyk. A test for the number of factors in an approximate factor model. *Journal of Finance*, 48:1263–92, 1993.

[19] I. H. G. S. Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, 2004.

[20] L. M. Cope, R. A. Irizarry, H. A. Jaffee, Z. Wu, and T. P. Speed. A benchmark for affymetrix genechip expression measures. *Bioinformatics*, 20(3):323–31, 2004.

[21] F. Crick. Central dogma of molecular biology. *Nature*, 227:561–63, 1970.

[22] X. Cui, J. T. G. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6:59–75, 2005.

[23] S. S. Dave, G. Wright, B. Tan, A. Rosenwald, R. D. Gascoyne, and *et al.* Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *New Engl J Med*, 351:2159–69, 2004.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc B*, 39:1–38, 1977.

[25] Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University and Novartis Institutes of BioMedical Research, R. Saxena, B. F. Voight, V. Lyssenko, and *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–6, 2007.

[26] M. O. Dorschner, M. Hawrylycz, R. Humbert, J. C. Wallace, A. Shafer, J. Kawamoto, J. Mack, R. Hall, J. Goldy, P. J. Sabo, A. Kohli, Q. Li, M. McArthur, and J. A. Stamatoyannopoulos. High-throughput localization of functional elements by quantitative chromatin profiling. *Nat Methods*, 1(3):219–25, 2004.

[27] S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

[28] M. L. Eaton and D. E. Tyler. On wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Ann Stat*, 19:260–271, 1991.

[29] W. F. Eddy and J. A. Hartigan. Variance of the number of false discoveries. *Ann Stat*, 5:370–4, 1977.

[30] B. Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J Am Stat Assoc*, 99:96–104, 2004.

[31] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of Computational Biology*, 96:1151–60, 2001.

[32] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns michael b. eisen. *Proc Natl Acad Sci USA*, 95(25):14863–68, 1998.

[33] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, 3 edition, 1968.

[34] K. R. Gabriel. Least squares approximation of matrices by additive and multiplicative models. *J Roy Stat Soc B*, 40:186–96, 1978.

[35] K. R. Gabriel. Generalized bilinear regression. *Biometrika*, 85:689–700, 1998.

[36] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, and *et al.* Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–37, 1999.

[37] K. L. Gunderson, F. J. Steemers, G. Lee, L. G. Mendoza, and M. S. Chee. A genome-wide scalable snp genotyping assay using microarray technology. *Nat Genet*, 37(5):549–54, 2005.

[38] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, M. Raffeld, and et *al.* Gene-expression profiles in hereditary breast cancer. *New Engl J Med*, 344:539–548, 2001.

[39] I. Hedenfalk, M. Ringer, A. Ben-Dor, Z. Yakhini, Y. Chen, and *et al.* Molecular classification of familial non-brca1/brca2 breast cancer. *Proc Natl Acad Sci USA*, 100(5):2532–37, 2003.

[40] M. Henriksen and J. R. Isbell. On the continuity of the real roots of an algebraic equation. *Proc Amer Math Soc*, 4:431–4, 1953.

[41] T. R. Hughes, M. Mao, A. R. Jones, J. Burchard, M. J. Marton, and *et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol*, 19(4):342–7, 2001.

[42] T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–34, 2001.

[43] T. Ideker, V. Thorsson, A. Siegel, and L. Hood. Testing for differentially-expressed genes by maximum-likelihood analysis of DNA microarray data. *Journal of Computational Biology*, 7:805–17, 2000.

[44] Ingenuity Systems. Ingenuity pathway analyis proprietary software, 2007.

[45] J. Inglese, D. S. Auld, A. Jadhav, R. J. Johnson, A. Simeonov, and *et al.* Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Nat. Acad. Sci. U.S.A.*, 103:11473–11478, 2006.

[46] R. Jaenisch and A. Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet*, 33:245–54, 2003.

[47] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann Statist*, 29:295–327, 2001.

[48] L. Klebanov, C. Jordan, and A. Yakovlev. A new type of stochastic dependence revealed in gene expression data. *Stat Appl Genet Mo B*, 5:7, 2006.

[49] W. J. Krzanowski. The algebraic basis of classical multivariate methods. *The Statistician*, 20:51–61, 1971.

[50] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, and *et al.* The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313:1929–35, 2006.

[51] J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3:e161, 2007.

[52] E. L. Lehmann. *Testing Statistical Hypotheses.* Springer, 1997.

[53] A. Lewbel. The rank of demand systems: Theory and nonparameteric estimation. *Econometrika*, 59:711–30, 1991.

[54] R. J. Lipshutz, D. Morris, M. Chee, E. Hubbell, M. J. Kozal, N. Shah, N. Shen, R. Yang, and S. P. Fodor. Using oligonucleotide probe arrays to access genetic diversity. *BioTechniques*, 19(3):442–7, 1995.

[55] R. Liu, X. Wang, G. Y. Chen, P. Dalerba, A. Gurney, and *et al.* The prognostic role of a gene signature from tumorigenic breast-cancer cells. *New Engl J Med*, 356:217–26, 2007.

[56] G. MacBeath and S. L. Schreiber. Printing proteins as microarrays for high-throughput function determination. *Science*, 289(5485):1760–3, 2000.

[57] J. Marchini, L. Cardon, M. Phillips, and P. Donnelly. The effects of human population structure on large genetic association studies. *Nat Genet*, 36:512–7, 2004.

[58] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis.* Academic Press, 1997.

[59] P. McCullagh and J. Nelder. *Generalized Linear Models.* Chapman and Hall, 1989.

[60] M. Morley, C. M. Molony, T. Weber, J. L. Devlin, K. G. Ewens, and *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430:743–747, 2004.

[61] E. E. Ntzani and J. P. Ioannidis. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet*, 362:1434–9, 2003.

[62] A. Owen. Variance of the number of false discoveries. *J Roy Stat Soc B*, 67:411–26, 2005.

[63] A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. P. Fodor. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA*, 91(11):5022–6, 1994.

[64] C. M. Perrou, S. S. Jeffrey, M. van de Rijn, C. A. Rees, M. B. Eisen, and *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci, U.S.A.*, 96:9212–17, 1999.

[65] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38:904–909, 2006.

[66] X. Qui, L. Klebanov, and A. Y. Yakovlev. Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. *Stat Appl Genet Mo B*, 4:34, 2005.

[67] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, and *et al.* Genome-wide location and function of DNA binding proteins. *Science*, 290:2306–09, 2000.

[68] D. R. Rhodes and A. M. Chinnaiyan. Integrative analysis of the cancer transcriptome. *Nat Genet*, 37:S31–7, 2005.

[69] G. E. Rodwell, R. Sonu, J. M. Zahn, J. Lund, J. Wilhelmy, and *et al.* A transcriptional profile of aging in the human kidney. *PLoS Biol*, 2(12):2191–2201, 2004.

[70] P. Rosenbaum. *Observational Studies.* Springer Series in Statistics, 2 edition, 2002.

[71] E. E. Schadt, S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, and *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422:297–302, 2003.

[72] J. Schäfer and K. Strimmer. A shrinkage approach to large covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mo B*, 4:32, 2005.

[73] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 207:467–70, 1995.

[74] B. Schweitzer, S. Wiltshire, J. Lambert, S. O'Malley, K. Kukanskis, Z. Zhu, S. F. Kingsmore, P. M. Lizardi, and D. C. Ward. Inaugural article: immunoassays with rolling circle DNA amplification: a versatile platform for ultrasensitive antigen detection. *Proc Natl Acad Sci USA*, 97(18):10113–9, 2000.

[75] G. A. F. Seber and A. J. Lee. *Linear Regression Analysis.* Wiley Series in Probability and Statistics, 2 edition, 2003.

[76] G. A. F. Seber and A. J. Lee. *Linear Regression Analysis.* Wiley, 2003.

[77] M. J. Shaw, C. Subrmaniam, G. W. Tan, and M. E. Wedge. Knowledge management and data mining for marketing. *Decision Support Systems*, 31:127–37, 2001.

[78] S. Solinas-Toldo, S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, and *et al.* Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. *Genes, Chromosomes and Cancer*, 20:399–407, 1997.

[79] J. D. Storey. A direct approach to false discovery rates. *J Roy Stat Soc B*, 64:479–98, 2002.

[80] J. D. Storey. The optimal discovery procedure: A new approach to simultaneous significance testing. *J Roy Stat Soc B*, 69:347–68, 2007.

[81] J. D. Storey, J. Y. Dai, and J. T. Leek. The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics*, 8:414–32, 2007.

[82] J. D. Storey, J. Madeoy, J. L. Strout, M. Wurfel, J. Ronald, and *et al.* Gene expression variation within and among human populations. *Am J Hum Genet*, 80:502–9, 2007.

[83] J. D. Storey, J. E. Taylor, and D. Siegmund. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J Roy Stat Soc B*, 66:187–205, 2004.

[84] J. D. Storey and R. Tibshirani. Statistical significance for genome-wide studies. *Proc Natl Acad Sci USA*, 100(16):9440–9445, 2003.

[85] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis. Significance analysis of time course microarray experiments. *Proc Natl Acad Sci USA*, 102(36):12837–12842, 2005.

[86] B. E. Stranger, M. S. Forrest, A. G. Clark, M. J. Minichiello, S. Deutsch, and *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genetics*, 1:e78, 2005.

[87] The C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: A platform for investigating biology. *Science*, 291:1304–51, 2001.

[88] The Celera Genomics Sequencing Team. The sequence of the human genome. *Science*, 291:1304–51, 2001.

[89] The Inflammation and the Host Response to Injury Glue Grant, 2007.

[90] The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature*, 409:934–41, 2001.

[91] R. Tibshirani. Immune signatures in follicular lymphoma. *New Engl J Med*, 352:1496–7, 2005.

[92] G. C. Tseng, M. K. Oh, L. Rohlin, J. C. Liao, and W. H. Wong. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*, 29:2540–57, 2001.

[93] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci, U.S.A.*, 98:5116–21, 2001.

[94] M. J. van der Laan, S. Dudoit, and K. S. Pollard. Multiple testing. part II. step-down procedures for control of the family-wise error rate. *Stat Appl Genet Mo B*, 3:14, 2004.

[95] J. A. Warrington, N. A. Shah, X. Chen, M. Janis, C. Liu, and *et al.* New developments in high-throughput resequencing and variation detection using high density microarrays. *Hum Mutat*, 19(4):402–9, 2002.

[96] Y. H. Yang, M. J. Buckley, and T. P. Speed. Analysis of cDNA microarray images. *Brief Bioinformatics*, 2(4):341–9, 2001.

[97] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, and *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15, 2002.

[98] G. Yvert, R. B. Brem, J. Whittle, J. M. Akey, E. Foss, and *et al.* Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. *Nat Genet*, 35(1):57–64, 2003.

VITA

Jeffrey Leek was born in Lawrence, Kansas and grew up primarily in Pocatello, Idaho. He earned a Bachelor of Science degree in Mathematics from Utah State University in 2003 and a Master of Science in Biostatistics from the University of Washington in 2005. In 2007 he earned a Doctor of Philosophy at the University of Washington in Biostatistics.