This is to certify that the
dissertation entitled

DEPENDENCY AND PURITY IN LARGE-SCALE STATISTICAL SIGNIFICANCE TESTING: A DNA MICROARRAY PERSPECTIVE

presented by

Keyur Hemantkumar Desai

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Electrical Engineering

Major Professor's Signature

09/24/2008

Date

DEPENDENCY AND PURITY IN LARGE-SCALE STATISTICAL SIGNIFICANCE TESTING: A DNA MICROARRAY PERSPECTIVE

By

Keyur Hemantkumar Desai

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Electrical Engineering

2008

ABSTRACT

DEPENDENCY AND PURITY IN
LARGE-SCALE STATISTICAL SIGNIFICANCE TESTING:
A DNA MICROARRAY PERSPECTIVE

By

Keyur Hemantkumar Desai

Statistical methods for detecting differential gene expression (DGE) in DNA microarray experiments are discussed. A comprehensive definition of DGE refers to "statistical dependence" between gene expression levels and biological conditions of interest, such as differing environments, treatments, time points, phenotypes, or clinical outcomes. The ability to detect genuine DGEs has become crucial in the effort to understand and cure difficult diseases like cancer, diabetes, and Alzheimer's disease. The statistical nature of microarray data necessitates the use of "large-scale significance testing" to detect DGE. Large-scale testing is qualitatively different than the usual one-at-a-time testing: implied information from the "other" cases can force its way into the decision rule. Two prominent attributes of gene expression data, "inter-gene dependency" and "purity," complicate the situation even further. The former refers to the statistical patterns of dependencies among gene expression levels, and the latter stems from the fact that a great majority of genes show no change in expression even under changing conditions.

Inter-gene dependency is perceived commonly as a harmful force cluttering the decision-making. Contrastingly, this research takes a more sympathetic view of inter-gene dependency when seen together with purity. We argue that their combination gives rise to not only accurate but also more powerful statistical reasoning. Our effort to combine the two has culminated in two contributions, each of which has both

applied and theoretical implications. The first is a method for a better assessment of the number of false positives in the presence of heavy inter-gene dependency and extreme sampling errors due to exceedingly small sample sizes. The second is a method for "exploiting" inter-gene dependency to yield substantially more powerful DGE detection procedures. The empirical evidence from real and simulated test cases suggests the usefulness of our ideas.

ACKNOWLEDGEMENT

This research was supported in part by a Michigan State University IRGP grant, a graduate fellowship from the MSU Quantitative Biology Program, and a research fellowship from the MSU Department of Electrical and Computer Engineering. The support of the MSU High Performance Computing Center is appreciated.

I am grateful to Professors Christina Chan, Hayder Radha, Justin McCormick, and Jack Deller for being on my Ph.D. committee. Their feedback and advice have been a valuable contribution to this work.

Thanks to Professor Jack Deller for being a nearly perfect advisor. I owe my successful transition from engineering to genomics to his guidance, encouragement, support, and constant assurance. Thanks to Professor Justin McCormick for his generosity in time to explain intricacies of cancer research to a naive engineering student. Thanks to Professor Hayder Radha for his continual interest in my progress.

I am grateful to my friends Shirish, Sharath, Amit, Akshay, Snehal, Darshak, Amit Gore, Koushik, Kiran, and Abhishek for their friendship, support, and motivation. I am indebted to Bina Desai for her interest in my early education. Many thanks to Professor S.D. Mahanti for his support and encouragement during my early days in the graduate school.

This dissertation is dedicated to my mother, father, brother, and grandparents. Their love, support, and sacrifices have carried me to where I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction and Background

The objective of this research was to improve the existing statistical techniques of differential gene analysis. The literature identifies "dependency among gene expressions" as the chief obstacle in drawing meaningful conclusions from micorarray data and "purity" as its vast untapped resource. This research focuses on understanding, correcting, and exploiting the effects of dependency by combining it with purity in a general setting. It has culminated in two major findings pertinent to the general approach for detecting "statistical linear dependence" between gene expression levels and measured biological states. The general approach consists of two main steps: (i) ranking the genes and (ii) evaluating the number of false positives for a significance cut-off. Our first finding establishes the importance of third moment skewness corrections in estimating the number of false positives. The second finding provides a general technique to revise a given gene-ranking for better statistical power. A good starting point is to discuss the role of statistical reasoning in interpreting the gene expression data.

Organization of the chapter. Section 1.1 reviews the foundations of modern biology. Section 1.2 begins with a "bird's eye view" of the problem of differential gene detection and the potential implications of the present work. An example of two-state microarray data and a listing on typical sources of within state gene expression variations are given in Section 1.3. A contrast between one-at-a-time and large-scale significance testing is drawn in Section 1.4. In Section 1.5, we discuss the possible scope and the key goals of large-scale inferences and decision-making germane to the expression profiling data. An outline of the entire dissertation appears in Section 1.6. A short overview of DNA microarray technology concludes this chapter.

1.1 Foundations of Modern Biology

Before we begin, it is helpful to recapitulate the foundations of modern biology. Refer to Watson et al. (2004) for a more complete discussion. Hunter (1993) provides a brief survey of biology for researchers trained in engineering or computer science.

- Life changes and develops through evolution and that all life forms known have a common origin.

- The cell is the fundamental unit of life. Cells arise from other cells through cell division, and in multicellular organisms, every cell in the organism's body is produced from a single cell in a fertilized egg and hence contains the same genotype.

- Biological form and function are passed on to the next generation by genes, which are the primary units of inheritance and made of DNA. All information flows from the genotype, the genetic makeup of the organism, to the phenotype,

the observable physical or biochemical characteristics of the organism.

- The total complement of genes in an organism or cell is known as its genome. When a gene is active, the DNA code is transcribed into an RNA copy of the gene's information.

- Gene expression refers to the transcription of DNA into messenger RNA (mRNA) by RNA polymerase. The mRNA is then translated into protein by the ribosome. In gene expression analysis, expression level refers to the amount of mRNA detected in a sample.

- Phenotypic differences can ultimately be understood in terms of the regulation of gene expression.

1.2 Statistical Reasoning in Differential Gene Expression Detection

Gene expression profiling experiments measuring the expression levels of thousands of genes at once to create a global picture of cellular function have become a vital component of modern biomedical research. They let us investigate complex human diseases—such as cancer, diabetes, Alzheimer's disease, etc.—in unprecedented genetic detail. However, in order to draw meaningful conclusions from the wealth of gene expression data, we must rely on statistical reasoning for two main reasons. First, the measurements are based on mRNA molecules drawn from a cell population. Second, there are several uncontrollable sources of noise and uncertainty. Consequently, the promise of profiling—better understanding of regulatory mechanisms, more effective therapeutic targets, accurate (sub) categorization of diseases, superior

clinical diagnosis-prognosis, etc.—relies heavily on pertinent statistical methodologies (Lander, 1999; Petricoin et al., 2002; Singh et al., 2002; Jones and Baylin, 2002). Surprisingly, despite a major research interest in the statistical methods for gene expression data analysis, several key issues have remained unresolved due to their rather difficult and unfamiliar nature (Frantz, 2005).

Statistical methods for detecting differential gene expression (DGE) between two or more biological samples have attracted considerable attention. The comprehensive definition of DGE refers to "statistical dependence" between gene expression levels and biological conditions of interest, such as differing environments, treatments, time points, phenotypes, or clinical outcomes. In practice, however, "linear dependence," also known as "second moment dependence," is pursued more frequently because its detection is readily subject to a statistically rigorous approach, and yet its existence often suggests a phenomenon of scientific interest and clinical utility.

For convenience, the biological conditions are referred as "states." The term "two-state" refers to binary valued states, such as healthy versus cancer; "multi-state" to multi valued states, such as different categories of breast cancer; and "continuous-state" to continuous valued states, such as the age of the subject or insulin dosage, etc.

The need for detecting DGE between two states is very common. The methods customized for two-state is the present focus. Note that even after the common goal being to detect linear dependence, as per the situation (two-state, multi-state, or continuous-state), the methods may differ in some mathematical and implementation aspects. In fact, detecting second moment dependence in a two-state study can also be interpreted in terms of detecting statistical mean difference between the two states, and, in turn, dealt more efficiently through the standard (unpaired) t-statistic.

Indeed the t-test and *simple linear regression* are mathematically equivalent (refer to Section 2.1) implying that the present ideas for two-state methods are also applicable to multi-state or continuous-state methods, perhaps with little or no modifications.

Additionally, some of the outcomes of this research also seem translatable to DGE notions other than "central tendency," for example fold change (Biotechnology, 2006) or tail-rank statistics (Coombes et al., 2008; Zheng and Pepe, 2007). These extensions will be the subjects of future research.

The overarching theme of this dissertation is to develop methods for combining inter-gene dependency (Section 2.4) and purity (Section 2.5) to improve statistical decision-making. The former refers to (statistical) dependencies among gene expression levels and the latter to an abstract notion about the test statistics of differentially expressed (DE) genes not falling in certain intervals. Previous understanding perceived inter-gene dependency as more of a harmful force cluttering the decision-making, but now newer understanding suggests that when combined with purity, they can yield not only accurate but also more powerful statistical inference (Benjamini, 2008; Morris, 2008; Cai, 2008; Efron, 2008; Klebanov and Yakovlev, 2007). The present scope covers methods which include only the measured expression and "assigned" state labels; however, the possibility of extending these ideas to integrate legitimate "prior" knowledge like "gene grouping," as in enrichment analysis (Subramanian et al., 2005), has also shown potential.

Note. To avoid spurious mix-ups, the term "dependence" is reserved for statistical dependence *between* gene expression levels and biological conditions, whereas the term "dependency" is reserved for statistical dependence *among* gene expression levels. Throughout this dissertation, the words null and non-differential are used inter-

changeably; the same holds true for the words non-null and differential.

1.3 Variability in Gene Expression Data

Table 1.1 shows typical data generated by a two-state study. A distinctive characteristic of these data sets is a huge number of cases, a "large m," and very few measurements per case, a "small n." This particular data set is from van't Wout et al. (2003) who measured the mRNA levels of 7680 genes on 8 separate microarrays, 1–4 assigned to healthy cells and 5–8 to HIV infected cells. The challenge for the statistical methods is to discover the true signal, for example the true mean difference, in the presence of substantial within state (per) gene expression variations.

There are at least four factors contributing to these within state gene expression variations. These are:

1. *Cross-hybridization noise* due to some mRNAs molecules cross-hybridizing to the probes that are supposed to detect another mRNA.

2. *Measurement noise* associated with the microarray technology and related biochemical processing.

3. *Biological variability*: The mRNA molecules are obtained from a cell population and the individual cells can have differing gene expression levels due to a variety of influences including (i) differences in the cells' micro-environments (e.g., nutrient and temperature gradients), (ii) the growth phase differences between cells in the culture, (iii) the phase variations, and (iv) the periods of rapid change in the gene expression and multiple additional stochastic effects that

	healthy	healthy	healthy	healthy	HIV	HIV	HIV	HIV
gene 1	9.609	7.323	5.328	13.63	8.757	21.72	0.4873	6.364
gene 2	2642	5034	537.2	766.8	1123	961.7	601.7	1016
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
gene 1000	42.03	13.94	60.91	70.26	84.03	25.91	103.9	99.96
gene 1001	135.1	115.2	111.8	134.1	151.7	145.1	135.9	166.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
gene 4000	1455	513.1	1159	438.1	1806	647.4	1921	759.3
gene 4001	555.2	909.3	216.8	902.8	775.9	1510	492.7	1981
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
gene 7679	15.31	41.25	16.82	14.23	45.49	55.23	10.16	32.86
gene 7680	47.73	109.7	31.99	151.7	63.14	99.14	33.04	44.3

Table 1.1: An example of two-state comparative experiments. The HIV study of van't Wout et al. (2003).

cannot be controlled (Hatfield et al., 2003; Baldi and Brunak, 2001; Watson et al., 2004; Alberts et al., 2002).

4. *Confounding variables*, such as age, sex, race, living style, genotype, etc. in studies with differing subjects. These can introduce large-scale variation (Leek and Storey, 2007).

1.4 Testing for Statistical Significance

Due to the unavoidable variability, "statistical" decision-making is necessary to detect DGE. The theory and methods of statistical hypothesis testing deal specifically with such situations (Lehmann and Romano, 2006), yet testing each gene separately can be very "agonizing." From the beginning itself, the validity of a reference null distribution (obtained either from theory or re-sampling, see Section 2.1) may pose a concern. Next, a smaller p-value (defined in Section 2.1) does contradict the null hypothesis of "no DGE," however its role as a comprehensive measure of the inherent ambiguity is often under question (Ioannidis, 2005). Both the celebrated Neyman-Pearson lemma and the Bayesian decision theory are of limited use as they require specified knowledge of distributions under the alternative hypothesis. Moreover, the latter also requires *prior* distributions of hypotheses occurrence. Maybe the problem in itself is ill-posed. Berger (2003) and accompanying "follow-up comments" highlight different point of views in this matter. Fortunately, these conceptual difficulties go away when considering several thousand genes in parallel. In this case, at least in principle, there is enough data to infer the null, the possible alternatives, and their relative proportions.

The statistical framework known as large-scale significance testing (LSST) is a natural extension of one-at-a-time significance testing. The theory and methods of LSST deal with thousands of simultaneous tests (Lehmann and Romano, 2006). In other words, this framework can deal with all rows of Table 1.1 simultaneously. However, earlier approaches in LSST were limited in their treatment of dependency among tests. In practice, most microarray data exhibit substantial inter-gene dependency among expression variations (Owen, 2005), which eventually gets manifested in dependency among tests. There are number of numerical studies which point out that the

LSST approaches not entertaining inter-gene dependency properly can lead to highly inaccurate reporting of the underlying facts (Qiu et al., 2005b; Kim and van de Wiel, 2008).

After deciding that a proper treatment of dependency is a worthwhile research endeavor, the other alternative which we explored is cluster analysis, e.g., Eisen et al. (1998). When tailored to report two clusters, then clustering in essence can deal with dependency through an appropriate similarity metric. The drawback is that clustering neglects valuable information like assigned state values. Also, the overall framework itself is not very conducive to deducing precise statistical guarantees. Therefore, this work focusses on extending existing LSST methods to include dependency.

Specifically, we focus on extending the relevant LSST methods to include commonly observed patterns of "inter-gene correlation" in expression variations. Recall that correlation is a "scale-free" measure of statistical linear dependence also known as second moment or second-order dependence. The present emphasis is on *correcting* and *exploiting* the effects of inter-gene correlation by combining it with purity—a statistical assumption stemming from the fact that in many comparative studies the activity of a great majority of genes remains unchanged with respect to the treatment of interest. This can also be looked as drawing second-order *conditional inferences* based on cases that are fundamentally less ambiguous than the others. The scope and the goals of large-scale inferences germane to expression profiling data are discussed next.

1.5 Large-scale Inferences

1.5.1 The Scope

- The tests are related in the sense that the noise and uncertainty associated with them are of similar statistical nature perhaps due to common origins.

- We assume: For convenience the statistical reasoning may be done around per-gene univariate summary statistics, $T_1, \ldots, T_m$, and related null hypotheses, $H_1, \ldots, H_m$; however, that the original measurements are available and hence inferring the statistical structure among T_i's is possible.

- By convention, the rejection (or significance) regions to be considered are "tail-areas." Recall that the term rejection-region refers to an interval in the space of test statistics wherein we are interested in rejecting the null hypotheses. Both "one-sided" and "two-sided" tail-areas are allowed. Working with tail-areas naturally yields the methods which are sequential in nature and begin by ordering the test statistics; for example, see the Benjamini and Hochberg method described in Section 2.3.

  The rationale behind tail-areas is that the evidence for a gene being null decreases monotonically as we move farther from the intervals wherein most of the probability mass of the null distribution is concentrated. Methods employing rejection-regions different than tail-areas are observed to yield results with certain interpretational difficulties (Storey, 2002b; Rice and Spiegelhalter, 2008).

- Moreover, the number of genes, denoted by m, may run in several thousands—for current comparative microarray experiments m typically ranges from 5000 to

25000. Large-scale methods and algorithms should anticipate and accommodate computational constraints and statistical convergence issues that are typical to this range.

- The methods are customized in the sense that they try to capture the scientific context affecting the data—the present emphasis is on incorporating and exploiting typical dependency structures by combining them with purity.

1.5.2 The Goals

Broadly speaking, the goal is to *reliably* identify as many DE genes as possible. To elaborate on more specific statistical goals, the following terminology is helpful: When a null hypothesis H_i is rejected, in other words when a gene is declared DE, then a "discovery" is made; when a gene declared as DE is indeed non-DE, then a "false discovery" is made. A broader statistical goal is to report as many discoveries as possible while not exceeding a specified "false discovery proportion." Sometimes a "ball park" figure for the number of discoveries is specified and then, the goal is to minimize the false discovery proportion as much as possible. When recast in the language of statistical inference, the former amounts to an accurate estimation of the number of false discoveries for a rejection region and the latter to revising the test statistics to increase statistical power.

1.6 An Outline

This dissertation is outlined as follows. Chapter 1 introduces the "problem." Chapter 2 presents a mathematical formulation, reviews the relevant literature, discusses inter-gene dependency and purity, and underlines our main contributions. The issue

of estimating the number of false discoveries is discussed in Chapter 3, whereas Chapter 4 presents a novel gene re-ranking method. Conclusion and future work appear in Chapter 5.

At the beginning of each chapter a separate paragraph called "organization of the chapter" is provided to guide the reader through that particular chapter.

Supplement: DNA Microarrays

A DNA microarray uses the selective nature of DNA-DNA or DNA-RNA hybridization. In this process two complementary strands of DNA (sometimes DNA and a complementary strand of RNA) bind. An array is a collection of microscopic DNA segments attached to a solid surface, such as glass, plastic, or silicon chip. The affixed DNA segments are known as probes, thousands of which can be placed in known locations on a single array (see Fig. 1.1). In a typical experiment, the microarray is washed with a sample containing fluorescent dyed mRNA transcripts. The probe and the target sequences will hybridize according to their complementary nature. The abundance of target molecules can then be identified based on the resulting fluorescence patterns. Typical numbers reported by a microarray appears in Table 1.1. Based on the way DNA microarrays are fabricated, two distinct microarray platforms have evolved, each with its own pros and cons.

Spotted microarrays or two-color microarrays. These arrays employ *clones* or *oligonucleotides* that are spotted onto glass slides and two distinctly labeled complementary DNA (cDNA) samples are hybridized together on a single array. The advantage is that the spotting process is relatively inexpensive and it is easier to make custom made arrays (Schena et al., 1995). This flexibility however comes at a

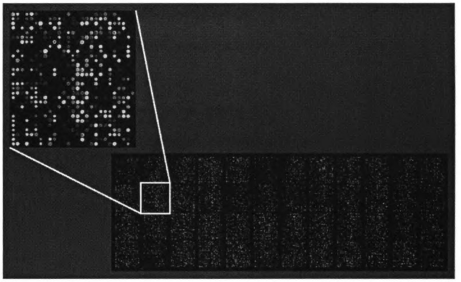


Figure 1.1: An example of an approximately 40,000 probe spotted oligonucleotide microarray with enlarged inset to show detail.

price: The spotting process is inherently variable. This limitation is circumvented by reporting relative mRNA expression levels instead of absolute expression levels(Woo et al., 2004). The relative levels are obtained by co-hybridizing a two-color array with two RNA samples, namely, *experimental* and *reference*. Here, the reference should not be confused with the control (for example, while comparing normal cells and their cancer mutants, the RNA sample from the normal cells is termed the control and from the cancer cells the treatment).

Affymetrix microarrays or one-color microarrays. These arrays employ short (25-mer) oligonucleotide probes deposited on a silicon substrate through high precision photolithography (Lockhart et al., 1996). This process is similar to the one used in manufacturing of electronic microchips and remarkably accurate in producing nearly identical arrays. However, the flexibility of customization is lost since for every custom array an expensive mask must be created. One-color microarrays report absolute

mRNA expression levels.

Normalization. Measurements reported by two separate microarrays, even if they belong to a single study, are seldom comparable in their original form. The major reasons are unequal quantities of starting RNA, differences in labeling or detection efficiencies between the fluorescent dyes used, and systematic biases in the measured expression levels (Quackenbush, 2002). A transformation, known as normalization, is necessary. Robust normalization methods for both one-channel and two-channel microarray data are available.

Cross-hybridization. Even after perfect normalization, the measured mRNA levels can still suffer from experimental noise due to *cross-hybridization*: some mRNAs may cross-hybridize probes in the array that are supposed to detect another mRNA. This problem has been alleviated somewhat by a systematic selection of expressed sequence tags (ETS) with high specificity-selectivity (Li and Stormo, 2001). Both microarray platforms have evolved to provide very similar data quality (Patterson et al., 2006).

Chapter 2

Large-Scale Significance Testing

The task of differential gene expression (DGE) detection is often recast as large-scale significance testing (LSST). Inter-gene dependency among gene expression variations weakens the LSST DGE detection formulation by introducing substantial dependency among test statistics. Any other formulation is equally prone to the effects of dependency. However, within the framework of LSST itself, dependency can be understood, corrected, and exploited by combining it with purity. This chapter lays down the necessary statistical groundwork to do so. A direct treatment of generic any-order dependency is impractical, but treating second-order dependency is both useful and mathematically tractable. A scale-free version of second-order dependency is correlation and how to include inter-gene correlation in the LSST formulation is discussed.

Organization of the chapter. Section 2.1 presents a statistical formulation of the "problem." A mapping from inter-gene correlation to inter-test correlation is given by Eqn. 2.3. Error measures, the false discovery proportion and the false discovery rate, appear in Section 2.2. Section 2.3 reviews the relevant literature and states the principal contributions of this work. A discussion on inter-gene dependency is

provided in Section 2.4. Purity and two related concepts, "identifiability" and "zero assumption" are discussed in Section 2.5.

2.1 A Statistical Formulation for DGE Detection

There are $m + 1$ random variables involved in this statistical reasoning: L and $(X_1, \ldots, X_m)$. Here, L refers to "state" defined in Section 1.2, and X_i to expression level of gene i. Their (unknown) underlying joint probability distribution is $\mathcal{F}(L, X_1, \ldots, X_m)$. A microarray experiment obtains n independent and identically distributed samples from this distribution.

A "random" data set is written as $(\mathbf{L}; \mathbf{X}_1; \ldots; \mathbf{X}_m)$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{in})$ denotes a random sample for gene i and $\mathbf{L} = (L_1, \ldots, L_n)$ the corresponding states. The subscript i in X_{ij} indexes the genes and j the microarrays. To increase normality, raw X_{ij}'s are converted to a logarithmic scale. The "realized" data set is written as $(l; \mathbf{x}_1; \ldots; \mathbf{x}_m)$ with $l = (l_1, \ldots, l_n)$ denoting the "assigned" states and $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})$ the measured expression for gene i.

The overarching goal is to test whether random variables L and X_i, $i = 1, \ldots, m$, are dependent. The existence of statistical dependence between L and X_i suggests that gene i may be crucial to the phenotypic distinctions being studied. A rigorous treatment of general statistical dependence measures like mutual information (Cover and Thomas, 1991) is often impractical. Instead most approaches test linear (or second moment) statistical dependence. A univariate statistic measuring linear dependence between L and X_i can be obtained through *simple linear regression* between $\mathbf{X}_i$ and $\mathbf{L}$.

The LSST for DGE detection involves m univariate per gene summary statis-

tics, $T_1, \ldots, T_m$, with T_i corresponding to gene i. These T_i's are assessed using $H_1, \ldots, H_m$, the null hypotheses of independence between L and X_i, $i = 1, \ldots, m$. A T_i is gauged with respect to its null distribution $p(T_i|H_i)$: The fact, it is "typical" under H_i, suggests that L and X_i are uncorrelated. Note, however, that other forms of statistical dependence between L and X_i may go undetected. Philosophically the bottleneck is caused by the fact that statistical independence has a concrete mathematical interpretation, but statistical dependence is just an arbitrary functional relation between random variables.

The two-state situation can be handled efficiently through the standard unpaired t-statistic:

$$T_i = (\bar{X}_{i;2} - \bar{X}_{i;1})/S_i, \tag{2.1}$$

where $\bar{X}_{i;k}$ is the mean of gene i in state k and S_i is the pooled within-state standard deviation of gene i. Mathematically the t-statistic is equivalent to performing *simple linear regression*, and hence, in essence it tests linear dependence. The usual interpretation, i.e., the t-statistic detecting a mean difference and H_i referring to the true mean difference being zero, is more obvious and well-known. If the additional assumption of normality of X_i is made, then $p(T_i|H_i)$ can be replaced by the Student's t-distribution. Doing so is not absolutely necessary: Permutation calculations can estimate a putative null cumulative density function (cdf) common to all genes, say G_0, which may represent the basic facts more accurately (Efron, 2007a).

Dependency among "truly null" genes has profound implications on the "significance cut-off," i.e., the threshold beyond which if a T_i occurs, then its corresponding gene is declared non-null. Dependency among null and / or non-null genes has profound implications on the possibility of mapping $(T_1, \ldots, T_m)$ to $(T_1^*, \ldots, T_m^*)$ for better statistical power. Again, a general treatment of dependency among T_i's is

impractical, but correlation among T_i's is tractable, both mathematically and computationally, and will be pursued.

When T_i's are obtained through *simple linear regression*, then correlation among T_i's is easy to track. Lemma 1 of Owen (2005) states that for T_i's measuring the "linear correlation" between L and X_i, if expression of gene i and gene i' is independent of state L, then

$$\text{Cor}(T_i, T_{i'} \mid \mathbf{x}_i, \mathbf{x}_{i'}) = \hat{\rho}_{ii'}. \tag{2.2}$$

In Eqn. (2.3) $\hat{\rho}_{ii'}$ is the "Pearson product-moment correlation coefficient" between samples $\mathbf{x}_i$ and $\mathbf{x}_{i'}$. It can be verified through computer simulations that the same fact applies to t-statistics obtained through *simple linear regression.*

The discussion surrounding Eqn. (2.3) suggests following: When expression of gene i and gene i' is independent of state L, and statistics T_i and $T_{i'}$ are due to *simple linear regression*, then

$$\text{Cor}(T_i, T_{i'}) = \rho_{ii'}, \tag{2.3}$$

where $\rho_{ii'}$ is the "theoretical" correlation between gene i and gene i'. For exceedingly small n, the sampling error in $\hat{\rho}_{ii'}$ is a concern. This concern, when making inferences for millions of inter-gene correlation coefficients simultaneously, can be circumvented through clever data processing (see Section 3.2.2). Owen (2005, Section 4) and Efron (2007a, Section 2) both adopt a similar point of view.

For analytical convenience, T_i's can be converted to z-values:

$$Z_i = \Phi^{-1}\left\{G_0(T_i)\right\}, \quad i = 1, \ldots, m, \tag{2.4}$$

where G_0 is the putative null cdf mentioned above and Φ^{-1} is the inverse cdf of $\mathcal{N}(0,1)$. For "truly null" cases we expect $Z_i \sim \mathcal{N}(0,1)$. Note that Eqn. (2.4) is a nonlinear order-preserving monotone transformation. The relationship between $\text{Cor}(T_i, T_{i'})$ and $\text{Cor}(Z_i, Z_{i'})$ must be calibrated. In practice, the approximation $\text{Cor}(Z_i, Z_{i'}) \approx \text{Cor}(T_i, T_{i'})$ is seen to work well (Efron, 2007a). Methods for direct estimation of $\text{Cor}(Z_i, Z_{i'})$ are discussed by Zou and Hall (2002).

Some DGE methods convert test statistics $T_1, \ldots, T_m$ to p-values, $P_1, \ldots, P_m$, through $p(T_1|H_1), \ldots, p(T_m|H_m)$. For example, a one-sided p-value corresponding to $T_i = t_i$ is obtained as $p_i = \min[\Pr(T_i < t_i \mid H_i), \Pr(T_i > t_i \mid H_i)]$; a two-sided p-value is obtained as $p_i = \Pr(|T_i| < |t_i| \mid H_i)$.

2.2 Role of the Number of False Discoveries in Measuring Quality

Recall the terminology introduced in Section 1.5. When a null hypothesis H_i is rejected, in other words when a gene is "declared" non-null, then a "discovery" is made; when a gene declared as non-null is indeed null, then a "false discovery" is made. Let for a random data set $(\mathbf{L}; \mathbf{X}_1; \ldots; \mathbf{X}_m)$ a decision rule report R number of discoveries. Let V $(\leq R)$ of these are false discoveries. The ratio V/R, known as the False Discovery Proportion (FDP) and interpreted as zero when $R = 0$, is of obvious appeal, especially in exploratory data analysis where statistical findings form a basis for further investigation.

Contrastingly, pre-FDP LSST methods focused on V-only-error-rates because the emphasis was more on confirmatory data analysis and stricter safe-guards against making erroneous positive statements were necessary. Notice, when both V and

R are involved, then the decision rule is highly data dependent: If data with less ambiguity is encountered then "relatively" more findings are reported and vice versa.

FDP and FDR. The usual convention assumes the random variable R as "observable" and V as "unobservable" and hence requiring estimation. In their breakthrough paper, Benjamini and Hochberg (1995) used an all-null-theoretical-expectation,

$$E_\Gamma(\dot{V}) = \sum_{i=1}^{m} \Pr\left(T_i \in \Gamma \mid H_i\right), \tag{2.5}$$

as an estimate of the "realized" v for an arbitrary rejection-region Γ. The "dot" in $E_\Gamma(\dot{V})$ emphasizes the fact that it is an "all null" expectation, which may be contrasted with $E_\Gamma(V)$, the true (unknown) expectation for that Γ. These two expectations start deviating appreciably as the proportion of null genes, π_0, reduces. If the putative null cdf G_0 is obtained through permutations, then Eqn.(2.5) is a part of the Yekutieli-Benjamini FDP Estimator described in (Qiu and Yakovlev, 2006, Section 3.2).

Benjamini and Hochberg showed that their procedure, which estimates the realized v/r through $E_\Gamma(\dot{V})/r$ with r referring to the realized value of R, can "bound" the average FDP, i.e., $E(\mathrm{V/R})$, at an arbitrary prescribed level. The average FDP is now widely known as the False Discovery Rate (FDR). The bounding of FDR is termed as "control of FDR" and discussed in the next section.

2.3 Literature Review and Contributions of This Research

2.3.1 Control.

"Control of an error measure" is a concept that developed in the Frequentist-framework of LSST. Here, the error measures are compound involving a hypothetical data ensemble which can be generated by the experiment. The goal is to develop methods that can keep the error measure *below* a prescribed "level." Among all the methods that can do so, the ones with more average power (the proportion of true discoveries which are reported) are sought after.

Controlling FDR: Independent tests. The sequential p-value algorithm of Benjamini and Hochberg (1995), which controls the FDR at level α, has following steps:

1. Let $H_1 \dots H_m$ be the null hypotheses and $P_1 \dots P_m$ their corresponding p-values
2. Order p-values in increasing order and denote them by $P_{(1)} \dots P_{(m)}$
3. Find $R = \underset{0 \le k \le m}{\text{argmax}} \left(P_{(k)} m\right) \le (k\alpha)$, where k is an integer
4. Reject all $H_{(i)}$ for $i = 1, \dots, R$.

This algorithm assures that the reported discoveries have an "estimated FDP" always below α. Benjamini and Hochberg proved the following: When truly null cases are statistically independent, then in the long run behavior the above procedure guarantees $E(V/R) \le \alpha$. Storey et al. (2004) provides an alternative and a more concise proof. That the equality holds, requires additional assumptions.

According to Benjamini (2008) the motivation behind their 1995 method was to deal efficiently with around 100 statistically independent hypotheses in a clinical setting. Their method is now applied to situations with thousands and sometimes even millions of cases and yet whenever independency holds there is no reason to disbelieve its efficacy. By the time it was realized that the method might be critically susceptible to inter-case dependency was realized, it had become a standard technique.

In DGE detection, dependency is intrinsic because the gene expression data represent the life process, a fundamentally coordinated activity. Yet the exact nature, form, and amount of dependency are still being debated—Section 2.4 discusses this in detail.

This issue of dependency has profoundly affected the LSST research. Despite strong interest, the issue remains unresolved. Not only very little is understood about how to exploit dependency, but even to determine the adverse effects of dependency on methods developed with the assumption of independency has been found very challenging. In relation to dependency, a major emphasis in the LSST research has been to establish "validity" of any proposed procedure. Validity refers to the fact that for a given procedure, the FDR can be bounded below an arbitrarily prescribed level.

Validity: Dependent tests. Benjamini and Yekutieli (2001) show that for tests with a positive-regression dependency condition, estimating the FDP as $E_\Gamma(\dot{V})/r$ still ensures validity, and if this FDP estimator is modified as $E_\Gamma(\dot{V}) \cdot \left(\sum_{i=1}^{m} i^{-1}\right)/r$, then validity is ensured for arbitrary dependency. However, this implies a very conservative FDP estimation; for example, with $m = 25000$ a scaling of $1/10.7$ is performed. A slightly different but more provocative interpretation is following: in order to ensure validity in the presence of dependency, the realized data are judged at a much more

stringent FDP level than what is needed. Benjamini (2008) argues that the real situations fall between positive regression dependency and arbitrary dependency and hence a less conservative scaling should work.

Instead van der Laan et al. (2004) propose to satisfy the probabilistic constraint $\Pr(\text{FDP} > \gamma) < \alpha$ by modifying the existing procedures that satisfy the probabilistic constraint $\Pr(V > \gamma) < \alpha$; this conversion does not require independent tests, but what happens to the average power is not clear. Lehmann and Romano (2005) propose two more methods to satisfy the probabilistic constraint $\Pr(\text{FDP} > \gamma) < \alpha$ under dependency. The first method can work under an abstract mathematical condition on dependency structure claimed to be reasonable. The second method can work for arbitrary dependency but the same problem as in Benjamini and Yekutieli (2001) with arbitrary dependency persists: The realized data is judged at a more stringent FDP level than necessary.

Since the existing FDR procedures that are capable of ensuring validity under an arbitrary dependency are highly conservative, most investigators still rely on the original Benjamini and Hochberg procedure.

Validity and variance. After proving validity of a procedure, for arbitrary dependency or a special dependency structure of interest, further evaluations are still needed. An important measure is the variance of the FDP estimation, $E(\text{FDP} - \alpha)^2$; here, the actual FDP is contrasted with the intended level α. Owen (2005) argues that this could be a serious issue because inter-gene dependency found in most microarray data tends to inflate the variance of V considerably compared to what it may be under independency; in the cases considered by Owen, the inflation is nearly 100 times.

Therefore, approximating the realized v by an all-null-theoretical-expectation for heavy dependency may be critically unrealistic. A simulation in (Efron, 2007a, Section 5) emphasizes this issue. Qui et al. (Qiu and Yakovlev, 2006; Qiu et al., 2005a, 2006, 2005b) also explore this effect and warn against neglecting the effect of dependency on an overall, as well as case-to-case, basis. An attempt to develop a better estimate of v may be worthwhile; however, the ways to do so are not obvious.

Conditional v estimates. Efron suggests "conditional v inferences" to account for inter-gene dependency. The notion of purity as in Section 2.5 is crucial for conditioning to work. Efron (2004) proposes a technique called "empirical null" to perform conditioning; this technique works solely on the realized t_i histogram and the dependency observed in the actual data is not entertained.

In a follow-up paper (Efron, 2007a), Efron develops a second moment theory for "null T_i histogram" to gain insight into the behavior of V. He makes a crucial observation that under heavy inter-gene correlation the behavior of the null T_i histogram is highly regulated in the sense that between its tail and its center there is extreme negative correlation.

Contribution 1. Based on Efron's observation, this work builds a moment-based estimator of the number of false discoveries. Sections 3.1—3.3 develop the methodology behind the proposed estimator and Section 3.4 uses real and simulated examples for verification. The key issue, the proposed estimator addresses, is the exceedingly small n situations as encountered in the HIV example of Section 1.3 where extracting any useful information about the dependency structure in itself is a major challenge. A notable feature of the proposed approach is that it naturally leads to an estimate of the entire probability distribution of the random variable V.

2.3.2 Power

For the present LSST formulation of DGE, the test statistics $T_1, \ldots, T_m$ are "designed" to capture linear dependence between random variables L and $X_1, \ldots, X_m$. The observation that there might be ways of mapping the original set of statistics $(T_1, \ldots, T_m)$ to a newer set $(T_1^*, \ldots, T_m^*)$ for better statistical power is very appealing. There are two notable and successful attempts in this regard, both, at some level, relying on the assumption of independent or weakly dependent T_i's.

Estimating proportion of null cases. Such attempts can be seen as identically scaling the original $(T_1, \ldots, T_m)$ in terms of their ability to detect linear dependence. The philosophy behind this adjustment is: if the proportion of null genes (π_0) is appreciably small, then the case for linear dependence is strengthen accordingly. Storey (2002a) and Pawitan et al. (2005) present and review successful attempts. Most π_0 estimators require the assumption of independency or weak dependency among T_i's.

Borrowing strength. The understanding that a better estimate of variance in the standard t-statistic [Eqn. 2.1] should yield more power, led to the idea of a "stabilized variance estimation," also known as "smooth t-test." The idea has been reintroduced in several forms. Baldi and Brunak (2001) use a Bayesian approach to trade-off between the sample variance for the gene of interest and a combined sample variance from similar looking samples; it is called a "shrinkage" estimate. Newton et al. (2001) use a Gamma-Poisson model which achieves a similar effect. The "fudge factor" of the SAM algorithm, Tusher et al. (2001), is also of the same variety. All these approaches make an implicit assumption that amount of inter-gene correlation is small enough so a combined sample variance from multiple genes will not suffer from excessive random

fluctuations.

Borrowing strength through signal-induced correlation. Tibshirani and Wasserman (2006) present a scheme to combine T_i's using the inter-T_i correlation for obtaining T_i^*'s with better statistical power. Their scheme averages a T_i with other T_i's in its "correlation neighborhood." They assume that the correlation among T_i's is primarily due to the treatment effect. Therefore, "truly" non-null T_i's would add up together and map to stronger T_i^*'s by reducing the noise through averaging, and truly null T_i's would largely be unaffect by the averaging. Philosophically their idea is to add-up signal strength among non-null cases through the device of sample correlation.

Storey et al. (2007) present an interesting, more general approach, of accomplishing a similar effect.

Exploiting residual correlation. Microarray samples that are assigned the same "state" do exhibit a substantial inter-gene correlation which suggests that the source of inter-gene correlation is more intrinsic since hypothetically there are no treatment differences for identical states. Therefore, we conclude that the source / sources of correlation affect both null and non-null genes. However, somehow, the possibility that this "intrinsic" correlation among T_i's can be "exploited" to yield a superior $(T_1^*, \ldots, T_m^*)$ is unexplored, at least in the LSST formulation of DGE. This work attempts to do so. Our contribution in this regard is described below.

Contribution 2. The basis of Contribution 1 is to perform conditioning on evidentially more likely null genes and while doing so incorporate the observed inter-gene correlation. Philosophically we employ the same idea but this time to yield a better $(T_1, \ldots, T_m) \rightarrow (T_1^*, \ldots, T_m^*)$ mapping. Sections 4.1—4.3 develop the methodology

behind the proposed re-ranking procedure and Section 4.4 uses real and simulated examples for verification. The approach relies on the residuals of *simple linear regression*. This contribution is complementary to the first as the proposed approach yields gains for moderate n situations, for example the Prostate cancer data of Singh et al. (2002) as in Section 4.4. If substantial null-null correlation is present and purity holds, then the gain in statistical power is notable.

2.4 Dependency in Gene Expression Data

Dependency refers to the fact that in a random data set the expression levels X_{ij} and $X_{i'j}$ are statistically dependent. Formally, $p\left(X_{ij}, X_{i'j}\right) \neq p\left(X_{ij}\right) p\left(X_{i'j}\right)$. It is helpful to think from the point view of mutual information between X_{ij} and $X_{i'j}$. Recall that the mutual information (Cover and Thomas, 1991) between random variables X and Y is defined as

$$I(X;Y) = \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p_1(x)\, p_2(y)}\right) dx\, dy.$$

That on-an-average how much dependency (or say mutual information) there is between two gene expressions measurements on a same micorarray, is the topic of a major scientific debate (Klebanov et al., 2006).

In practice, method-development for general dependency measures like mutual information is impractical; however, second-order dependency, i.e., correlation between X_{ij} and $X_{i'j}$, is mathematically tractable. A careful examination of various microarray data sets implies the existence of substantial on-an-average correlation between two arbitrary gene samples. Owen (2005) examines four different data sets — (i) expression levels in different human tissue (ii) a Kidney aging data (iii) human stress

data and (iv) gene expression in yeast — and concludes the existence of substantial inter-gene correlation.

Qiu et al. (2005b), Qiu et al. (2006), Almudevar et al. (2006), and Klebanov et al. (2006) argue that commonly observed inter-gene correlation is intrinsic and must be accounted for. However, we can just speculate about the possible sources of inter-gene correlation, such as co-regulations of the genes, spatially correlated measurement errors (Reiner-Benaim, 2007), confounding variables introducing long-rage dependency (Leek and Storey, 2007), etc. Klebanov and Yakovlev (2007) call for to improve statistical inference by incorporating correlation structures.

2.5 Purity, Identifiability and Zero Assumption

The assumption of "purity" states that we will not discover anything interesting near the center of the null distribution (Genovese and Wasserman, 2004). Efron (2008), while working with z-values, rephrases this as the "zero assumption" (ZA):

Zero assumption most of the z-values near 0 come from null genes.

Efron (2006) discusses the use of the ZA in a variety of differential analysis approaches. It plays a central role in the literature on estimating the proportion of null genes, as in Pawitan et al. (2005) and Langaas et al. (2005). The ZA is equally crucial for the two-group model approach developed in the Bayesian microarray literature, as in Lee et al. (2000), Newton et al. (2001), and Efron et al. (2001). The ZA is also important for the "empirical null" technique of Efron (2004).

For the purposes of statistical reasoning, purity allows to impose "identity" on null genes. This is known as "identifiability." In the present context, this notion refers

to the act of incorporating the genes with test statistics near the center of the null distribution as *explicitly* null in the inference. Here, the premise is that the statistical risk of imposing identifiability may weigh less than not doing so. Empirical evidence supports this assumption.

The assumption of purity is more believable if the proportion of null genes π_0 is huge. For a variety of gene expression studies this seems to be a common situation as the number of genes behaving differently in closely compared phenotypic distinctions is thought to be a small proportion. Intuitively, a significant number of total genes are involved in "basic pathways" that are crucial to the overall functioning and survival of the cell, and hence, by and large, their expression may remain unchanged with respect to the treatment of interest. A good part of statistical literature assumes $\pi_0 \geq 0.9$ for method-development purposes; see (Efron, 2004) and references therein.

Chapter 3

Estimating the Number of False Discoveries

The previous chapter emphasized that the number of false discoveries is central to large-scale significance testing. This chapter discusses the issue of estimating the number of false discoveries in the presence of substantial inter-gene correlation. In particular, we focus on exceedingly small sample sizes wherein estimating the inter-gene correlation structure becomes very challenging. Such situations are very common in cost-constrained microarray investigations which typically involve only 3-4 replicates per biological state. We lay the statistical groundwork for a method which, in principle, can estimate an entire distribution of a random variable model of the number of false discoveries. This distribution is interpretable from both Bayesian and Frequentist points of view. A distinctive feature of the present method is that it first summarizes the effect of millions of pair-wise correlation coefficients in a single parameter β, then explicitly incorporates this parameter in the inference. Doing so offers the possibility of a sophisticated yet practical inferential technique capable of

handling exceedingly small sample sizes.

Organization of the chapter. Key challenges in estimating the number of false discoveries and the intuition behind the proposed method appear in Section 3.1. Section 3.2 develops the method. Algorithm 1 in Section 3.3 summarizes the main steps. Section 3.4 applies the method to real and simulated test data. A discussion and potential extensions appear in Section 3.5.

3.1 Overview

Recall the basic "large-scale statistical significance testing" formulation from Section 2.1: Genes are represented by summary statistics and corresponding null hypotheses,

$$H_1, H_2, \ldots, H_m$$
$$T_1, T_2, \ldots, T_m$$
$$t_1, t_2, \ldots, t_m,$$

where T_i refers to a "random value" and t_i the "realized value." The magnitudes of the t_i's establish a gene-ranking, and the top $r \ll m$ genes with the largest t scores are reported as statistically significant discoveries. The cut-off between non-null and null genes is determined on the basis of the ratio v/r, where v refers to the number of false discoveries in r discoveries. This ratio, known as the false discovery proportion, is detailed in Section 2.2. In Section 2.3 it is pointed out that a key issue in large-scale significance testing is an accurate assessment / estimation of the unknown quantity v.

Similarly to the pair (t_i, T_i), it is also helpful to visualize the quantity v as the realized value of a random variable V associated with a hypothetical data ensemble. Formally,

$$V = \#\{\text{null } T_i \in \Gamma\},$$

where Γ is the rejection-region under consideration. Here, "#" is read as "the number of." The sample space and the probability distribution of V can be deduced from the joint distribution of null T_i's. Since the identity of truly null genes is unknown we restrict ourselves to a conservative all-null-calculation where all the genes are assumed to be null. Such "all null treatment" is very common in the microarray literature in which the goal is to identify a relatively small set of interesting non-null genes. For example, see Owen (2005), (Efron, 2007b), and Efron (2008); the discussion in Section 2.5 also has some reference to this point. The "FDR terminology" appears in Section 2.3.

In their breakthrough paper, Benjamini and Hochberg (1995) used an all-null-theoretical-expectation, $E_\Gamma(\dot{V}) = \sum_{i=1}^{m} \Pr(T_i \in \Gamma \mid H_i)$, Eqn. (2.5), as an estimate of the "realized" v. In fact, Benjamini and Hochberg (1995) showed that this intuitive estimator ensures that the FDR stays below an arbitrarily prescribed level. Note that when the proportion of null genes is close to one, then $E_\Gamma(\dot{V})$ is a good approximation of the true (unknown) expectation $E(V)$. It was later realized that when null cases are highly correlated, the variance of V,

$$\text{Var}(V) = E\{V - E(V)\}^2,$$

is greatly inflated. Owen (2005) presents a mathematical analysis of this fact and emphasizes that for certain situations the increase can be nearly 100 fold.

A direct implication of inflated $\text{Var}(V)$ is that the realizations of the random value V deviate substantially from the expected value $E(V)$. Which casts doubts on the effectiveness of approximating the realized v with $E_\Gamma(\dot{V})$. The recent literature calls for a better treatment of inter-gene correlation (see Section 2.4 for details). Therefore, "to build a realistic $\hat{v}$ for highly correlated tests" is the purpose of the research described in this chapter. The symbol $\hat{v}$ should be read as "an estimate of the realized v."

The present emphasis is on exceedingly small sample sizes, only 3-4 replicates per biological state, similarly to the HIV example of Section 1.3. The inherent difficulty with small sample situations is that extracting any useful information about inter-gene correlation structure in itself is a major challenge (Owen, 2005; Efron, 2007a).

Intuitively, the approach draws conditional inferences based on the identifiable information. Mathematically, if I is the information from identifiability, then we seek $\hat{v} = E(V|I)$. Section 2.5 discusses the concept of identifiability in detail. Formalizing this intuition is not straightforward.

3.2 The Proposed Method

The proposed method models the "histogram binning" of random values T_i's. While pursuing this modeling, inter-gene correlation enters into the statistical inference in a subtle way. For convenience, we work with the z-values by mapping T_i's to Z_i's (see Section 2.1 and Section 3.2.1 for details). Efron (2007a) has developed a second-moment theory for null Z_i "histogram binning" to gain insight into the behavior of V. At some level, in order to construct a moment-based $\hat{v}$, we effectively extend the "moment framework" by including third moment skewness corrections.

Due to purity the center of the Z_i histogram will be populated mostly by null cases. Therefore, we can place a small zero-symmetric bin designated as "center-area" in the sample space of Z_i's and posit that the count in center-area is primarily due to null cases. Consequently, the null count in center-area, denoted by C, is observable. Next, the observed value, say c, of the random variable C can be used to condition V to yield a theoretically better estimate of v.

The most complete probabilistic relationship between V and C is given by the joint distribution $P(V, C)$. We estimate this distribution. To do so we estimate the first three moments of the random variables (V, C) and later fit the maximum entropy distribution (*maxent*) to these moments. Section 3.2.1 discusses the moment estimation and Section 3.2.3 the *maxent* fitting. Naturally, this approach yields an estimate of $P(V|c)$. We claim that the proposed method yields an entire distribution of the number of false discoveries. The entire distribution is potentially more useful than a point estimate because of the noisy nature of large-scale inferences (Owen, 2005). If one wishes to skip conditioning altogether, then this method can report an estimate of $P(V)$ and is therefore still useful. Compared to purely histogram-curve-fitting techniques like "empirical null" (Efron, 2004), this approach enjoys the attractive feature that correlation is separately estimated and later explicitly incorporated in the inference through the moments of (V, C).

A crucial observation is that for most microarray data there is extreme negative correlation between random variables V and C (Efron, 2007a). Efron (2007a) provides an explanation for this phenomenon and then uses the gained insights to build a Poisson model based second-order $\hat{v}$ which is also reliant on the notion of a center-area. However, an hypothesis of this research was that characterizing the correlation between V and C using moments could be more helpful in correcting the effect of

inter-T_i dependency. Indeed computing the second moments is straightforward, but unfortunately purely second-order estimators apparently suffer from potentially hazardous "over / under estimation events." Both the present second moment $\hat{v}$ (see the results of Section 3.4) and Efron (2007a) $\hat{v}$ show this effect.

Three observations explain these estimation issues: (i) the random variable V is bounded below by zero, (ii) the expectation of V is small, and (iii) correlation causes the variance of V to inflate. These observations suggest that third moment skewness corrections are vital. Indeed Owen (2005) encourages further investigation of this point.

However, a third moment extension under severe sampling error is nontrivial. The chief contribution of these developments is an inferential technique to estimate the empirical density of 3×3 covariances enabling realistic estimates of third moments. In effect, it is possible, to within a useful degree, to fix the estimation issues inherent in second-order approaches.

The present extension of the moment framework, in principle, admits any-order moments, but computational challenges have prevented inclusion of higher than third moments. The inclusion of the third moments provides significant improvement for a range of real and simulated examples (see Section 3.4). Hence, a key observation relating to v estimation methodologies is that techniques which rely on identifiability can be enhanced by including third moment skewness corrections if they do not already do so.

Section 3.2.1 discusses the moment framework and derives the necessary formulae. Extraction and modeling of correlation information is discussed in Section 3.2.2. The maximum entropy technique of distribution fitting is described in Section 3.2.3.

3.2.1 Estimating the Moments

The goal of the work reported in this section is to derive mathematical formulae for estimating the moments of *random variables* V and C. This includes the individual moments (e.g., $E(V)$,$E\left(C^2\right)$, etc.) as well as the joint moments (e.g., $E(VC)$, $E\left(V^2C\right)$, etc.). The work required to meet the goal will extend, however, into Section 3.2.2. Broadly speaking, the approach is to perform a normal theory analysis of the effect of "pair-wise gene expression correlations" on the distribution of null counts.

We begin by transforming test statistics $T_1, \ldots, T_m$ to z-values:

$$Z_i = \Phi^{-1}\left\{G_0(T_i)\right\}, \, i = 1, \ldots, m, \tag{3.1}$$

where G_0 is the putative null cdf of the test statistic and Φ^{-1} is the inverse cdf of $\mathcal{N}(0,1)$. The z-values provide the analytical convenience of multivariate normal form in describing the joint null statistic behavior. For these calculations it is assumed that all genes are null

$$Z_i = \mathcal{N}(0,1), \, i = 1, \ldots, m, \tag{3.2}$$

so that the theoretical null distribution $\mathcal{N}(0,1)$ is individually correct.

Formally, the quantities of interest are:

$$V = \#\{Z_i : \, Z_i \leq \delta_1\} \tag{3.3a}$$

$$C = \#\{Z_i : |Z_i| \leq \delta_0\}. \tag{3.3b}$$

In Eqn. (3.3), the interval $[-\delta_0, \delta_0]$ coincident with random count C is known as "center-area" and the interval $[-\infty, \delta_1]$ coincident with random count V is known as

"left-sided tail-area." Mostly to be consistent with (Efron, 2007a) we work with a left-sided tail-area, however, the alternative choices, right-sided tail-area or double-sided tail-area, are equally valid. Throughout this work $\delta_1 = -2.5$ and $\delta_0 = 1$ unless stated otherwise.

Central to these developments is the premise of Efron (2007a) that the Z_i's falling in center-area represent null hypotheses, in turn, making count C *nearly* observable. A similar assumption plays a central role in the literature on estimating the proportion of null genes (Pawitan et al., 2005; Langaas et al., 2005). The "empirical null" corrections of Efron (2004) too, rest on a similar logic. The observation of purity in gene expression data motivates this premise. Refer to Section 2.5 for further details.

In order to use the above premise, the proposed strategy is to:

1. Estimate the moments of (V, C)
2. Infer a $P(V, C)$ based on the above moments
3. Report $P(V|C)$ based on determined $P(V, C)$
4. Use determined $P(V|C)$ to find a $\hat{v}|c$, i.e., v conditioned on c

Small improvements may be possible by the incorporation of an estimate of the proportion of truly null cases (Langaas et al., 2005; Storey et al., 2004).

Additionally we assume bivariate and trivariate normality of Z_i's. Efron (2007a) and Owen (2005) in their developments assume bivariate normality, therefore trivariate normality is an added assumption. Through bivariate and trivariate normal assumptions yield a second-order approximation of the "true" $p(Z_1, \ldots, Z_m)$. Since Z_i's are individual $\mathcal{N}(0, 1)$, a second-order approximation is useful. The empirical evidence of Section 3.4 provides support.

The z-value histogram. It is convenient and computationally efficient to approach the moments of (V, C) through the moments of the "Z_i-histogram." This is done by partitioning $\mathcal{Z}$, the sample space of z-values, into K disjoint bins,

$$\mathcal{Z} = \bigcup_{k=1}^{K} \mathcal{Z}_k,$$

where the kth bin has center $z[k]$ and width Δ (constant with k). The histogram counts are:

$$\begin{aligned} Y_k &= \#\{Z_i \in \mathcal{Z}_k\}, && \text{for } k = 1, \ldots, K \\ &= \sum_{i=1}^{m} I_k[i], && \text{for } k = 1, \ldots, K, \end{aligned}$$

where $I_k[j]$ is the indicator random variable for the Z_j falling in bin k. The moment expression derived are for "central moments" for convenience in using the *maxent* approach.

The expectation of Y_k is estimated as:

$$\begin{aligned} \langle Y_k \rangle \equiv E(Y_k) &= E\left\{ \left(\sum_i I_k[i] \right) \right\} \\ &= \sum_{i=1}^{m} \Pr(Z_i \in \mathcal{Z}_k) \\ &= m \cdot \int_{z[k]-\frac{\Delta}{2}}^{z[k]+\frac{\Delta}{2}} \varphi(x)\, dx \\ &= m\Delta\varphi(z[k]) + O(\Delta) \\ &\approx m\Delta\varphi(z[k]), \quad \text{where } \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}. \end{aligned} \tag{3.4}$$

The second moment of the pair (Y_k, Y_l), where k may equal l, is obtained as:

$$\begin{aligned}\mu_2[k,l] &= E\left\{(Y_k - \langle Y_k\rangle)(Y_l - \langle Y_l\rangle)\right\} \\ &= E\left\{\left(\sum_i I_k[i]\right)\left(\sum_j I_l[j]\right)\right\} - \langle Y_k\rangle\langle Y_l\rangle \\ &= \sum_{i\neq j} \Pr(Z_i \in \mathcal{Z}_k, Z_j \in \mathcal{Z}_l) + \sum_i \Pr(Z_i \in \mathcal{Z}_k, Z_i \in \mathcal{Z}_l) - \langle Y_k\rangle\langle Y_l\rangle. \quad (3.5)\end{aligned}$$

If r_{ij} models the correlation of a z-value pair (Z_i, Z_j), then, by adding bivariate normality, Eqn. (3.5) can be approximated as:

$$\mu_2[k,l] \approx \sum_{i\neq j} \frac{\Delta^2}{2\pi\sqrt{1-r_{ij}^2}} \exp\left(\frac{2r_{ij}z[k]z[l] - z[k]^2 - z[l]^2}{2(1-r_{ij}^2)}\right) - \langle Y_k\rangle\langle Y_l\rangle + 1_{k=l}\langle Y_k\rangle, \quad (3.6)$$

where $1_{k=l}$ equals 1 if $k = l$ otherwise 0.

Now, let $g(r)$ denote the empirical density of the pair-wise correlations of $\binom{m}{2}$ (Z_i, Z_j) pairs. This quantity together with Eqn. (3.6) yield a useful approximation for the second moments as stated in the following proposition.

Proposition 3.2.1. *The (central) second moment of a histogram count pair (Y_k, Y_l), where k may equal l, is given by*

$$\mu_2[k,l] \approx m^{\underline{2}}\Delta^2 G_g(z[k], z[l]) - \langle Y_k\rangle\langle Y_l\rangle + 1_{k=l}\langle Y_k\rangle,$$

where

$$G_g(x,y) = \int_{-1}^{+1} \frac{g(r)}{2\pi\sqrt{1-r^2}} \exp\left(\frac{2rxy - x^2 - y^2}{2(1-r^2)}\right) dr$$

and $m^{\underline{k}} = m(m-1)\cdots(m-k+1)$.

The error in the above approximation is of $O(\Delta^2)$.

The third moment of the triplet (Y_k, Y_l, Y_j), where k, l, and j may be equal, is:

$$\begin{aligned}\mu_3[k,l,j] &= E\left\{(Y_k - \langle Y_k\rangle)(Y_l - \langle Y_l\rangle)(Y_j - \langle Y_j\rangle)\right\} \\ &= E(Y_k Y_l Y_j) - \langle Y_k\rangle\langle Y_l\rangle\langle Y_j\rangle - \Big(\langle Y_k\rangle\mu_2[l,j] + \langle Y_l\rangle\mu_2[k,j] + \langle Y_j\rangle\mu_2[k,l]\Big),\end{aligned} \tag{3.7}$$

where:

$$\begin{aligned}E(Y_k Y_l Y_j) &= E\left\{\left(\sum_{p=1}^{m} I_k[p]\right)\left(\sum_{q=1}^{m} I_l[q]\right)\left(\sum_{r=1}^{m} I_j[r]\right)\right\} \\ &= \sum_{p\neq q\neq r} \Pr(Z_p \in \mathcal{Z}_k, Z_q \in \mathcal{Z}_l, Z_r \in \mathcal{Z}_j) \\ &+ 1_{k=j}\sum_{p\neq q} \Pr(Z_p \in \mathcal{Z}_k, Z_q \in \mathcal{Z}_j) + 1_{k=l}\sum_{p\neq q} \Pr(Z_p \in \mathcal{Z}_k, Z_q \in \mathcal{Z}_l) \\ &+ 1_{l=j}\sum_{p\neq q} \Pr(Z_p \in \mathcal{Z}_l, Z_q \in \mathcal{Z}_j) + 1_{k=l=j}\sum_{p} \Pr(Z_p \in \mathcal{Z}_k).\end{aligned}$$

$1_{k=l=j} = 1$ if $k = l = m$, else $1_{k=l=j} = 0$.

Let

$$\mathbf{R}_3[i,j,k] = \begin{pmatrix} 1 & r_{ij} & r_{ik} \\ r_{ij} & 1 & r_{jk} \\ r_{ik} & r_{jk} & 1 \end{pmatrix}$$

denote the 3×3 covariance of the triplet (Z_i, Z_j, Z_k) where $i \neq j \neq k$. Notice that this matrix is an element of the space of 3×3 correlation matrices, say $\mathcal{R}^3$. Notice that the matrices in $\mathcal{R}^3$ are always positive (semi) definite.

Now let $h(\mathbf{R}_3)$ denote the empirical density of all such $\mathbf{R}_3[i,j,k]$ for the true

$p(Z_1, \ldots, Z_m)$. Together with Eqns. (3.4)–(3.7), this yields an approximation for the third moment which is again embodied in a proposition.

Proposition 3.2.2. *The (central) third moment of a histogram count triplet* (Y_k, Y_l, Y_j), *where* k, l, *and* j *may be equal, is given by*

$$\begin{aligned}\mu_3[k,l,j] \approx\ & m^3\Delta^3 H_h(z[k], z[l], z[j]) \\ & + m^2\Delta^2 \left[1_{k=j} G_g(z[k], z[j]) + 1_{k=l} G_g(z[k], z[l]) + 1_{l=j} G_g(z[l], z[j])\right] \\ & + 1_{k=l=j}\langle Y_k\rangle - \langle Y_k\rangle\langle Y_l\rangle\langle Y_j\rangle - \left[\langle Y_k\rangle\mu_2[l,j] + \langle Y_l\rangle\mu_2[k,j] + \langle Y_j\rangle\mu_2[k,l]\right],\end{aligned}$$

where

$$H_h(x,y,z) = \int_{\mathcal{R}^3} \frac{h(\mathbf{R}_3)}{(2\pi)^{3/2}|\mathbf{R}_3|^{1/2}} \exp\left(-\tfrac{1}{2}[x,y,z]\mathbf{R}_3^{-1}[x,y,z]^T\right) d\mathbf{R}_3, \qquad (3.8)$$

and $G_g(x,y)$ *and* $\mu_2[i,j]$ *are defined in Proposition 3.2.1. The error in this approximation is of* $O(\Delta^3)$.

The integral in Eqn. (3.8) is computed over three dimensions. A kth moment calculation would involve $\binom{k}{2}$-D integral and require integration in $\mathcal{R}^{\binom{k}{2}}$.

Next, to get the moments of (V, C) we combine the moments of the corresponding

Y_k's. For example,

$$E\,(V - E(V))^2 = \sum_{\{k,l:\mathcal{Z}_k,\mathcal{Z}_l \subset \Gamma_r\}} \mu_2[k,l] \tag{3.9a}$$

$$E\,(C - (C))^2 = \sum_{\{k,l:\mathcal{Z}_k,\mathcal{Z}_l \subset \Gamma_c\}} \mu_2[k,l] \tag{3.9b}$$

$$E\,\{(V - E(V))(C - (C))\} = \sum_{\{k,l:\mathcal{Z}_k \subset \Gamma_r,\mathcal{Z}_l \subset,\Gamma_c\}} \mu_2[k,l] \tag{3.9c}$$

$$\vdots \tag{3.9d}$$

$$\vdots \tag{3.9e}$$

In Eqn. (3.9): $\Gamma_r \equiv [-\infty, \delta_1]$, tail-area (also called the rejection-region); $\Gamma_c \equiv [-\delta_0, \delta_0]$, center-area.

The key quantities in Proposition 3.2.1 and 3.2.2 are empirical correlation densities—obtaining those in the presence of severe sampling errors is discussed next.

3.2.2 Estimating Correlation Densities

Severe sampling fluctuations create technical challenges: The current methods can recover only $g(r)$, and $h(\mathbf{R}_3)$ requires informed approximations based on $g(r)$. For this reason as well as for ease in the calculations of Proposition 3.2.1 and 3.2.2, we seek to parameterize $g(r)$. Fortunately, for most real examples, a single omnibus parameter β is found to be sufficient. This omnibus measure based approach has an added benefit: Whenever inter-Z_i correlation is inaccessible, due to a complicated definition of the test statistic, the investigator can still exercise judgement to incorporate correlation among the test statistics.

Similar to (Efron, 2007a), we also normalize the columns of $[\mathbf{x_1}; \ldots; \mathbf{x_m}]$ to mean

zero and variance one (but not quantile normalized). This is a usual practice to negate "brightness" disparities among microarrays (Bolstad et al., 2003; Qiu et al., 2005a). Column standardization forces the sum of covariances to be zero. This allows fitting a zero centered density on $g(r)$, which in turn has profound consequences for the form of $h(\mathbf{R}_3)$.

To estimate $g(r)$, we require $\tilde{g}(\rho)$—the empirical density of $\binom{m}{2}$ correlation coefficients between gene expression levels. The mapping between r_{ij} and ρ_{ij} is needed to calibrate $g(r)$. However, for the usual two-sample t-statistic, assuming the independent columns in $[\mathbf{x}; \ldots; \mathbf{x}_m]$, $r_{ij} \approx \rho_{ij}$, and hence nearly the same density $\tilde{g}(\rho)$ and its extension $\tilde{h}(\mathbf{P}_3)$ apply to the Z_i's. Here, the symbol "tilde" distinguishes measurement correlation density from Z_i correlation density. The fact $r_{ij} \approx \rho_{ij}$ can readily be verified through computer simulations. See (Efron, 2007a, Remark A) for a similar discussion. Moreover, recall that this issue of the relationship between r and ρ was discussed in detail in Section 2.1.

Obtaining $\tilde{g}(\rho)$. Let $\hat{\rho}_{ij}$ be the sample correlation coefficient between rows i and j of the residual matrix $[\tilde{\mathbf{x}}_1; \ldots; \tilde{\mathbf{x}}_m]$, obtained by subtracting off each gene's average response within each treatment group. The cumulative sampling errors are removed by transforming to

$$\hat{\tau}_{ij} = \frac{1}{2} \log \frac{1 + \hat{\rho}_{ij}}{1 - \hat{\rho}_{ij}}; \tag{3.10}$$

assuming a translation model $\hat{\tau}_{ij} = \tau_{ij} + \varepsilon, \quad \tau_{ij} \sim \tilde{z}(\tau)$ on this scale; fitting a normal density $\mathcal{N}(0, \sigma^2)$ on $\hat{\tau}_{ij}$ histogram; letting $\varepsilon \sim \mathcal{N}(0, 1/(n-3))$ on the basis of bivariate normality (Owen, 2005); recovering $\tilde{z}(\tau)$ as $\mathcal{N}(0, \sigma^2 - 1/(n-3))$; and transforming back to the ρ scale. If necessary the same calculation can be done in the resampling mode as described in Efron (2007a, Remark A).

Next, on the ρ scale, the modified Beta density is fitted:

$$\tilde{g}(\rho) \propto \left(1-\rho^2\right)^{\beta} = \left(\tfrac{1}{2}(\rho+1)\right)^{\beta}\left(1-\tfrac{1}{2}(\rho+1)\right)^{\beta}, \qquad |\rho| \leq 1. \tag{3.11}$$

Efron (2007a) works with $(0, \alpha^2)$; The present β and Efron's α are related as:

$$\text{Var}(\rho) = \alpha^2 = 4\frac{\beta^2}{(2\beta)^2(2\beta+1)} \Rightarrow \beta = \frac{1-\alpha^2}{2\alpha^2}. \tag{3.12}$$

The rest of this section pursues $\tilde{h}(\mathbf{P}_3)$—a 3×3 extension of $\tilde{g}(\rho)$. For the usual two-sample t-statistic, $h(\mathbf{R}_3) \approx \tilde{h}(\mathbf{P}_3)$ because $r_{ij} \approx \rho_{ij}$. Here, $\mathbf{P}$ should be read as Rho, relating to ρ.

Obtaining $\tilde{h}(\mathbf{P}_3)$. To model $\tilde{h}(\mathbf{P}_3)$, we seek a joint density on $\mathcal{P}^3$—the space of all 3×3 correlation matrices—such that all the inherent marginal densities (i.e., the density of $\rho_{ij}(i \neq j)$, the (i,j)th entry of $\mathbf{P}_3$) are equivalent to $\tilde{g}(\rho)$. Such a density can be obtained from the inverse-Wishart density whose marginalization properties are especially helpful while deducing the probability density of a subset of random variables.

Let the underlying covariance $\mathbf{\Sigma}$ of $\tilde{\mathbf{X}}$ come from the standard inverse-Wishart density $\mathcal{W}_m^{-1}(I, \nu)$, $\nu \geq m$:

$$f_m(\mathbf{\Sigma} \mid \nu) \propto |\mathbf{\Sigma}|^{-(\nu+m+1)/2} \exp\left(-\text{tr}\left\{\mathbf{\Sigma}^{-1}\right\}/2\right), \tag{3.13}$$

where ν is the single parameter that characterizes the density. The goal is to relate ν to β and deduce the probability density of any 3×3 covariance sub-matrices of $\mathbf{\Sigma}$. Recall that while forming a covariance / correlation sub-matrix row and column selection must be identical. To do so we follow the "separation strategy" of Barnard

et al. (2000):

$$\mathbf{\Sigma} = \mathbf{S} \cdot \mathbf{P} \cdot \mathbf{S}, \tag{3.14}$$

where $\mathbf{S}$ is the diagonal matrix whose ith diagonal element, s_i, is the standard deviation of gene i, and $\mathbf{P}$ is the $m \times m$ correlation matrix of $\tilde{\mathbf{X}}$. By back-to-back marginalization yields the density of any $\kappa \times \kappa$ correlation sub-matrix of $\mathbf{P}$:

$$f_m(\mathbf{P}_\kappa \mid \nu) \propto |\mathbf{P}_\kappa|^{(\nu-m+\kappa-1)(\kappa-1)/2-1} \left(\prod_{i=1}^{m} |(\mathbf{P}_\kappa)_{ii}| \right)^{-(\nu-m+\kappa)/2}, \tag{3.15}$$

where $(\mathbf{A})_{ii}$ is the ith principal sub-matrix of $\mathbf{A}$. The reasoning connecting Eqn. (3.13) to Eqn. (3.15) follows.

Under the transformation $\mathbf{\Sigma} \rightarrow (\mathbf{S}, \mathbf{P})$, the Jacobian is given by $2^m \left(\prod_i s_i\right)^m$ [see Theorem 3, (Olkin, 1953)]. Thus, after marginalization over $\mathbf{S}$

$$f_m(\mathbf{P} \mid \nu) \propto |\mathbf{P}|^{-(\nu+m+1)/2} \prod_{i=1}^{m} \int_0^\infty s_i^{-(\nu+1)} \exp\left(-\frac{\rho^{ii}}{2s_i^2}\right) ds_i, \tag{3.16}$$

where ρ^{ii} is the ith diagonal element of $\mathbf{P}^{-1}$. The product occurs because of independence of s_i's. Similarly to (Barnard et al., 2000), substituting $\omega_i = \rho^{ii}/2s_i^2$ yields

$$f_m(\mathbf{P} \mid \nu) \propto |\mathbf{P}|^{-(\nu+m+1)/2} \left(\prod_{i=1}^{m} \rho^{ii} \right)^{-\nu/2} \left(\prod_{i=1}^{m} \int_0^\infty \omega_i^{(\nu-2)/2} \exp\left(-\omega_i\right) d\omega_i \right), \tag{3.17}$$

which leads to an expression for the probability density of the correlation matrix $\mathbf{P}$:

$$f_m(\mathbf{P} \mid \nu) \propto |\mathbf{P}|^{(\nu-1)(m-1)/2-1} \left(\prod_{i=1}^{m} |\mathbf{P}_{ii}| \right)^{-\nu/2}, \tag{3.18}$$

where $\mathbf{P}_{ii}$ is the ith principal sub-matrix of $\mathbf{P}$, and $\rho^{ii} = |\mathbf{P}_{ii}|/|\mathbf{P}|$, with $|\mathbf{A}|$ denoting the determinant of $\mathbf{A}$.

For $\mathbf{P}$ with probability density Eqn. (3.18), the marginal density of an arbitrary $\kappa \times \kappa$ correlation sub-matrix, denoted by $\mathbf{P}_\kappa$, is obtained as follows. Let the $\kappa \times \kappa$ covariance sub-matrix $\mathbf{\Sigma}_\kappa$ undergo the transformation $\mathbf{\Sigma}_\kappa \rightarrow (\mathbf{S}_\kappa, \mathbf{P}_\kappa)$. Then, due to the marginalization property of inverse-Wishart, $\mathbf{\Sigma}_\kappa \sim \mathcal{W}_\kappa^{-1}(I, \nu - m + \kappa)$. Steps Eqns. (3.13)–(3.18) for $\mathbf{\Sigma}_\kappa$ yield:

$$f_m(\mathbf{P}_\kappa \mid \nu) \propto |\mathbf{P}_\kappa|^{(\nu-m+\kappa-1)(\kappa-1)/2-1} \left(\prod_{i=1}^{m} |(\mathbf{P}_\kappa)_{ii}| \right)^{-(\nu-m+\kappa)/2} \qquad \square.$$

Substituting $\kappa = 2$ in (3.15) yields

$$f_m(\mathbf{P}_2 \mid \nu) \equiv f_m(\rho_{12} \mid \nu) \propto \left(1 - \rho_{12}^2\right)^{(\nu-m-1)/2}, \quad |\rho_{12}| \leq 1,$$

which has the same parametric form as Eqn. (3.11). By setting $\nu - m = 2\beta + 1$ we can force the inherent marginal densities of $\mathbf{P}$ entries $[\rho_{ij}\ (i \neq j)]$ to equal $\tilde{g}(\rho)$—the specific aim with which the derivation began. Finally, substituting $\kappa = 3$ in (3.15) gives:

$$\tilde{h}(\mathbf{P}_3) \propto \frac{\left(1 - \rho_{12}^2 - \rho_{23}^2 - \rho_{13}^2 + 2\rho_{12}\rho_{23}\rho_{13}\right)^{2(\beta+1)}}{\left[(1 - \rho_{12}^2)(1 - \rho_{23}^2)(1 - \rho_{13}^2)\right]^{\beta+2}}. \tag{3.19}$$

It should be noted that even though for large m inverse-Wishart, Eqn. (3.13), is a tenuous assumption, its use is strictly to estimate $\tilde{h}(\mathbf{P}_3)$ from the $\tilde{g}(\rho)$, there is no concern for the entire $\mathbf{P}$. Also, since single parameter probability densities on a positive definite matrix space are very few, this one is chosen for its useful "marginalization" property. The justification of this choice rests in the fact that

Eqn. (3.19) is an empirically realistic estimate as evident in the results of Section 3.4. The "Bayesian correlation priors" point of view (Liechty et al., 2004) was especially helpful in formulating these ideas. Exploring other ways to obtain $\tilde{h}(\mathbf{P_3})$ is a subject for future research.

3.2.3 Fitting the Maximum Entropy Distribution

$P(V,C)$ is inherently a discrete distribution with support $\mathcal{D} = \{(i,j) : 0 \leq i \leq m,\, 0 \leq j \leq m,\, 0 \leq i+j \leq m\}$, and any moment based inference would invariably involve computation over $\mathcal{S}$. Growing cardinality of $\mathcal{D}$ $\left(\propto m^2\right)$ makes dealing with large m difficult. However, computation can be reduced substantially by truncating $\mathcal{D}$ to

$$\begin{aligned}
&\mathcal{D}_t = \{(i,j) : V_{\min} \leq i \leq V_{\max},\, C_{\min} \leq j \leq C_{\max},\, 0 \leq i+j \leq m\}, \text{ where} \\
&V_{\min} = \max\left(\lfloor \langle V\rangle - l \cdot \mathrm{Std}(V) \rfloor, 0\right), \quad V_{\max} = \min\left(\lceil \langle V\rangle + l \cdot \mathrm{Std}(V) \rceil, m\right); \\
&C_{\min} = \max\left(\lfloor \langle C\rangle - l \cdot \mathrm{Std}(C) \rfloor, 0\right), \quad C_{\max} = \min\left(\lceil \langle C\rangle + l \cdot \mathrm{Std}(C) \rceil, m\right).
\end{aligned}$$

Here, Chebyshev's inequality guides the choice of parameter l: For $l \geq 6$, the loss of accuracy due to truncation is negligible.

Computation can be reduced even further by recognizing the fact that the distributions imposed on the basis of a small number of moment constraints often enjoy a high-level of regularity and a sparser mesh should suffice. In fact, the computation–accuracy trade-off is easy to deal with, if the task is changed to learning a continuous probability density $p(x,y)$ over continuous support

$$\mathcal{S}_t = \left\{(x,y) : x \in \mathrm{range}(V), y \in \mathrm{range}(C), \frac{-\langle V\rangle - \langle C\rangle}{m} \leq x+y \leq 1 - \frac{\langle V\rangle + \langle C\rangle}{m}\right\}.$$

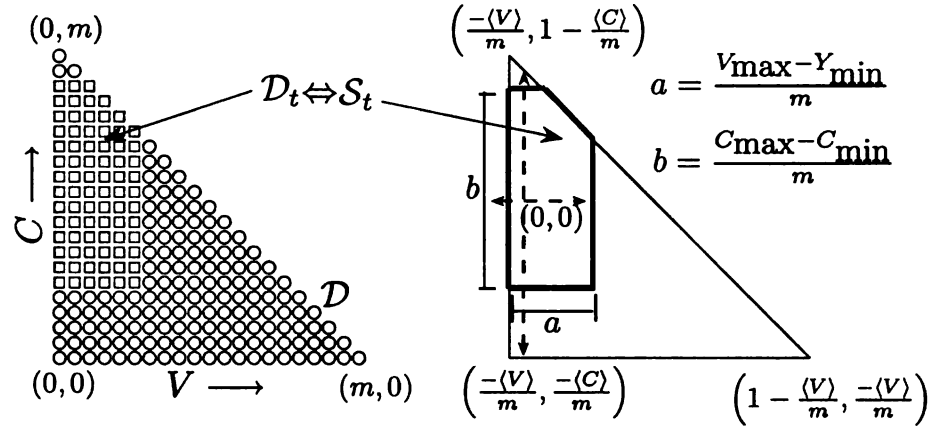


Figure 3.1: Discrete support $\mathcal{D}_t$ (□ markers) versus continuous support $\mathcal{S}_t$ (solid boundary). $\mathcal{S}_t$ is standardized to improve numerical stability.

range(Y) is $\left[\frac{Y_{\min} - \langle Y\rangle}{m}, \frac{Y_{\max} - \langle Y\rangle}{m}\right]$ (see Figure 3.1).

The above standardization offers numerical stability, but the moment constraints must be scaled appropriately. Let $\mathcal{P}_C$ denote the space of feasible $p(x, y)$'s, then $\forall p(x, y) \in \mathcal{P}_C$:

$$\int_{\mathcal{S}_t} x^i y^j p(x, y)\, dx\, dy = \mu^{ij}, \quad \text{and } 0 \leq i + j \leq K; \tag{3.20}$$

$$\text{where } \mu^{ij} = E\{(V - \langle V\rangle)^i (C - \langle C\rangle)^j\}/m^{i+j}. \tag{3.21}$$

In Eqn. (3.21) (i, j)=$(0, 0)$ corresponds to the constraint $\int_{\mathcal{S}_t} p(x, y)\, dx\, dy = 1$, while (i, j)=$(1, 0)$ together with (i, j)=$(0, 1)$ imply that $\forall p \in \mathcal{P}_C$ has mean $(0, 0)$.

Selection of a unique $p(x, y)$ relies on the principle of entropy maximization (*maxent*) which seeks a $p(x, y) \in \mathcal{P}_C$ with maximum information entropy (Jaynes, 2003). The information entropy essentially measures the spread of the distribution, and hence, *maxent* can be seen as a criterion, which within the knowledge constraints, chooses a least "assuming" probability density—arguably, a correct approach in the framework of statistical inference.

The application of *maxent* leads to the following optimization problem:

$$p^*(x,y) = \underset{p(x,y)\in\mathcal{P}_c}{\operatorname{argmax}} \left\{ -\int_{\mathcal{S}_t} p(x,y)\ln p(x,y)\,dx\,dy \right\}. \tag{3.22}$$

The solution takes the following exponential form:

$$p_\lambda(x,y) = \frac{\exp\left(\sum_{1\le i+j\le K}\lambda_{ij}x^i y^j\right)}{\int_{\mathcal{S}_t}\exp\left(\sum_{1\le i+j\le K}\lambda_{ij}x^i y^j\right)dx\,dy}. \tag{3.23}$$

Solving (3.22) requires concepts from the Calculus of variations and Lagrange multipliers. The reasoning leading to the exponential form Eqn. (3.23) and a procedure to determine optimal λ_{ij}'s will now be discussed.

In addition to the fact that information entropy functional is concave (Cover and Thomas, 1991), the constraints in Eqn. (3.21) are also linear in $p(x,y)$. Thus, the problem in Eqn. (3.22) is a convex program which can readily be solved in a Lagrangian dual framework, where one works with an unconstrained upper-bound (lower-bound if minimization) that is easy to optimize. More importantly, in the present case, the framework allows the conversion of the original infinite dimension problem of functional variation into a finite dimensional problem with as few variables as the number of constraints.

Proposition 3.2.3. *The dual $\Psi(\lambda)$ of the concave optimization problem Eqn.* (3.22) *is given by:*

$$\Psi(\lambda) = \ln\left[\int_{\mathcal{S}_t}\exp\left(\sum_{1\le i+j\le K}\lambda_{ij}x^i y^j\right)dx\,dy\right] - \sum_{2\le i+j\le K}\lambda_{ij}\mu^{ij}, \tag{3.24}$$

*where λ_{ij} is the Lagrange multiplier corresponding to the (ij)*th *constraint and μ^{ij},*

and i, j, K are defined in Eqn. (3.21).

Proof. By the definition of the Lagrangian dual function, the Lagrangian $\Psi(\lambda) =$

$$\sup_{p(x,y)\in\mathcal{P}} \left[-\int_{\mathcal{S}_t} p(x,y) \ln p(x,y)\, dx\, dy + \sum_{i+j\leq K} \lambda_{ij} \left(\int_{\mathcal{S}_t} x^i y^j p(x,y)\, dx\, dy - \mu^{ij} \right) \right]. \tag{3.25}$$

Taking the functional variation of the square bracketed term in Eqn. (3.25) with respect to the unknown density $p(x,y)$ and using the fact that $\int_{\mathcal{S}_t} p(x,y)\, dx\, dy = 1$, the maximizer of Eqn. (3.25) is obtained,

$$p_\lambda(x,y) = \frac{\exp\left(\sum_{1\leq i+j\leq K} \lambda_{ij} x^i y^j\right)}{\int_{\mathcal{S}_t} \exp\left(\sum_{1\leq i+j\leq K} \lambda_{ij} x^i y^j\right) dx\, dy}. \tag{3.26}$$

Inserting Eqn. (3.26) into Eqn. (3.25), yields $\Psi(\lambda)$

$$\begin{aligned} &= \int_{\mathcal{S}_t} p(x,y) \ln \left[\int_{\mathcal{S}_t} \exp\left(\sum_{1\leq i+j\leq K} \lambda_{ij} x^i y^j \right) dx\, dy \right] dx\, dy - \sum_{1\leq i+j\leq K} \lambda_{ij} \mu^{ij} \\ &= \ln \left[\int_{\mathcal{S}_t} \exp\left(\sum_{1\leq i+j\leq K} \lambda_{ij} x^i y^j \right) dx\, dy \right] - \textstyle\sum_{2\leq i+j\leq K} \lambda_{ij} \mu^{ij}, \end{aligned} \tag{3.27}$$

where the facts $\int_{\mathcal{S}_t} p(x,y)\, dx\, dy - \mu^{00} = 0$, $\mu^{10} = 0$, and $\mu^{01} = 0$ from Eqn. (3.21) have been used. □

It is easy to verify that the Hessian of Eqn. (3.24) is positive definite and hence $\Psi(\lambda)$ is convex. Suppose λ^* is the minimum of $\Psi(\lambda)$, then the corresponding *primal solution* $p_{\lambda^*}(x,y)$—obtained via Eqn. (3.26)—indeed maximizes Eqn. (3.22). To verify this, let $p^o(x,y)$ be the maximizer of Eqn. (3.22), then from Eqn. (3.25) $\Psi(\lambda) \geq \mathcal{H}\{f^o(x,y)\}, \forall\lambda$. Now from the general optimization theory, the functional

variation of Lagrangian with respect to $p(x,y)$ evaluated at $p^o(x,y)$ must be zero, which implies $p^o(x,y)$ could be written in the form Eqn. (3.26) for some λ^o. But then, $\Psi(\lambda^*) \leq \Psi(\lambda^o) \Rightarrow \Psi(\lambda^*) \leq \mathcal{H}\{f^o(x,y)\}$; consequently, $\Psi(\lambda^*) = \mathcal{H}\{f^o(x,y)\} \Rightarrow \lambda^* = \lambda^o$; hence $p_{\lambda^*}(x,y)$ is the maximizer of Eqn. (3.22).

What remains is the need for an efficient method of minimizing $\Psi(\lambda)$. "Newton's method" is used. According to Newton's method, if a multi-variable function $\Psi(\lambda)$ is twice differentiable and the initial guess point λ_0 is in the "neighborhood" of λ^*, then the sequence

$$\lambda_{n+1} = \lambda_n - \gamma[H\{\Psi(\lambda_n)\}]^{-1}\Delta\Psi(\lambda_n), \quad n \geq 0 \tag{3.28}$$

converges to λ^*. In (3.28) $\Delta\Psi(\lambda_n)$ denotes the gradient of $\Psi(\lambda)$ evaluated at λ_n and $H\{\Psi(\lambda_n)\}$ the Hessian. The parameter $\gamma > 0$ allows finer control of step sizes to avoid numerical instabilities. At the nth iteration, $\Psi(\lambda)$ is replaced by its second-order Taylor expansion around λ_n and then minimized exactly, which produces the minimum λ_{n+1}. At the $(n+1)^{th}$ iteration, λ_{n+1} becomes the point of expansion and the method continues until it convergences.

The elements of the gradient $\Delta\Psi(\tilde{\lambda})$ are given by:

$$\frac{\partial\Psi(\tilde{\lambda})}{\partial\lambda_{ij}} = \int_{\mathcal{S}_t} x^i y^j \left\{ \frac{\exp\left(\sum_{1\leq i+j\leq K} \tilde{\lambda}_{ij} x^i y^j\right)}{\int_{\mathcal{S}_t} \exp\left(\sum_{1\leq i+j\leq K} \tilde{\lambda}_{ij} x^i y^j\right) dx\, dy} \right\} dx\, dy - \mu^{ij} \tag{3.29}$$

$$= \int_{\mathcal{S}_t} x^i y^j \tilde{p}(x,y)\, dx\, dy - \mu^{ij} = \tilde{\mu}^{ij} - \mu^{ij}, \tag{3.30}$$

where $\tilde{\mu}^{ij}$ denotes the (ij)th central moment of the distribution given by $\tilde{\lambda}$ via (3.26).

Similarly, the elements of the Hessian are given by:

$$\frac{\partial^2 \Psi(\tilde{\lambda})}{\partial \lambda_{kl} \partial \lambda_{ij}} = \int_{\mathcal{S}_t} x^{i+k} y^{j+l} \tilde{p}(x,y)\, dx\, dy - \int_{\mathcal{S}_t} x^k y^l \tilde{p}(x,y)\, dx\, dy \cdot \int_{\mathcal{S}_t} x^i y^j \tilde{p}(x,y)\, dx\, dy$$
$$= \tilde{\mu}^{(i+k)(j+l)} - \tilde{\mu}^{kl} \tilde{\mu}^{ij}. \tag{3.31}$$

From (3.30) and (3.31) we observe that the gradient calculations are done as a part of the Hessian calculations which essentially involve terms requiring integration on $\mathcal{S}_t$. A plethora of advanced techniques to carry out numerical integration on a quadrangle like $\mathcal{S}_t$ is available in the literature. An equal-spaced rectangular mesh turns out to be sufficient for the present purpose. The sequence Eqn. (3.28) is initialized with $\lambda = 0$ which implies a uniform distribution over $\mathcal{S}_t$.

3.3 Summary of the Method

Algorithm 1: Inferring the distribution of the number of false discoveries.

Input: $(\mathrm{l}; \mathbf{x_1}; \ldots; \mathbf{x_m})$ (Section 2.1) and δ_1 [Eqn. (3.3)].

Output: An estimate of $P(V|c)$, where (random) tail-count $V = \#\{\text{null } Z_i \in [-\infty, \delta_1]\}$ and center-count $c = \#\{z_i \in [-1, 1]\}$.

1. Compute $z_1, \ldots, z_m$: $z_i = \Phi^{-1}\{G_0(t_i)\}$, Eqn. (3.2). t_i's are linear regression test statistics (Section 2.1), Q is the test statistic cdf, and Φ^{-1} is the standard normal inverse cdf.

2. Determine β, Eqn. (3.12), summarizing the inter-Z_i correlations.

3. Partition $[-5, 5]$, the Z_i sample space, into $K = 100$ equal width bins ($\Delta = 0.1$) and use Eqn. (3.4), Proposition 3.2.1, and Proposition 3.2.2 to compute the first three moments of the Z_i histogram

4. Use determined Z_i histogram moments to compute the first three moments of V and C, Eqn. (3.9)

5. Use determined moments with the *maxent*, Section 3.2.3, to estimate $P(V, C)$

6. Use the estimated $P(V, C)$ to determine $P(V|c)$

3.4 Test Cases

MATLAB code for the algorithm developed above can be requested via email at keyurdesai@gmail.com. The approach was tested on two real data sets, both showing a significant amount of inter-gene correlation and unusual null count behavior. Calculations below are for $\delta_1 = -2.5$ and $\delta_0 = 1$. Comparisons with the second-order $\hat{v}$ of Efron (2007a) are also made.

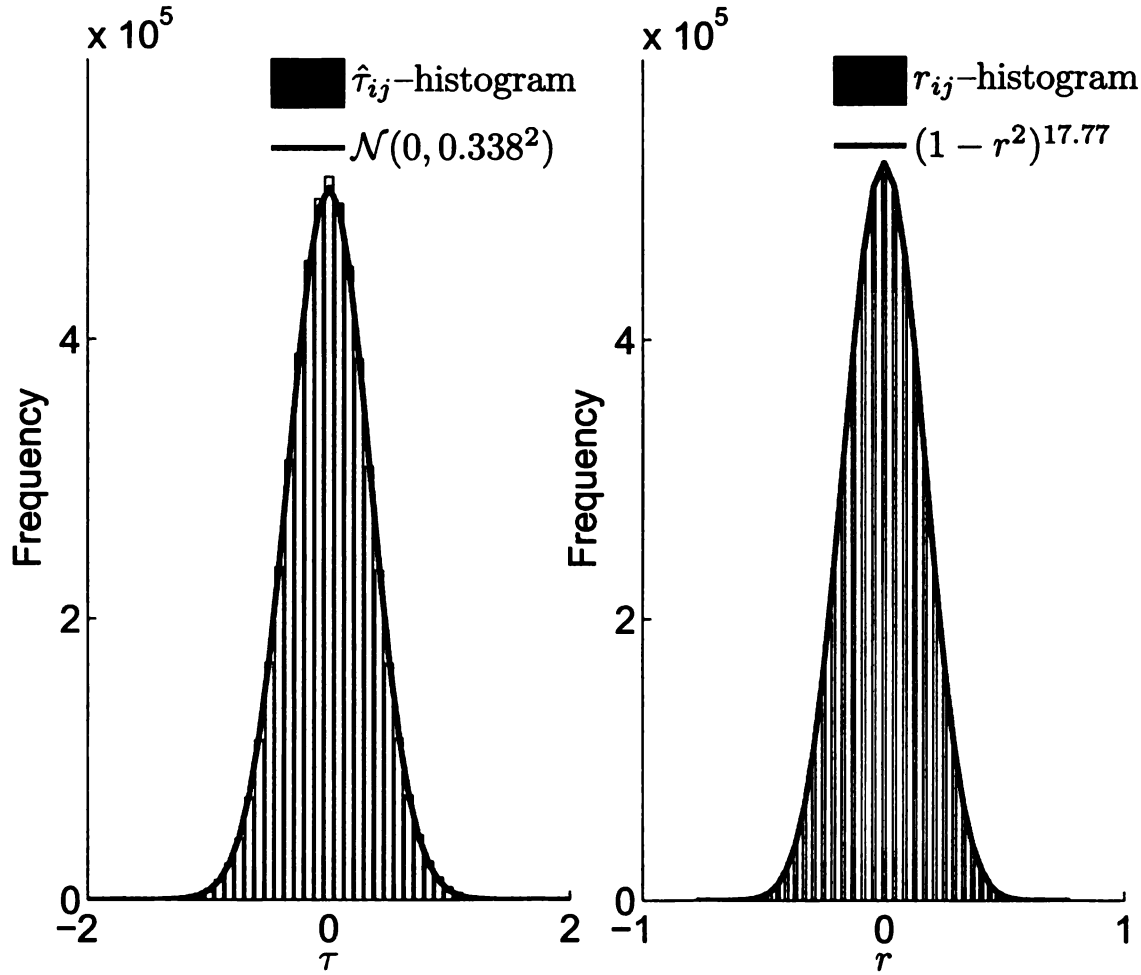
3.4.1 Real Data

The BReast CAncer (BRCA) study of Hedenfalk et al. (2001) has $m = 3226$ and $n = 15$ with 7 samples assigned to BRCA1 mutations and 8 to BRCA2. The study sought to identifying genuine mRNA activity differences between these two categories. The study used two-color microarrays and hence the measurements are in terms of "ratios." The logarithms of these ratios are used to raise normality (Tsai et al., 2003).

The HIV study of van't Wout et al. (2003) has $m = 7680$ and $n = 8$ with 4 samples assigned to an HIV infected condition and the remaining 4 to the control. The control (CD4-T cell lines) was infected by the HIV-1_{BRU} virus. This study reported raw mRNA levels which were also converted to logarithms for the present purpose.

The approach reduces the entire data matrix to just two parameters: The observed C and the β. As evident in Figure 3.2, the parametrization of Section 3.2.2 is realistic. For BRCA example β=17.77 and for HIV β–3.51. Additional details are provided in the caption.

The next step is to compute the moments of (V, C) per Propositions 1 and 2. These



(a)

Figure 3.2: Caption appears on the next page below Figure 3.2.

calculations require m, β, c, δ, and Δ. $\Delta = 0.1$ is selected. A *maxent* distribution was fit to these moments. Figure 3.3 reports the moments and the corresponding *maxent* distribution for the BRCA data. Here, (V, C) show strong negative correlation of -0.89; a similar figure is reported by (Efron, 2007a, Table 1). Furthermore, V shows significant positive skewness, which causes C to show negative skewness. This is not surprising as V is bounded below by 0 and yet has small mean but inflated variance. In effect, third-moment provides additional detail about the joint behavior of (V, C).

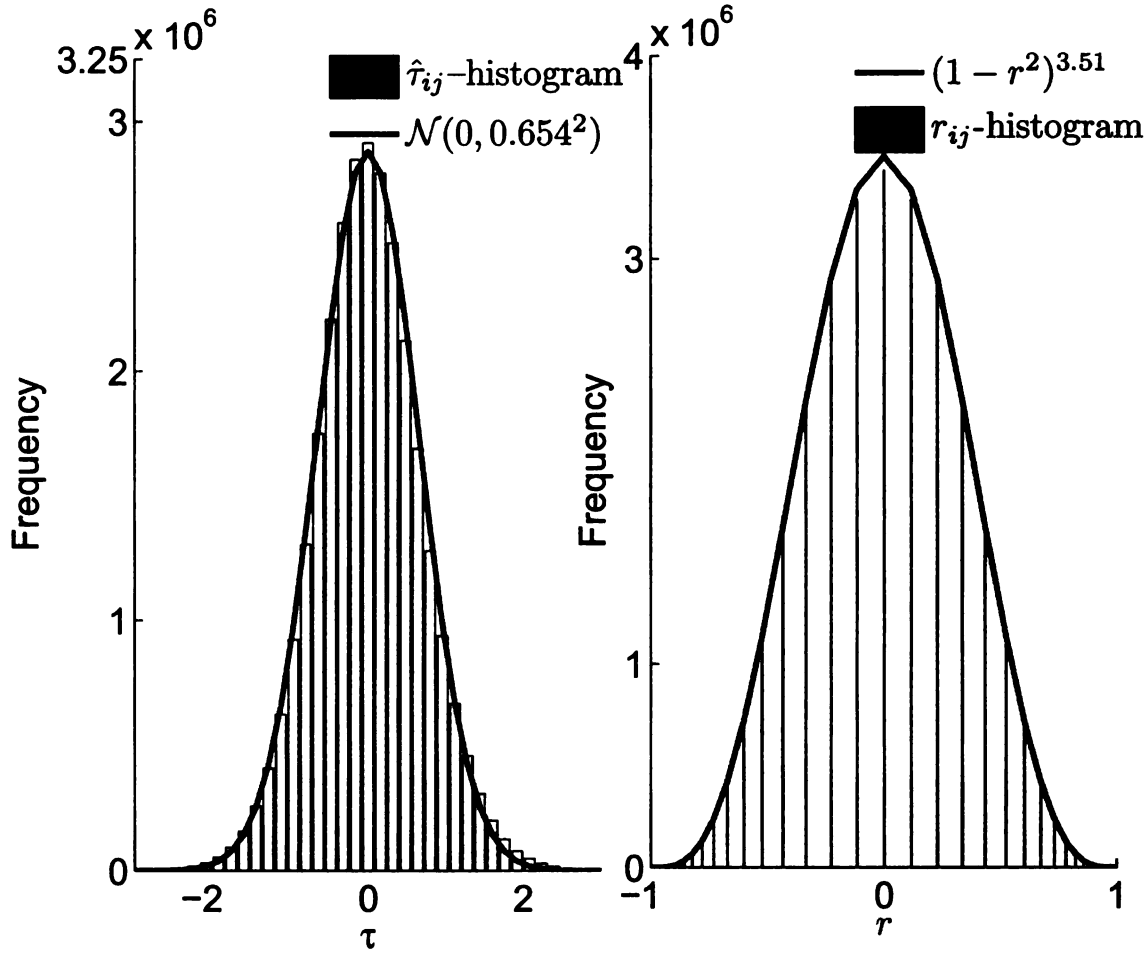


Figure 3.2: Effect of sampling fluctuations on the empirical correlation density. (a) BRCA example (b) HIV example. For each sub figure: Left panel is the histogram of sample correlations after applying the Fisher transformation (3.10) and a normal distribution (heavy curve) fit to it; Right panel is the histogram of de-noised correlations and a modified beta distribution fit to it (heavy curve). This summarizes the cumulative effect of $\binom{m}{2}$ gene-gene correlations in a single parameter β.

During the *maxent* numerical optimization a 100 × 500 equal-spaced mesh was found sufficient for the BRCA study; however, for the HIV study, it was necessary to expand the mesh to 400×2000 because of both larger m and heavier (V, C) correlation. The BRCA optimization took 30 iterations to converge, whereas the HIV optimization

$\langle V \rangle$	19.9
$\langle C \rangle$	2203.0
$E\{(V - \langle V \rangle)^2\}$	289.7
$E\{(C - \langle C \rangle)^2\}$	33113.3
$E\{(V - \langle V \rangle)(C - \langle C \rangle)\}$	-2732.8
$E\{(V - \langle V \rangle)^3\}$	10299.8
$E\{(C - \langle C \rangle)^3\}$	-1399069.6
$E\{(V - \langle V \rangle)^2(C - \langle C \rangle)\}$	-66210.9
$E\{(V - \langle V \rangle)(C - \langle C \rangle)^2\}$	388225.9

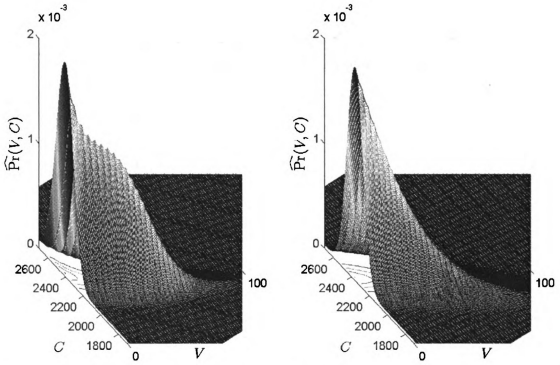


Figure 3.3: BRCA example: Estimated (V, C) moments and a *maxent* distribution fit to it. Third moment estimate of $P(V, C)$ (left) exhibits finer details than the second moment estimate (right).

took 70.

Figure 3.4 reports the estimated $P(V|c)$. Second moment and third moment estimates are shown separately. In the framework of statistical inference, such a distribution is the ultimate goal. Point estimates and associated confidence intervals

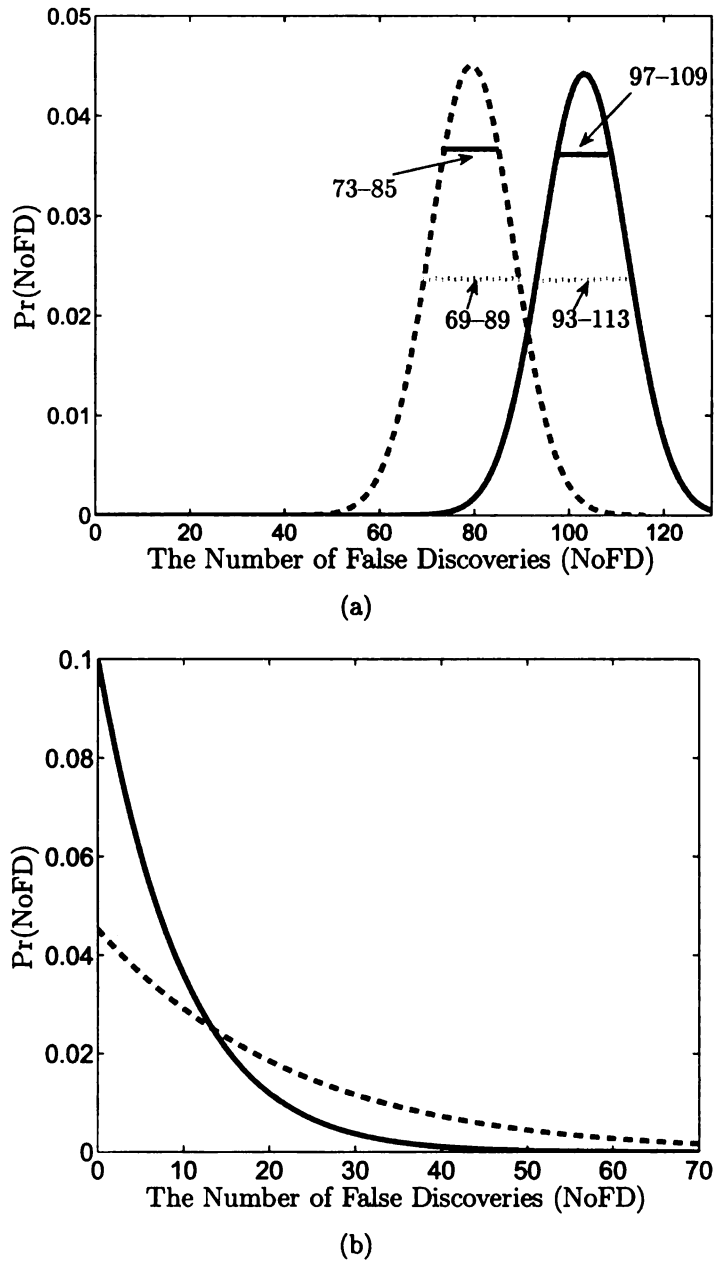


Figure 3.4: The distributions of the number of false discoveries; Panel (a) the BRCA study, Panel (b) the HIV study. To show the effect of skewness corrections, the third moment distribution (solid curve) is compared to its second moment counterpart (dashed curve). For BRCA the second moment mean estimate is 79 compared to 104 for the third-moment; while for HIV these are 19 and 8. The BRCA panel also shows 50% (solid line) and 75% (dotted line) confidence intervals.

can be extracted in accordance with a 'loss function."

If the mean of the estimated $P(V|c)$ is used to determine the realized v, then for the BRCA study, third moment calculations suggests 104 false discoveries versus 79 for second moment. Contrastingly, the "all-null-theoretical-expectation" yields only 20 false discoveries. These numbers must be put in perspective by noting that the actual z_i count falling in the left-sided tail-area $[\infty, -2.5]$ is 116. The standard Benjamini and Hochberg 1995 procedure evaluates the FDP coincident with the left-sided tail-area $[\infty, -2.5]$ as 0.1724, the second moment as 0.68, the third moment as 0.9, and the Benjamini and Yekutieli 2001 procedure as 1 (actually as 1.4759).

For HIV example, third moment calculates 8 false discoveries compared to 19 for second-moment. Contrastingly, the "all-null-theoretical-expectation" suggests 48 false discoveries. This time the actual z_i count in the left-sided tail-area $[\infty, -2.5]$ is 46. In this case, the FDP coincident with the left-sided tail-area $[\infty, -2.5]$ is evaluated by Benjamini and Hochberg 1995 procedure as 1 (actually as 1.0435), the second moment as 0.41, the third moment as 0.17, and the Benjamini and Yekutieli 2001 procedure as 1 (actually as 9.94).

That the conservative scaling in Benjamini and Yekutieli 2001 procedure to ensure "validity" can cause a significant loss in statistical power (Section 2.3), becomes apparent in the HIV example. Whereas "all-null-theoretical-expectation" of Benjamini and Hochberg 1995 procedure can let through a potentially powerless data set, becomes apparent in the BRCA example. However, extensive treatment of inter-gene correlation combined with identifiability may lead to more realistic conclusions.

The second-order $\hat{v}$ developed in (Efron, 2007a) found 77 false discoveries for the BRCA study. Efron compares that to the results of nonparametric analysis and concludes underestimation, but the issue is left unexplored. Our main finding is that

moments higher than second are important in describing the null Z_i histogram.

3.4.2 Simulated Data

It is helpful to test the method on simulated data where the true answer is known. In simulation below all genes are attributed as "null" and hence no treatment effect is added. The objective of this all-null simulation is to evaluate the estimation accuracy of the developed method. The estimated tail-counts are contrasted with the true tail-counts. In each trial, a 3226×15 matrix with entries simulating raw mRNA levels is generated based on the Gamma-Gamma model described below. First 7 columns are assigned "state 1" and the remaining 8 are assigned "state 2." The standard two-sample t-statistic is used enroute to the z-values, $z_1, \ldots, z_{3226}$. The number of z_i's falling in the tail-area is the "true count" which is compared with the "estimated count" from the method.

Let the mRNA level X_{ij} of gene i, measured by jth microarray, be

$$X_{ij} \sim \text{Gamma}(k, \theta_i), \quad \text{for } j = 1, \ldots, n, \tag{3.32}$$

where $\text{Gamma}(k, \theta)$ is the Gamma distribution with the shape parameter $k > 0$ and the scale parameter $\theta > 0$, similarly to the Gamma-Gamma model in (Newton et al., 2001). In (3.32) the shape parameter k is common to all genes. Note that the index variable k of Section 3.2.1 has no connection with this k. The θ_i scale parameters characterize the underlying mRNA levels which vary from gene to gene:

$$\theta_i \overset{\text{i.i.d.}}{\sim} \text{Gamma}(k_0, \theta_0), \quad \text{for } i = 1, \ldots, m. \tag{3.33}$$

The intuition that genes with bigger underlying mRNA levels should have higher variance is consistent with model (3.32) since the mean of the ith gene is $k\theta_i$ and variance is $k\theta_i^2$. The parameters (k, k_0, θ_0) may be chosen on the basis of the overall gene expression histogram of some real microarray data set.

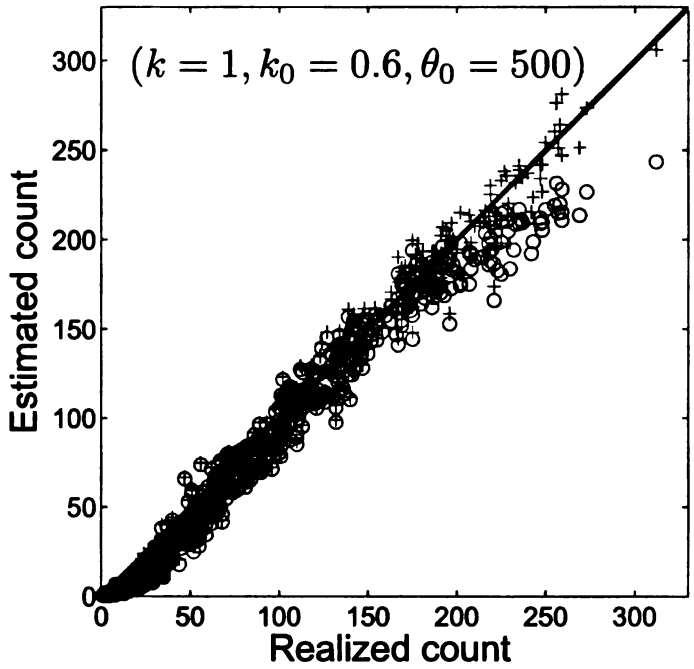
Three sets of results, $(1, 0.6, 500), (2, 0.39, 384)$, and $(3, 0.33, 300)$, are presented. These numbers were chosen to preserve the total sample variance. These particular values are based on the HIV data of van't Wout et al. (2003) which were collected using Affymetrix microarrays. In particular, case k=1 implies that the X_{ij}'s are exponentially distributed, while case k=2 implies a unimodal distribution with heavy tails and a noticeable departure from Gaussianity. Case k=3 is characterized by a more normal (Gaussian) looking distribution, however, with slightly heavier tails.

Following technique was employed to add substantial row-wise correlation: Through the normal cdf, map the entries of a $m \times n$ matrix of correlated Gaussian random variables

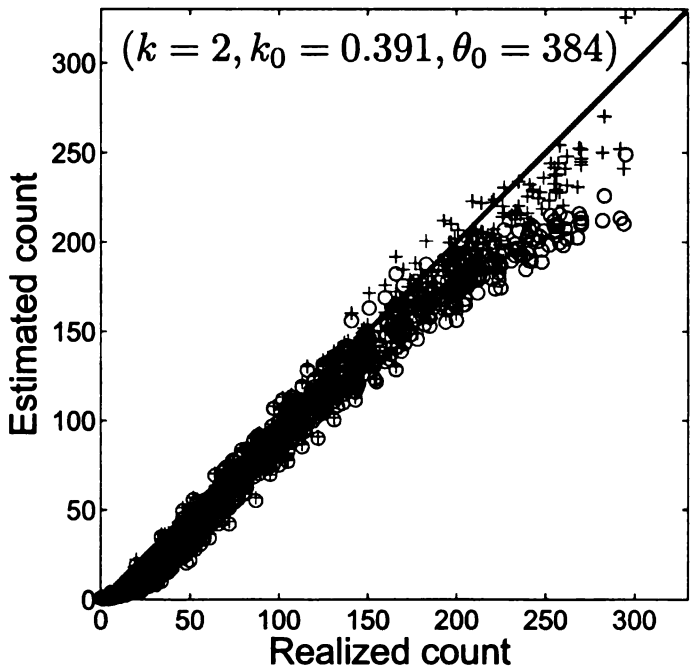
$$\mathbf{Z}^c = \mathbf{L}^T\mathbf{Z}, \quad \text{where } Z_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{3.34}$$

to their p-values $P_{ij} = \Phi(Z_{ij}^c)$, and further map these p-values into X_{ij}'s through the inverse Gamma cdf's as in accordance with Eqn. (3.32). In Eqn. (3.34), $\mathbf{L}^T\mathbf{L} = \mathbf{R}$ is the Cholesky factorization of the correlation matrix $\mathbf{R}$ in which $\mathbf{L}$ is a lower triangular matrix.

Developing proper correlation matrices for simulation is a challenging problem. Even the methods described in the comprehensive work of Marsaglia and Olkin (1984) fail to generate arbitrarily dense correlation matrices for a large m ($m > 1000$). Instead we take a novel approach based on the recent work of Higham (2002) whose basic idea is to generate a symmetric matrix and covert it to a nearest correlation matrix based on Frobenius norm minimization. There are many ways to solve the

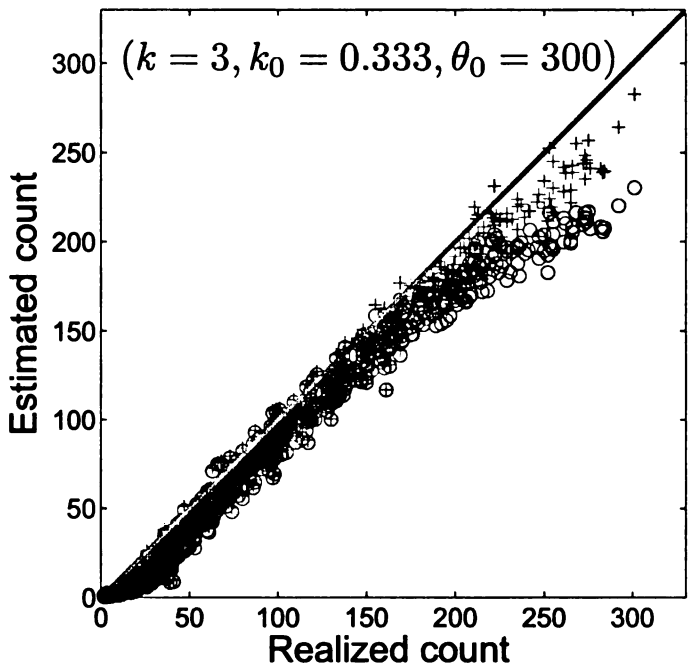


(a)



(b)

Figure 3.5: $\delta_1 = -2.0$. For the caption, refer to panel (c) on the next page.

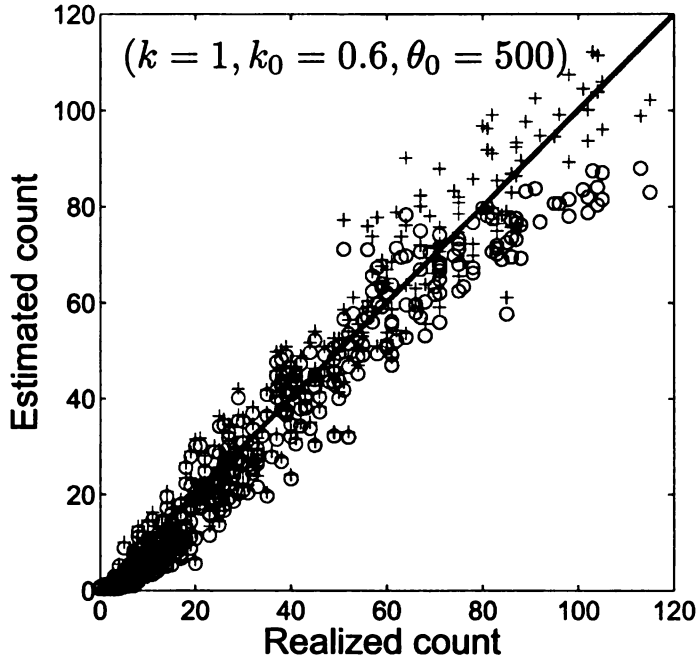


(c)

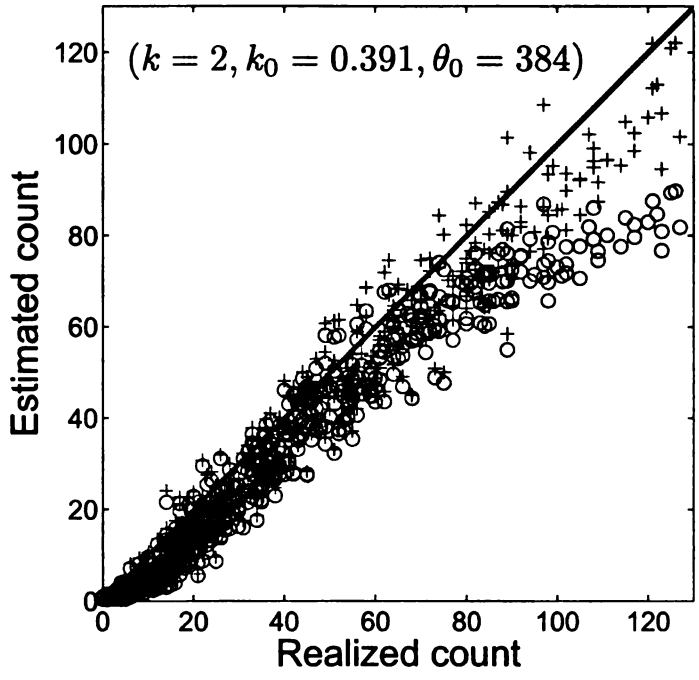
Figure 3.5: Simulation experiments comparing conditional estimates: third moment estimates + marker and second moment o marker. This figure corresponds to the left-sided tail-area with $\delta_1 = -2.0$. Substantial row-wise correlation is present. The abscissa is the realized count while the ordinate is the estimated count. For the significance of different (k, k_0, θ_0)'s refer to the main text. Third moment skewness corrections enhance the estimation accuracy.

numerical optimization proposed in Higham (2002); however, recently proposed Newton method of Qi and Sun (2006) is noteworthy for its speed and accuracy. Higham (2002) provides a lucid exposition of the problem of finding the nearest correlation matrix.

Figures 3.5 and 3.6 compare the second moment estimates versus the third moment estimates for $\delta = -2.0$ and $\delta = -2.5$, respectively. Both cases use $\delta_0 = 1$ for the center-area boundary. The choice of a center-area is discussed in Section 3.5. For each sets of (k, k_0, θ_0), 800 test data were simulated. On each test data the approach was applied in its entirety and no additional knowledge was assumed. Throughout the

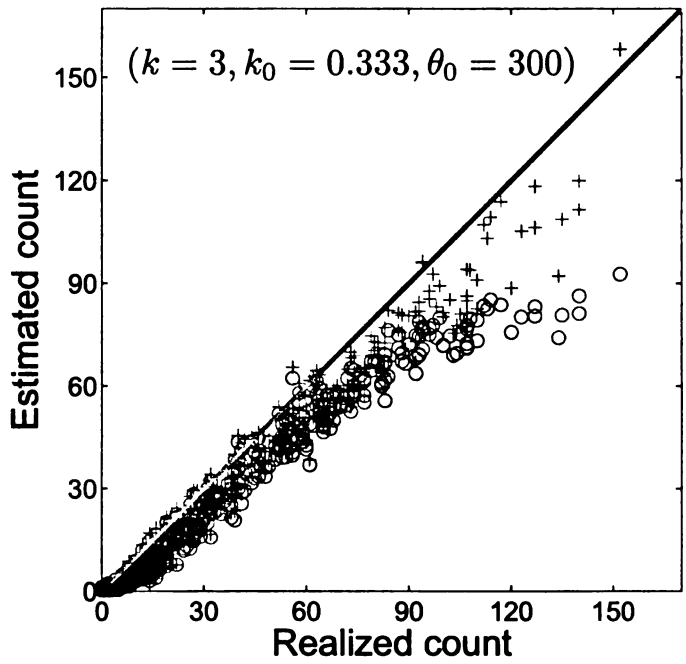


(a)



(b)

Figure 3.6: For the caption, refer to panel (c) on the next page.



(c)

Figure 3.6: Left-sided tail-area with $\delta_1 = -2.5$, else the caption for Figure 3.5 is applicable.

mean of the estimated $P(V/c)$ was used as the final estimate of the tail-area count. Note that the all-null-theoretical-estimates are always 20 for $\delta_1 = -2.5$ and 73 for $\delta_1 = -2.0$, regardless of the test data analyzed.

For all three sets of (k, k_0, θ_0), third moment skewness corrections enhance the estimation accuracy. Evidently, the third moment estimates are significantly less prone to "over / under estimation events" than the second moment estimates. A positive results in three very different simulations emphasize the over-all usefulness of the developed method.

3.5 Discussion

Improved DNA microarray technology, refined standardization procedures, and a careful execution of laboratory protocols collectively lead to testing situations with individually accurate but strongly dependent null hypotheses. Inter-test dependency arising from intrinsic gene-gene interactions cannot be circumvented by experimental design. Therefore, the effects of dependency on the accuracy of decision-making must be carefully analyzed. In particular, due to dependency, the "realized" v/r may vary substantially on a case-to-case basis and the control of $E(V/R)$ may no longer ensure the quality of reported discoveries.

Treating general dependency is impractical, but admitting second-order dependency is possible and widely attempted. The developments in this chapter emphasize that an "explicit" combination of estimated correlation with identifiability can mitigate the unfavorable effects of inter-gene correlation, and that the moment theory of null statistic histogram allows to do so.

It is tempting to conclude that a large number of vectors drawn from the distribution $\mathcal{N}(0, \mathbf{R})$ can yield the required moment estimates and the mathematical work in Sections 3.2.1—3.2.2 is unneeded. However, such a conclusion neglects excessive sampling errors that are present in sample correlation coefficients. With substantial sampling errors, a realistic estimate of the underlying $\mathbf{R}$ cannot be obtained but the quantities $g(r)$ and $h(\mathbf{R}_3)$ are still obtainable.

Permutation calculations, as in Section 4 of (Efron, 2007a), present an alternative way to estimate the moments. They too can run into computational difficulties, especially when the test statistic is computation intensive. An even bigger difficulty is innate when samples are few: For a two-state study like HIV, 4–4 samples each, only 70 unique permutations are available. Nevertheless the permutation based approach

is promising as a subject of further research.

When a direct extraction of inter-test correlation is not feasible, the "single omnibus parameter models" remain useful: They allow the user to systematically exercise judgement by selecting different values for β to examine a range of "correlation effects."

The distribution of interest $P(V, C)$ tends to have complicated support like $\mathcal{D}$ (Figure 3.1), and the *maxent* algorithm is well-suited to such complicated support regions. At a more fundamental level, *maxent* is intended to minimize the amount of unintentional prior information brought into the inference.

For the present approach, apart from the numerical parameters Δ (bin width) and the mesh resolution in *maxent*, the only open choice is δ_0, the center-area boundary. The choice $\delta_0 = 1$ in this work is based on the first eigenvector analysis of (Efron, 2007a) which suggests that (within certain approximations) the interval $[-1, 1]$ has completely opposite count behavior from the rest of the $\mathcal{Z}$ space.

Does more inter-Z_i correlation translate into more extreme $\mathrm{Cor}(V, C)$? The answer is surprisingly no. In the BRCA example, $\mathrm{Cor}(V, C)$ is -0.89, whereas in the HIV example it is reduced to -0.75. Further insight into this behavior should be a useful contribution.

Here, the aim was to mitigate the unfavorable effects of inter-gene correlation when estimating the number false discoveries. In fact, inter-gene correlation can be "exploited" to yield a superior gene-ranking. This is the topic of the next chapter.

Chapter 4

Exploiting Correlation to Improve Gene-ranking

In the LSST DGE detection formulation (Section 2.1), genes are represented by univariate summary statistics. Moreover, there is correlation among test statistics due to inter-gene correlation among gene expression variations. The observation that in high-dimensional inference there can be ways of mapping the original set of statistics $(t_1, \ldots, t_m)$ to a newer set $(t_1^*, \ldots, t_m^*)$ with more statistical power is very appealing. The research described in this chapter develops a technique to obtain one such mapping. The key idea is to combine correlation with purity through a distance metric that can account for the effect of correlation on the joint distribution of T_i's. As a special case, we develop a method that builds upon the widely used two-sample t-statistic approach suitable for two-state studies. The method uses the Mahalanobis distance as the distance metric. An extension accommodating multi-state and continuous-state studies is also discussed.

Organization of the chapter. Section 4.1 provides an overview of the proposed approach. Section 4.2 obtains closed form expressions for the minimum Mahalanobis distance estimates. Section 4.3 builds on the theory of Section 4.2 to develop the tEllipsoid gene-ranking method. In Section 4.4, we apply tEllipsoid to the prostate cancer data of Singh et al. (2002) and evaluate the gain in statistical power. Section 4.5 discusses the implications of these results.

4.1 Overview

Detecting differentially expressed genes in the presence of substantial inter-gene correlation is a challenging problem. Research has focused largely on understanding the harmful effects of correlation on the threshold settings demarcating null and non-null genes. The research discussed in Chapter 3 developed a method which explicitly admits the observed sample correlation in the analysis and offers a more accurate assessment of significance cut-offs. The research described in this chapter shows that correlation can, in fact, be exploited to share information across tests, which, in turn, can increase statistical power. The key contributions in Chapter 3 were in part motivated by the exceedingly small sample situations ($n \sim 5$–10), whereas the work in this chapter is intended to benefit situations with $n \geq 20$.

It is helpful to think in terms of mapping the original set of observed statistics $(t_1, \ldots, t_m)$ to a newer set $(t_1^*, \ldots, t_m^*)$ for better statistical power (Section 2.3). The literature is not devoid of attempts to develop such mappings that exploit correlation among (the random variable interpretation of these) test statistics, but such efforts have not produced compelling results. We posit that the limitations of such developments are due, at least in part, to neglecting *identifiability*—the act of incorporating

the genes whose test statistics fall near the center of the null distribution as *explicitly* null in the inference. The statistical risk of imposing identifiability weigh less than not doing so because of the fact that in most comparative studies the activity of a great majority of genes remains unchanged with respect to the treatment of interest (Section 2.5).

This chapter presents a framework to obtain a better gene-ranking by combining correlation and identifiability through an optimization criterion. The framework builds upon the widely-used two-sample t-statistic approach and uses the Mahalanobis distance as the optimality criterion. Although the initial motivation was to improve statistical inference in two-state microarray studies, the framework readily generalizes to multi-states and continuous-states as well as to other multiple comparison applications. The connection between the standard t-statistic and *simple linear regression* as discussed in Section 2.1 is crucial to this generalization.

Recall the basic LSST formulation from Section 2.1: Genes are represented by summary statistics and the corresponding null hypotheses,

$$H_1, H_2, \ldots, H_m$$

$$T_1, T_2, \ldots, T_m$$

$$t_1, t_2, \ldots, t_m,$$

where T_i refers to a "random value" and t_i the "observed value." The magnitudes of t_i's establish a gene-ranking, and the top $r \ll m$ genes with the largest t scores are reported as statistically significant discoveries. The investigator can either explicitly supply r or rely on the *false discovery rate* (FDR) calculations to find a maximal r with the allowable FDR. The present discussion assumes that r is fixed.

The issue of correctly estimating the FDP in the presence of correlation has received much recent attention because highly correlated tests increase the variance of the FDP leading to unreliable results (Owen, 2005). As discussed in Efron (2007a) and Chapter 3, for "over powered" data sets, there may be significantly fewer tail-area null counts than expected, while for "under powered" data sets, the situation can worsen with many more tail-area null counts than expected. Importantly though, techniques for correctly estimating the FDP do not change the gene-ranking, but only the size of the reported list.

The research discussed in this chapter was motivated by the notion that, for "under powered" data sets, it might be possible to exploit correlation among test statistics to establish a gene ranking that has better statistical power than the original t_i based ranking. The method that resulted from an exploration of this question indeed seems to improve the statistical power of *all* data sets. The proposed method uses, (i) a vector of the observed statistics $\mathbf{t} = [t_1, t_2, \ldots, t_m]^T$ and (ii) an estimate of the covariance matrix of the vector $\mathbf{T} = [T_1, T_2, \ldots, T_m]^T$, to output a substantially revised version of $\mathbf{t}$, denoted $\mathbf{u}|\mathbf{t}$, whose corresponding entries can be used to establish an improved gene-ranking.

The method is summarized as follows. Let $\mathbf{T} \sim (\mathbf{u}, \Sigma)$. Note that no distributional information nor higher order statistics of $\mathbf{T}$ are assumed. Now based on the observed value $\mathbf{t}$, we can estimate $\mathbf{u}|\mathbf{t}$, but while doing so we invoke the *zero assumption* (ZA) (Efron, 2008) that the *smallest* $P_0(\%)$ of t_i's are associated with null genes. Based on the ZA, we can set corresponding entries of $\mathbf{u}$ to zero. For the remaining entries of $\mathbf{u}$ we obtain *minimum Mahalanobis distance* (Mahalanobis, 1936) estimates.

Inter-gene correlation causes the vector $\mathbf{T}$ to distribute around the center of mass $\mathbf{u}$ in an hyperellipsoidal manner, and the Mahalanobis distance is a natural way to

measure vector distances in such a distribution. In fact, to emphasize the geometric intuition of tracking the center of an ellipsoid, the method is named as "tEllipsoid." Through extensive experimentation with both real and simulated data, it has been found that for a truly null t_i which happens to be in tail-area, the corresponding $u_i|\mathrm{t}$ consistently tends to zero (its theoretical value).

Two prior research efforts, Storey et al. (2007) and Tibshirani and Wasserman (2006), were particularly useful in formulating the present approach. Interestingly, both approaches aim at exploiting the signal structure among non-null tests, whereas the present approach aims at exploiting the structure among null tests; see Section 2.3 for details.

4.2 The Proposed Method

An $m \times n$ matrix of gene expressions, for m genes and n samples, is given. For the present discussion we assume that the samples fall into two states $k = 1$ and $k = 2$ and there are n_k samples in group k with $n_1 + n_2 = n$. The generalization to multi-state and continuous-state is discussed at the end of the section. We start with the standard (unpaired) t-statistic:

$$t_i = \frac{\bar{x}_{i;2} - \bar{x}_{i;1}}{s_i}, \tag{4.1}$$

where $\bar{x}_{i;k}$ is the mean of gene i in group k and s_i is the pooled within-group standard deviation of gene i. If the i^{th} gene is indeed null, then we expect the random variable $T_i \sim (0, \nu/(\nu-2))$. Here, the degrees of freedom ν is obtained from either the unpaired t-test theory or the permutation null calculations as discussed in Section 2.1. Note that T_i is a random variable and t_i is its observed value. If gene i is non-null, then

we expect $T_i \sim (u_i, \sigma_i^2)$. For non-null genes, the values of u_i and σ_i depend on the amount of up / down regulation, the number of samples in each group, and ν.

Without loss of generality, we may assume that the genes are indexed so that

$$|t_1| \leq |t_2| \leq \cdots \leq |t_m|. \tag{4.2}$$

Then, a reasonable way to impose identifiability on null genes is through the ZA, namely, that $P_0(\%)$ of the genes—those with the smallest t statistics—are null. Section 2.5 provides a comprehensive discussion regarding the rationale behind the ZA. The use of the ZA is justified in the present situation as long as P_0 is sufficiently small so that the bottom $P_0(\%)$ genes would almost certainly be declared null for reasonable FDR's. Potentially, the statistical risk of "imposing" identifiability weigh less than not imposing it, especially when purity holds. Empirical evidence of Section 4.4 supports this intuition.

Formally, the ZA is stated as follows: Let c be the largest integer (gene index) such that $c/m \leq P_0/100$, denoted

$$c = \lceil 0.01 m P_0 \rceil, \tag{4.3}$$

then genes with indices $1, 2, \ldots, c$ are assumed null. Let us partition the set of t statistics into those corresponding to genes declared null under the ZA, $\{t_1, t_2, \ldots, t_c\}$, and those for the remaining $m - c$ genes which continue to *compete* for the non-null designation, $\{t_{c+1}, t_{c+2}, \ldots, t_m\}$. (The present c is different from the c in Chapter 3.) For convenience, we introduce the following vector notation,

$$\mathbf{t} = \left[\begin{array}{cc} \mathbf{t}_{1:c}^T & \mathbf{t}_{c+1:m}^T \end{array} \right]^T = \left[\begin{array}{cc} \mathbf{t}_{(0)}^T & \mathbf{t}_{(1)}^T \end{array} \right]^T.$$

Then the random vector $\mathbf{T}$ is distributed in the following way:

$$\mathbf{T} \sim (\mathbf{u}, \boldsymbol{\Sigma}), \tag{4.4}$$

where $\mathbf{u}$ is the underlying mean vector and $\boldsymbol{\Sigma}$ the covariance matrix. The corresponding partitions of $(\mathbf{u}, \boldsymbol{\Sigma})$ are denoted

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_{(0)} \\ \mathbf{u}_{(1)} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{(00)} & \boldsymbol{\Sigma}_{(01)} \\ \boldsymbol{\Sigma}_{(10)} & \boldsymbol{\Sigma}_{(11)} \end{pmatrix}.$$

The central hypothesis here is that there is a vector, say $\mathbf{u}|\mathbf{t}$, whose elements represent a reordering of the elements of the $\mathbf{t}$, such that gene-ranking represented by $\mathbf{u}|\mathbf{t}$ has better statistical power for detecting non-null genes than that based on $\mathbf{t}$ itself.

The present effort focuses mainly on the second moment distributional characteristics of $\mathbf{t}$. However, in fact, if the gene expressions are normally distributed, then, perhaps, $\mathbf{t}$ is described more accurately by the multivariate Student distribution. Exploitation of this additional structure will be considered in future work.

4.2.1 Choosing a Distance Metric

We are interested in obtaining an estimate of the vector mean $\mathbf{u}$ based on the observation $\mathbf{t}$. This requires an appropriate metric in the space of $\mathbf{t}$ vectors, with which to quantify the distance of the observed $\mathbf{t}$ from the center of mass $\mathbf{u}$, say $\mathbf{dist}(\mathbf{t}, \mathbf{u})$. The ℓ_2 norm induces a useful metric between $\mathbf{t}$ and $\mathbf{u}$ provided that we first decorrelate

the vector elements as $\Sigma^{-1/2}(\mathbf{t}-\mathbf{u})$, thus yielding

$$\begin{aligned}\mathbf{dist}(\mathbf{t},\mathbf{u}) &= \sqrt{||\,\Sigma^{-1/2}(\mathbf{t}-\mathbf{u})\,||^2} \\ &= \sqrt{(\mathbf{t}-\mathbf{u})^T\Sigma^{-1}(\mathbf{t}-\mathbf{u})}. \end{aligned} \tag{4.5}$$

This weighted Euclidean distance is sometimes called the *Mahalanobis distance* in the pattern classification literature as in Deller et al. (1999) and Dejviver and Kittler (1982).

We can relate $\mathbf{t}$ to $\mathbf{u}$ through the Mahalanobis distance but while doing so we invoke the ZA, which, in turn, implies that the first c entries of $\mathbf{u}$ are zero. This yields the estimate

$$\mathbf{u}^* = \begin{pmatrix} \mathbf{0} \\ \mathbf{u}^*_{(1)} \end{pmatrix}, \quad \text{where} \quad \mathbf{u}^*_{(1)} = \underset{\mathbf{u}_{(1)}\in\mathbb{R}^{m-c}}{\operatorname{argmin}} \tag{4.6}$$
$$\begin{pmatrix} \mathbf{t}_{(0)}-\mathbf{0} \\ \mathbf{t}_{(1)}-\mathbf{u}_{(1)} \end{pmatrix}^T \begin{pmatrix} \Sigma_{(00)} & \Sigma_{(01)} \\ \Sigma_{(10)} & \Sigma_{(11)} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{t}_{(0)}-\mathbf{0} \\ \mathbf{t}_{(1)}-\mathbf{u}_{(1)} \end{pmatrix}.$$

In effect, $\mathbf{u}^*_{(1)}$ combines the identifiability information based on the ZA with the information about the covariance structure of $\mathbf{T}$ which too can be obtained from the measured X itself. Notably the optimization in Eqn. (4.6) enjoys closed form solution:

$$\mathbf{u}^*_{(1)} = \mathbf{t}_{(1)} - \Sigma_{(10)}\Sigma^{-1}_{(00)}\mathbf{t}_{(0)}. \tag{4.7}$$

The derivation leading from Eqn. (4.6) to Eqn. (4.7) follows.

Suppose that

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \text{ and } \tilde{\mathbf{u}}_{(1)} = \mathbf{t}_{(1)} - \mathbf{u}_{(1)}. \tag{4.8}$$

We also have $\mathbf{C} = \mathbf{B}^T$. Substituting these in Eqn. (4.6) yields:

$$\tilde{\mathbf{u}}^*_{(1)} = \underset{\tilde{\mathbf{u}}_{(1)} \in \mathbb{R}^{m-c}}{\operatorname{argmin}} \mathbf{t}^T_{(0)} \mathbf{A} \mathbf{t}_{(0)} + 2\tilde{\mathbf{u}}^T_{(1)} \mathbf{C} \mathbf{t}_{(0)} + \tilde{\mathbf{u}}^T_{(1)} \mathbf{D} \tilde{\mathbf{u}}_{(1)}. \tag{4.9}$$

In Eqn. (4.9), by setting the gradient w.r.t $\tilde{\mathbf{u}}_{(1)}$ to 0, we obtain:

$$\tilde{\mathbf{u}}^*_{(1)} = -\mathbf{C}^T \mathbf{D}^{-1} \mathbf{t}_{(0)}. \tag{4.10}$$

Now for $\boldsymbol{\Sigma}^{-1}$, we can appeal to the matrix inversion lemma (Golub and Van Loan, 1996):

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}^{-1}_{(00)} \left(1 + \boldsymbol{\Sigma}_{(01)} \mathbf{Q}^{-1} \boldsymbol{\Sigma}_{(10)} \boldsymbol{\Sigma}^{-1}_{(00)}\right) & -\boldsymbol{\Sigma}^{-1}_{(00)} \boldsymbol{\Sigma}_{(01)} \mathbf{Q}^{-1} \\ -\mathbf{Q}^{-1} \boldsymbol{\Sigma}_{(10)} \boldsymbol{\Sigma}^{-1}_{(00)} & \mathbf{Q}^{-1} \end{pmatrix},$$

where $\mathbf{Q} = \boldsymbol{\Sigma}_{(11)} - \boldsymbol{\Sigma}_{(10)} \boldsymbol{\Sigma}^{-1}_{(00)} \boldsymbol{\Sigma}_{(01)}$. Plugging this in Eqn. (4.10) yields:

$$\tilde{\mathbf{u}}^*_{(1)} = \boldsymbol{\Sigma}_{(10)} \boldsymbol{\Sigma}^{-1}_{(00)} \mathbf{t}_{(0)}. \tag{4.11}$$

Combining Eqn. (4.11) with Eqn. (4.8) provides the desired expression:

$$\mathbf{u}^*_{(1)} = \mathbf{t}_{(1)} - \boldsymbol{\Sigma}_{(10)} \boldsymbol{\Sigma}^{-1}_{(00)} \mathbf{t}_{(0)} \quad \square.$$

4.2.2 An Intuitive Interpretation

If $\mathbf{t}_{(1)}$ is written as $\mathbf{u}_{(1)} + \mathbf{e}_{(1)}$, then Eqn. (4.7) can be seen as

$$\mathbf{u}^*_{(1)} = \mathbf{u}_{(1)} + \left[\mathbf{e}_{(1)} - \boldsymbol{\Sigma}_{(10)}\boldsymbol{\Sigma}^{-1}_{(00)}\mathbf{t}_{(0)}\right].$$

Then, the proposed method is interpreted in the following way: The method uses the identifiable null cases to predict and, in turn, suppress the "null energy" in the competing cases.

4.2.3 Estimating the Covariance

Notice that Eqn. (4.7) involves theoretical covariances whose estimates must be obtained from the data set at hand. To estimate the required entries of $\boldsymbol{\Sigma}$, we make two observations. The validity of these observations can be established through computer simulations. An intuitive argument also appears. Note that Eqn. (4.7) does not require the covariance between two non-null T_i's. The realization of the fact that the t-statistic is a "standardized" statistic helps understand these covariance—correlation relationships.

Observation 1. If genes i and i' both are null, then

$$\text{Cov}\left(T_i, T_{i'}\right) \approx \text{Cor}\left(X_i, X_{i'}\right) \frac{\nu}{\nu - 2} \tag{4.12}$$

This observation maybe intuitive to the reader from Eqn. (4.1) itself, or it is easily verified through a computer simulation. Efron (2007a), Owen (2005), and the method developed in Chapter 3 all use this observation for their respective conditional FDR calculations.

Observation 2. Similarly, if the gene i is null and i' non-null (or conversely), then

$$\text{Cov}\left(T_i, T_{i'}\right) \approx \text{Cor}\left(\tilde{X}_i, \tilde{X}_{i'}\right) \frac{\nu}{\nu - 2}, \tag{4.13}$$

where $\text{Cor}(\tilde{X}_i, \tilde{X}_{i'})$ denotes the "residual" correlation between gene i and i'. Residual correlation is the correlation in the fraction of gene expression that is unaffected by the treatment of interest. For null genes the fraction is one because by definition the gene expression X_i for a null gene is independent of the state variable L. For a non-null gene, the fraction is determined by how strongly that gene is affected by the treatment.

Equations (4.12) and (4.13) suggest to use "residual" sample correlations to estimate $\text{Cov}(t_i, t_{i'})$:

$$\widehat{\text{Cov}}\left(T_i, T_{i'}\right) \propto \frac{\sum_j \tilde{x}_{ij}\tilde{x}_{i'j}}{\sqrt{\left(\sum_j \tilde{x}_{ij}^2\right)\left(\sum_j \tilde{x}_{i'j}^2\right)}}, \tag{4.14}$$

where $\tilde{x}_{ij}$ denotes the (observed) residual corresponding to the i^{th} gene and the j^{th} microarray. The scale factors cancel in the terms $\boldsymbol{\Sigma}_{(10)}$ and $\boldsymbol{\Sigma}_{(00)}^{-1}$, so that estimating $\nu/(\nu\text{-}2)$ is not required. These arguments hold for two-state, multi-state, and even continuous-state studies provided the test statistics are obtained through *simple linear regression.* Here, the entries of the residual matrix are the differences between the observed expression values and the corresponding predicted expression values from regression equations $X_i = a + bL + \epsilon, i = 1, \ldots, m$. Note that the two-sample t-statistic can be interpreted as simple linear regression with a binary-valued covariate. The situations where test statistics are not from *simple linear regression* will be considered in future work.

4.2.4 The Final Equation

In light of Eqn. (4.14), Eqn. (4.7) takes the practical form

$$\widehat{\mathbf{u}}^*_{(1)} = \mathbf{t}_{(1)} - \widetilde{\mathbf{C}}_{(10)}\widetilde{\mathbf{C}}^{-1}_{(00)}\mathbf{t}_{(0)}, \tag{4.15}$$

where $\widetilde{\mathbf{C}}$ is the sample correlation matrix of the residuals. In most cases computing the full matrix inverse $\left(\widetilde{\mathbf{C}}^{-1}_{(00)}\right)$ is not necessary and solving the term $\widetilde{\mathbf{C}}^{-1}_{(00)}\mathbf{t}_{(0)}$ through an efficient linear solver reduces the computation considerably.

4.3 Implementation Details

This section outlines a self-contained differential analysis algorithm based on the ideas discussed in Section 4.2. Its name tEllipsoid was coined to emphasize the geometric intuition of tracking the center of an hyper-ellipsoid.

TEllipsoid takes a gene expression matrix $\mathcal{X}$ and assigned biological conditions and provides a specified number, say r, of the most differentially expressed genes. In principle, the ranking is based on the set $\{u_i^*\}$ from Eqn. (4.6). In practice, we rely on the estimates $\{\hat{u}_i^*\}$ from Eqn. (4.15).

Two-state implementation begins by re-indexing the genes based on their two-sample t-statistics [Eqn. (4.2)]. Then, based on the ZA, the first c genes are identified as null, as specified in Eqn. (4.3). By default, P_0 is set to 50(%). Although the choice 50% is somewhat arbitrary, this fraction has worked well empirically in the data sets tested. A P_0 as low as 10(%) is found to enhance the power. Future research may yield more rigorous methods for choosing P_0.

In the two-state implementation, in order to nullify any genuine treatment differ-

ences, $\mathcal{X}$ is converted to $\widetilde{\mathcal{X}}$ by subtracting each gene's average response within each treatment group. The residual sample correlation matrix $\widetilde{\mathbf{C}}$ of $\widetilde{\mathcal{X}}$ is computed subsequently. The crucial step is to compute $\widehat{\mathbf{u}}^*_{(1)}$ based on Eqn. (4.15). The elements of $(\widehat{\mathbf{u}}^*)^T = \left[\mathbf{0}^T_{c\times 1} \left(\widehat{\mathbf{u}}^*_{(1)}\right)^T\right]$ determine the gene-ranking: A gene with bigger $|\widehat{u}^*_i|$ is assigned a higher rank. The first r genes are reported as top r statistical discoveries.

The multi-state / continuous-state implementation beings by re-indexing the genes based on their *simple linear regression* t-statistic. The entries of the residual matrix $\widetilde{\mathcal{X}}$ are the differences between the observed expression values and the corresponding predicted expression values from regression equations $X_i = a + bL + \epsilon, i = 1, \ldots, m$.

4.3.1 Numerical stability and Computational Complexity

Because the number of samples n is often less than the number of genes m, the residual sample correlation matrix turns out to be singular and hence non-invertible. Therefore, we add a very small correction term $\left(= 10^{-10}\right)$ to its diagonal entries to make it nonsingular, and in effect, invertible. After this correction, tEllipsoid shows excellent numerical stability.

Equation (4.15) involves matrix inversion, which, if performed in a naive way, could be a prohibitive operation, since microarray data sets may have several tens of thousand genes. Indeed, solving the term $\widetilde{\mathbf{C}}^{-1}_{(00)}\mathbf{t}_{(0)}$ as a system of simultaneous linear equations $\left(\widetilde{\mathbf{C}}_{(00)}\mathbf{x} = \mathbf{t}_{(0)}\right)$ is much faster than explicitly computing $\widetilde{\mathbf{C}}^{-1}_{(00)}$. In particular, we can employ the Cholesky decomposition to exploit the fact that the matrix $\widetilde{\mathbf{C}}_{(00)}$ is symmetric and positive definite. MATLAB implementation of tEllipsoid uses the in-built `linslove` with appropriate settings, which, in turn, uses highly optimized routines of LAPACK (Linear Algebra PACKage—`http://www.netlib.org/lapack/`).

For the Prostate data (used in Section 4.4) with 12625 genes and 102 samples, tEllipsoid, running on a computer with a 2.2 GHz dual-core AMD Opteron processor with 8 GB of RAM and MATLAB version R2006b, requires just under 40 seconds to report the final gene list. For the same settings, the implementation with explicit matrix inversions takes $\sim$ 10 minutes.

4.3.2 Algorithm for Two-State Studies

---

tEllipsoid: An enhanced gene-ranking for differential gene expression detection

Input: $\mathcal{X}$ = Labeled $m \times n$ gene expression matrix; r = Size of gene list

Output: The gene list containing top r differentially expressed genes

1. Calculate two-sample (unpaired) t-statistics: $t_i = (\bar{x}_{i;2} - \bar{x}_{i;1})/s_i$
2. Reindex genes such that $|t_1| \leq |t_2| \leq \cdots \leq |t_m|$
3. Gather first $c = \lceil 0.01 m P_0 \rceil$ t_i's in a vector $\mathbf{t}_{(0)}$; By default P_0=50(%)
4. Convert $\mathcal{X}$ to $\widetilde{\mathcal{X}}$ by subtracting each gene's average response within each treatment group
5. Compute $\widetilde{\mathbf{C}}$ = the (residual) sample correlation matrix of $\widetilde{\mathcal{X}}$
6. Find $(\widehat{\mathbf{u}}^*)^T = \left[\mathbf{0}_{c\times 1}^T \left(\widehat{\mathbf{u}}_{(1)}^*\right)^T\right]$, where $\widehat{\mathbf{u}}_{(1)}^* = \mathbf{t}_{(1)} - \widetilde{\mathbf{C}}_{(10)}\widetilde{\mathbf{C}}_{(00)}^{-1}\mathbf{t}_{(0)}$
7. Determine gene-ranking: Assign a gene with bigger $|\widehat{u}_i^*|$ a higher rank
8. Report top r genes as statistical discoveries

4.3.3 Algorithm for Multi-State and Continuous-State Studies

---

tEllipsoid: An enhanced gene-ranking for differential gene expression detection

Input: $\mathcal{X}$ = Labeled $m \times n$ gene expression matrix; r = Size of gene list

Output: The gene list containing top r differentially expressed genes

1. Calculate per gene *simple linear regression* t-statistics
2. Reindex genes such that $|t_1| \leq |t_2| \leq \cdots \leq |t_m|$
3. Gather first $c = \lceil 0.01 m P_0 \rceil$ t_i's in a vector $\mathbf{t}_{(0)}$; By default P_0=50(%)
4. Obtain the residual matrix $\widetilde{\mathcal{X}}$ whose entries are the differences between the observed expression values and the corresponding predicted expression values
5. Compute $\widetilde{\mathbf{C}}$ = the sample correlation matrix of $\widetilde{\mathcal{X}}$
6. Find $(\widehat{\mathbf{u}}^*)^T = \left[\mathbf{0}_{c\times 1}^T \left(\widehat{\mathbf{u}}_{(1)}^*\right)^T\right]$, where $\widehat{\mathbf{u}}_{(1)}^* = \mathbf{t}_{(1)} - \widetilde{\mathbf{C}}_{(10)}\widetilde{\mathbf{C}}_{(00)}^{-1}\mathbf{t}_{(0)}$
7. Determine gene-ranking: Assign a gene with bigger $|\widehat{u}_i^*|$ a higher rank
8. Report top r genes as statistical discoveries

4.4 Test Cases

To appreciate the increase in statistical power attributable to "exploiting" correlation, the performance of tEllipsoid is contrasted with three leading techniques. The first is the raw t-statistic itself and the other two are SAM [Significance Analysis

of Microarrays (Tusher et al., 2001)] and EDGE [Extraction and analysis of Differential Gene Expression (Leek et al., 2006)]. SAM adds a small exchangeability factor s_0 to the pooled sample variance when computing the two-sample t-statistic: $d_i = (\bar{x}_{i;2} - \bar{x}_{i;1})/(s_i + s_0)$; whereas EDGE is based on a general framework for sharing information across tests (see Storey et al. (2007)). EDGE is reported to show substantial improvement (in terms of statistical power) over five of the leading techniques including SAM (Storey et al., 2007). The other four are: (i) t/F–test of Kerr et al. (2000) and Dudoit et al. (2002); (ii) Shrunken t/F–test of Cui et al. (2005); (iii) The empirical Bayes local FDR of Efron et al. (2001); (iv) The *a posteriori* probability approach of Lonnstedt and Speed (2002). It should be noted that tEllipsoid can also serve as an additional layer to SAM and EDGE and enhance their power.

To determine the performance quality of various techniques, we focus primarily on the empirical FDR in the reported gene list: Empirical FDR = NoFP$/r$, where NoFP = the number of false positives. Broadly speaking, smaller the FDR better the technique.

4.4.1 Case I: Real Data With Induced Differences

Singh et al. (2002) studied $m = 12625$ genes on $n = 102$ oligonucleotide microarrays, comparing $n_1 = 50$ healthy males with $n_2 = 52$ prostate cancer patients. The purpose of their study was to identify genes that might anticipate the clinical behavior of Prostate cancer. We downloaded the .CEL files from `http://www-genome.wi.mit.edu/MPR/prostate`. The software RMAExpress (Irizarry et al., 2003) was used to obtain high quality gene expressions from these .CEL files. We let RMAExpress apply its in-built background adjustment, however, the quantile normalization was skipped. Each gene was represented in the final expression matrix $\mathcal{X}$ by the logarithm

(base 10) of its expression level. Taking the log is thought to increase normality and stabilize across group standard deviations (Tsai et al., 2003).

Algorithm testing required an expression matrix $\mathcal{X}$ with the knowledge of truly non-differential genes. At the same time, we wanted the inter-gene correlation in $\mathcal{X}$ to resemble that in the real microarray data. These two seemingly conflicting requirements were satisfied concurrently by row standardizing a real $\mathcal{X}$. The prostate cancer matrix $\mathcal{X}$ was transformed to $\underline{\widetilde{\mathcal{X}}}$ by subtracting each gene's average response within each treatment group, and by normalizing within group sample mean squares. That is, for each group $k \in \{1, 2\}$, $(1/n_k) \sum_j \underline{\widetilde{\mathcal{X}}}_{ij} = 0$ and $(1/n_k) \sum_j \underline{\widetilde{\mathcal{X}}}^2_{ij} = 1$. Here, the sum runs over corresponding n_k samples only. With this transformation, all genes have equal energy and yet the same within group inter-gene correlation structure as the original X.**Note.** Normalizing within group sample mean squares to unity is not implemented in the tEllipsoid algorithm.

To generate a test data set from $\underline{\widetilde{\mathcal{X}}}$, its 102 columns were randomly divided into groups of 50 ($=n_1$) and 52 ($=n_2$). Next m_u (m_d) genes were randomly chosen for up (down) regulation by adding a positive (negative) offset x_u (x_d) to the corresponding entries in group 2. Various choices of (m_u, m_d, x_u, x_d) were tested to represent a range of differential analysis scenarios encountered in practice.

Two cases were studied. In the first, the proportion of truly differential genes, say p_1, was taken to be relatively small: $p_1 \sim 0.01$–0.05. The second case employed a larger $p_1 \sim 0.1$. The former simulates microarray studies seeking genes that distinguish subtypes of cancer, diabetes, etc., whereas the latter resembles studies comparing healthy versus diseased cell activity.

Results were obtained using the subroutines `samr.r` from the package "samr" and `statex.r` from the package "edge." Both routines compute their native gene summary

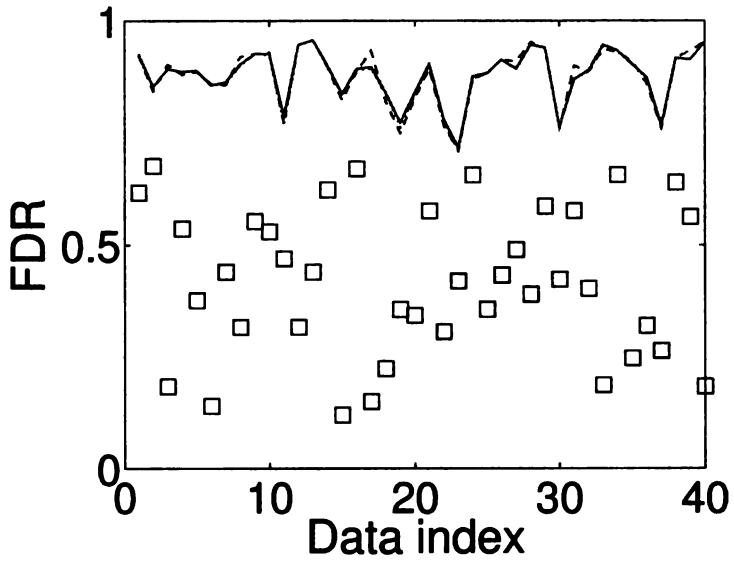
statistics which, in turn, can be used to determine top r genes.

Case 1 [$p_1 \approx 0.025$, m_u=200, m_d=100, x_u=0.1, and x_d=-0.1]. Figure 4.1(a) shows plots of the FDRs for 40 different data sets with the size of the reported list, r=300. A large value of r coincides with an attempt to extract as many differential expressions as possible, a desired goal especially in microarray studies performed to identify genes that are to be explored further – experimentally or computationally – to gain better understanding of underlying gene networks. Since the differential signal x_u=0.1 and x_d=-0.1 is rather weak, recovering a good list is not easy as evident from the results – among all methods only tEllipsoid achieved sufficiently low FDRs to rescue a few $\mathcal{X}$'s.

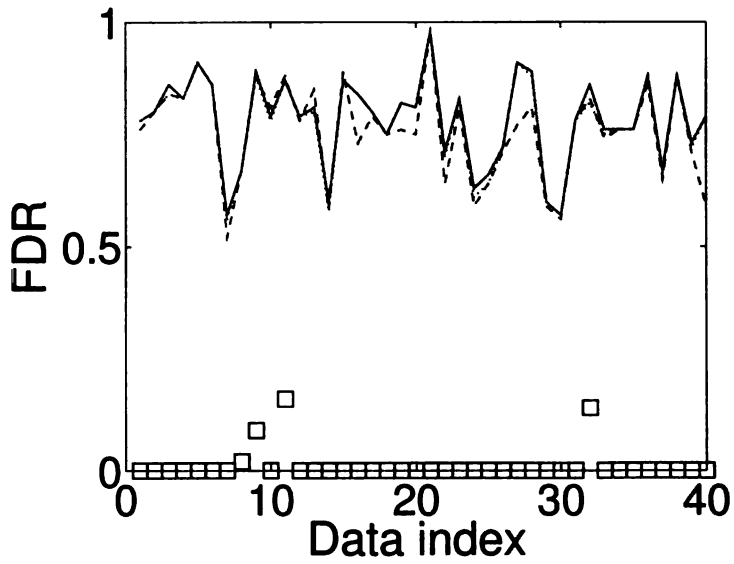
Figure 4.1(b) presents results for r=100. A smaller r would be chosen to identify high-quality class distinguishing features for gene-expression-profiling-based clinical diagnosis and prognosis, where the goal is to build accurate classifiers and predictors. Whereas Singh et al. (2002) build a classifier around only 16 of 12625 features, they do discuss the need to include as many reliable features as possible. Remarkably, for 37 out of 40 $\mathcal{X}$ matrices, tEllipsoid reports gene lists with no false discoveries at all, while the other techniques fail to obtain a single gene list with the FDR < 0.5.

Case 2a [$p_1 \approx 0.1$, m_u=600, m_d=600, x_u=0.02, and x_d=-0.02]. In this set of experiments, p_1 is increased, but the differential signal is reduced. This situation also proves to be challenging for the existing techniques. However, tEllipsoid provides the FDR of ~ 0.5 for r=1200, and, again for r=300, while it reports most gene lists with no false discoveries at all.

Case 2b [$p_1 \approx$0.1, m_u=600, m_d=600, x_u=0.1, and x_d=-0.1]. This subcase is designed to assess the effects of small sample sizes on performance. n_1 and n_2 are both reduced to 20. We randomly chose 20 columns per group from the original prostate

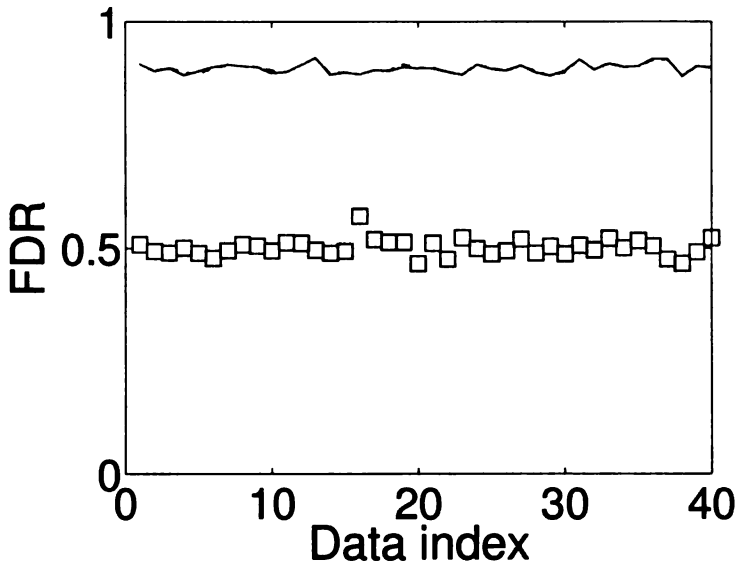


(a)

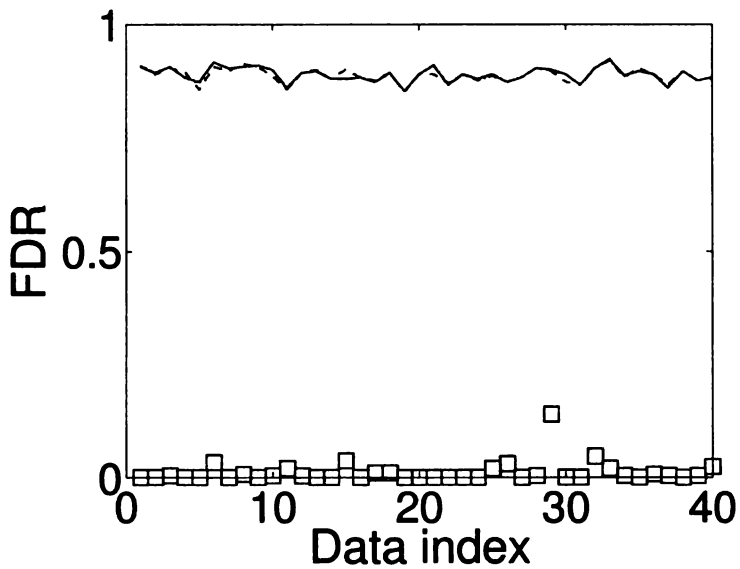


(b)

Figure 4.1: FDRs for Case 1. The number of truly differential gene is 300. Panel (a) r=300; Panel (b) r=100. "Exploiting correlation" considerably enhances the statistical power. Square (□) marker = tEllipsoid. Lines: solid = raw t-statistic; dotted = SAM; dashed = EDGE.



(a)



(b)

Figure 4.2: FDRs for Case 2a. The number of truly differential gene is 1200. Panel (a) r=1200; Panel (b) r=300. "Exploiting correlation" appreciably enhances the statistical power. Square (□) marker = tEllipsoid. Lines: solid = raw t-statistic; dotted = SAM; dashed = EDGE.

cancer $\mathcal{X}$, and then applied the data generation process (including row standardization) detailed in Subsection 4.4.1. Reduction in the number of samples is compensated by increase in the differential signal. The FDRs for tEllipsoid, Fig. 4.3, are excellent suggesting that tEllipsoid increases power of small sample data sets too.

4.4.2 Case II: Simulated Data

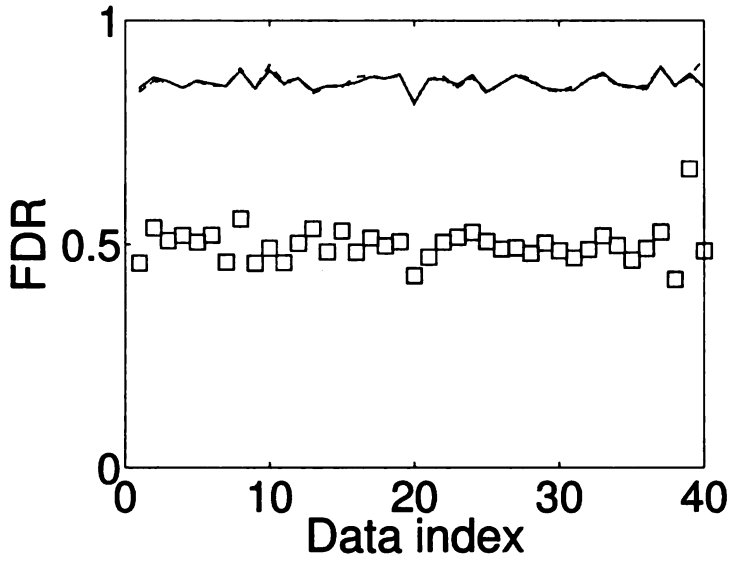
Before devising the prostate data test setup, tEllipsoid was tested on several simulated data sets. Below we discuss some simulation results that shed further light on the small sample behavior.

Let us denote by $\mathcal{X}_{(i)}$ the i^{th} column of a simulated expression matrix $\mathcal{X}$. We assume that the random vector $\mathcal{X}_{(i)}$ is multivariate Gaussian with mean $\mathbf{0}$ and covariance matrix $\mathbf{W}$. Each such column represents m=3226 genes with a covariance matrix $\mathbf{W}$ that introduces roughly the same amount of correlation as found in the BRCA data of Hedenfalk et al. (2001). We choose $m_u = 50, m_d = 50, x_u = 1, x_d = -1, n_1 = 10$, and $n_2 = 10$. Figure 4.4 shows plots of the FDRs for r=50 and r=100. Table 4.1 shows results for some data point from Fig. 4.4(b). Shown are the top 100 values of $\hat{u}_i^*$ and each corresponding original t_i with concomitant rank. With smaller n, preeminence of tEllipsoid with respect to existing techniques scales down a bit. Nevertheless, for r=50 case, for 25 out of 40 simulated $\mathcal{X}$ realizations, tEllipsoid achieves a low FDR of ~ 0.1 or less.

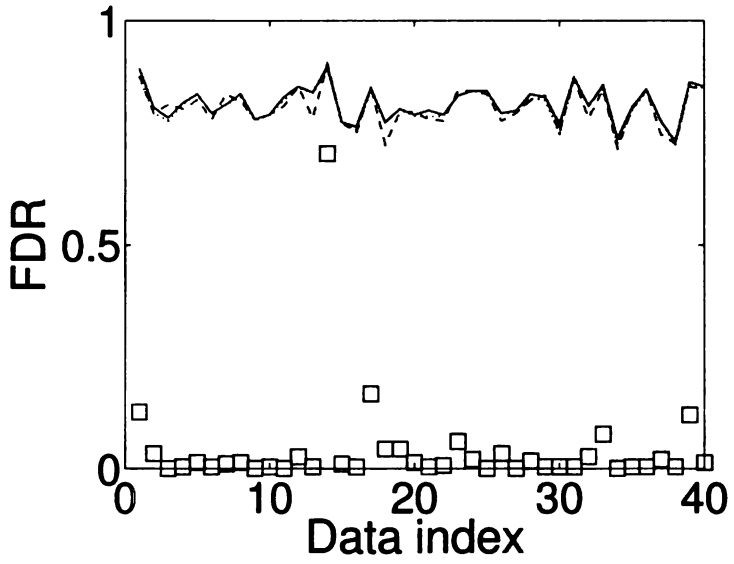
Interestingly, with a smaller n, SAM outperforms the other two techniques. This is not entirely surprising as a smaller n can make the noise in the per gene pooled variance s_i (and possibly the equivalent quantity in the EDGE algorithm) more prominent. Nevertheless, SAM does mitigate this issue in some measure by using the exchangeability factor s_0 to adjust the effective pooled variance (Tusher et al., 2001).

1–50				51–100			
$\hat{u}_i^*$ rank	$\hat{u}_i^*$	t_i	t_i rank	$\hat{u}_i^*$ rank	$\hat{u}_i^*$	t_i	t_i rank
1	4.22	5.87	1	**51**	**2.18**	**3.45**	**23**
2	-4.17	-5.55	2	**52**	**2.17**	**3.05**	**42**
3	-3.93	-4.26	5	53	2.16	2.80	82
4	-3.74	-4.12	7	54	-2.15	-2.57	122
5	-3.58	-4.49	4	55	2.15	1.96	357
6	-3.49	-3.34	28	56	-2.14	-1.47	751
7	-3.45	-4.25	6	57	2.13	2.25	229
8	3.35	3.87	10	58	2.13	1.77	486
9	-3.33	-3.20	35	59	2.13	1.44	785
10	3.33	3.77	13	60	2.12	2.14	273
11	3.25	3.42	25	61	-2.11	-1.48	744
12	-3.16	-2.18	260	**62**	**2.10**	**1.80**	**453**
13	3.14	4.54	3	63	2.09	2.60	114
14	3.10	2.87	65	64	-2.09	-2.05	312
15	-3.08	-3.54	17	65	2.09	2.70	96
16	-3.07	-2.80	80	66	2.09	2.23	237
17	3.06	3.49	20	67	2.08	2.34	188
18	3.02	2.29	213	68	-2.08	-2.24	232
19	-2.99	-3.34	27	69	-2.06	-2.53	130
20	-2.93	-3.13	38	70	-2.04	-2.11	283
21	-2.92	-2.92	57	**71**	**-2.04**	**-2.95**	**54**
22	2.86	3.26	31	**72**	**-2.03**	**-3.08**	**40**
23	-2.83	-2.82	74	73	-2.02	-2.30	210
24	2.82	2.37	180	**74**	**-2.01**	**-3.67**	**15**
25	-2.81	-2.13	276	75	2.00	2.62	109
26	**2.81**	**3.48**	**21**	76	-1.98	-2.38	171
27	-2.79	-3.01	47	77	1.98	1.43	795
28	2.70	2.87	64	78	1.96	1.69	549
29	-2.66	-3.15	37	79	-1.95	-1.47	746
30	**-2.58**	**-3.85**	**11**	80	1.95	1.95	361
31	-2.56	-2.84	71	**81**	**1.95**	**2.81**	**77**
32	-2.55	-1.72	524	82	-1.94	-1.41	813
33	-2.54	-2.63	106	**83**	**1.94**	**3.40**	**26**
34	-2.54	-2.69	98	84	1.94	1.30	948
35	2.53	2.30	209	**85**	**-1.94**	**-3.27**	**30**
36	2.48	2.45	148	86	-1.93	-1.11	1190
37	-2.47	-2.29	212	87	-1.93	-1.37	872
38	-2.46	-3.21	33	**88**	**-1.93**	**-3.44**	**24**
39	-2.43	-2.44	154	**89**	**-1.92**	**-3.07**	**41**
40	2.43	2.71	94	90	-1.92	-1.50	726
41	2.40	2.86	66	**91**	**1.90**	**3.62**	**16**
42	-2.34	-2.60	115	**92**	**-1.90**	**-2.82**	**75**
43	-2.34	-2.98	50	93	1.89	1.25	1007
44	**-2.33**	**-3.80**	**12**	**94**	**1.87**	**3.89**	**8**
45	-2.32	-2.06	306	**95**	**-1.86**	**-3.49**	**19**
46	2.29	1.81	444	**96**	**-1.86**	**-2.08**	**300**
47	-2.27	-1.17	1110	97	1.85	1.20	1074
48	2.26	1.97	347	**98**	**-1.83**	**-2.90**	**60**
49	**2.24**	**3.75**	**14**	99	1.83	1.39	833
50	**2.20**	**3.88**	**9**	100	-1.82	-1.95	367

Table 4.1: TEllipsoid in action with Top 100 $\hat{u}_i^*$'s. Corresponding t_i's and their rank are also shown. TEllipsoid = 22 NoFPs; raw t-statistics = 68 NoFPs. Truly null genes are printed in bold-sans typeface.

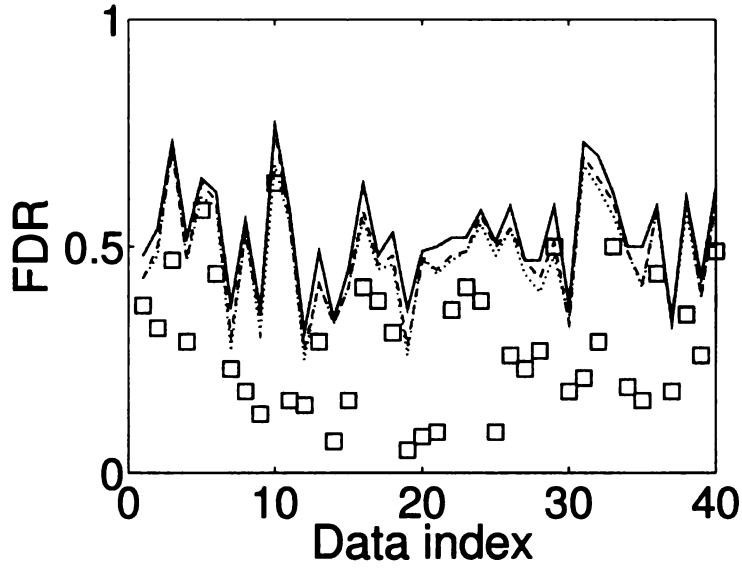


(a)

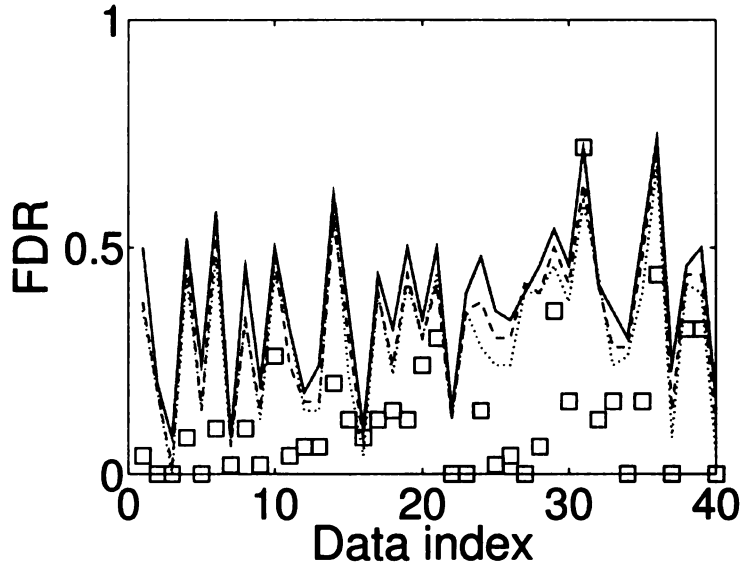


(b)

Figure 4.3: FDRs for Case 2b. Panel (a) r=1200; Panel (b) r=300. The sample size is smaller than that in Cases 1 & 2a and yet "Exploiting correlation" has apparent benefits. Square (□) marker = tEllipsoid. Lines: solid = raw t-statistic; dotted = SAM; dashed = EDGE.



(a)



(b)

Figure 4.4: FDRs for simulated data. Panel (a) r=100; Panel (b) r=50. (Small) sample sizes: n_1=10, n_2=10. Yet, "Exploiting correlation" considerably enhances the statistical power. Square (□) marker = tEllipsoid. Lines: solid = raw t-statistic; dotted = SAM; dashed = EDGE.

4.5 Discussion

By allowing researchers to examine the simultaneous expressions of enormous numbers of genes, microarrays promised to revolutionize the understanding of complex diseases and usher in an era of personalized medicine. However, the shift in perception of that promise is palpable in the literature. A 1999 *Nature Genetics* article (Lander, 1999) is entitled "Array of hope," but a 2005 *Nature Reviews* article (Frantz, 2005) is entitled "An array of problems." It is not unusual for impacts of new technologies to be overestimated when first deployed, then to have the expectations moderated as the technologies reveal new complexities in the problems they are designed to solve. In the study of microarray data, the need for exceeding care in the design and regularization of experiments and data collection are understood to be critical, but the biggest hindrance to progress has been the data interpretation. In particular, the biggest challenge seems to be the treatment of *intrinsic* inter-gene correlation.

In most microarray data there are at least three vital resources: (i) identifiability (ii) immense parallel structure, and (iii) inter-gene correlation itself. In this light, tEllipsoid can be viewed as exploiting more than correlation as a means of sharing information across tests, as it also involves identifiability.

A crucial step in formulating tEllipsoid was the comprehension of the effects of inter-gene correlation on $\mathrm{Cov}(T_i, T_{i'})$. In light of Observations 1 and 2, the choice of the Mahalanobis distance was intuitive, as it is already known to give computationally attractive solutions through the matrix inversion lemma.

Limited time and resources—and perhaps also the necessity for scientific focus—often require biomedical researchers to work on only a small number of "hot (gene) prospects." Even under such highly conservative conditions, however, misleading results can occur, as evident in the results of Figs. 4.1–4.4. For all their careful

development and statistical power, even state-of-the-art tools that do not account for correlation can report spurious gene lists. The extra statistical power available by exploiting inter-gene correlation promises to further guard against anomalous results that can have serious consequences for the trajectory of a study of gene function, causation, and interaction.

In summary, this chapter has reported the development and testing of a novel framework for the detection of differential gene expression. The framework combines the exploitation of inter-gene correlation to share information across tests, with identifiability – the fact that in most microarray data sets, a large proportion of genes can be identified *a priori* as non-differential. When applied to the widely used two-sample t-statistic approach, this viewpoint yielded an elegant differential analysis technique, which requires as inputs only a gene expression matrix, related two-sample labels, and the size of desired gene-list r. Empirical evidence suggests that exploiting correlation substantially enhances statistical power. Usually, with increase in microarray samples, power tends to increase considerably, but, even for small sample sizes, the performance improvement is noticeable.

Chapter 5

Concluding Remarks

5.1 Summary and Concluding Remarks

This work has investigated the effects of inter-gene dependency on statistical methods for differential gene analysis. Differentially expressed genes are pivotal in understanding highly intricate genotype–phenotype relations and devising appropriate therapeutic interventions. The effort focused on understanding, correcting, and exploiting the effects of second-order inter-gene dependency within the large-scale significance testing formulation for differential gene analysis. It was shown that combining inter-gene dependency with gene expression purity yields more accurate and powerful statistical inferences yielding a more accurate list of differentially expressed genes. The main statistical theme of this work has been to draw second-order conditional inferences based on cases that are theoretically more likely to be null.

This research resulted in novel ways of combining inter-gene correlation with purity. A method for mitigating unfavorable effects of correlation on the false discovery rate calculations was developed. A framework for exploiting beneficial effects of cor-

relation on gene-ranking enhancement was also discovered. The findings are very general in that they are applicable to any test statistics obtained using *simple linear regression.*

In particular, a novel approach to the false discovery calculations was explored. The approach has produced an inferential technique capable of handling exceedingly small sample sizes and reporting an entire distribution of a random variable model of the number of false discoveries. The technique first summarizes the effect of millions of pair-wise correlation coefficients in a single parameter, then explicitly incorporates this parameter in the inference of the number of false discoveries.

The possibility of exploiting correlation to improve gene-ranking was also explored. This effort culminated in a powerful framework employing statistical distance measures which can account for the effect of correlation on the joint distribution of test statistics. The extra statistical power made available by exploiting inter-gene correlation promises to further guard against anomalous results that can have serious consequences for the trajectory of a study of gene function, causation, and interaction.

The urgency for more accurate differential gene analysis methods motivated this research. However, the principal contributions are firmly rooted in the fundamental theory and methods of large-scale significance testing and enjoy broader applicability. Large-scale significance testing has become a key statistical tool for exploratory data analysis in single nucleotide polymorphism detection and other high-throughput genomic research endeavors. A proper treatment of dependency in most large-scale significance testing applications seems necessary to draw meaningful conclusions.

5.2 Future Work

Research topics for future work concerning the FDR estimation method of Chapter 3 are listed below:

1. Further investigation of $\mathrm{Cor}(V, C)$ across a range of microarray data sets.

2. Evaluation of the effect of center-area boundary δ_0 on $\mathrm{Cor}(V, C)$ and on overall accuracy.

3. Extension of the single parameter correlation model to a multi-parameter correlation model to increase modeling accuracy.

4. Exploration of connections between the *maxent* exponential density and the exponential parametric form of "empirical null density" as in Efron (2004).

Research topics for future work concerning the gene-ranking framework of Chapter 4 are listed below:

1. Investigation of the role of P_0 in overall performance.

2. Empirical and / or theoretical understanding of the relation between the residual correlation and the gain in statistical power; the eigenvalue spread of residual correlation matrix seems crucial here.

3. Empirical and / or theoretical understanding of the relation between the number of samples and the gain in statistical power.

4. Developing false discovery rate estimates for the proposed gene-ranking.

5. Incorporation of *prior* information, such as "gene grouping" or *a priori* non-null identity.

Bibliography

B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and JD Watson. *Molecular biology of the cell.* 2002.

A. Almudevar, L.B. Klebanov, X. Qiu, P. Salzman, and A.Y. Yakovlev. Utility of correlation measures in analysis of gene expression. *NeuroRX*, 3(3):384–395, 2006.

P. Baldi and S. Brunak. *Bioinformatics: the machine learning approach.* MIT Press Cambridge, MA, USA, 2001.

J. Barnard, R. McCulloch, and X.L. Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1311, 2000.

Y. Benjamini. Comment: Microarrays, empirical Bayes and the two-groups model. *Statist. Sci*, 23(1):23–28, 2008.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B.*, 57(1):289–300, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

J.O. Berger. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1):1–32, 2003.

N. Biotechnology. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24:1151–1161, 2006.

BM Bolstad, RA Irizarry, M. Astrand, and TP Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

T.T. Cai. Comment: Microarrays, empirical Bayes and the two-group model. *Statistical Science*, 23(1):29–33, 2008.

K.R. Coombes, J. Wang, and K.A. Baggerly. The tail-rank statistic for finding biomarkers from microarray data, with application to prostate cancer. *preprint at* ***http://bioinformatics.mdanderson.org/TailRank***, 2008.

T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley New York, 1991.

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(1):59–75, 2005.

PA Dejviver and J. Kittler. *Pattern recognition: a statistical approach*. Prentice Hall International, 1982.

J.R. Deller, Jr., J.H.L. Hansen, and J.G. Proakis. *Discrete-time processing of speech signals*. IEEE Press, 2000.

S. Dudoit, Y.H. Yang, M.J. Callow, and T.P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1):111–139, 2002.

B. Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–105, 2004.

B. Efron. Microarrays, empirical Bayes, and the two-groups model. *Preprint, Dept. of Statistics, Stanford University*, 2006.

B. Efron. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477):93–103, 2007a.

B. Efron. Size, power and false discovery rates. *Annals of Statistics*, 35:1351–1377, 2007b.

B. Efron. Microarrays, empirical Bayes and the two-group model. *Statistical Science*, 23(1):1–22, 2008.

B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456): 1151–1161, 2001.

M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863, 1998.

S. Frantz. An array of problems. *Nature Reviews Drug Discovery*, 4:362–363, 2005.

C. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *Annals of Statistics*, 32(3):1035–1061, 2004.

G.H. Golub and C.F. Van Loan. *Matrix computations*. Johns Hopkins University Press, 1996.

G.W. Hatfield, S. Hung, and P. Baldi. Differential analysis of DNA microarray gene expression data. *Molecular Microbiology*, 47(4):871–877, 2003.

I. Hedenfalk, D. Duggan, Y. Chen, et al. Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344(8):539–548, 2001.

N.J. Higham. Computing the nearest correlation matrix–a problem from finance. *IMA Journal of Numerical Analysis*, 22(3):329–343, 2002.

L. Hunter. Molecular biology for computer scientists. *Artificial Intelligence and Molecular Biology, Cambridge*, 1993.

JP Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8): e124, 2005.

R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, and T.P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4): e15, 2003.

ET Jaynes. *Probability theory: The logic of science*. Cambridge University Press, 2003.

P.A. Jones and S.B. Baylin. The fundamental role of epigenetic events in cancer. *Nat Rev Genet*, 3(6):415–428, 2002.

M.K. Kerr, M. Martin, and G.A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7(6):819–837, 2000.

Kyung In Kim and Mark van de Wiel. Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*, 9(1):114, 2008.

L. Klebanov and A. Yakovlev. Diverse correlation structures in microarray gene expression data and their utility in improving statistical inference. *Annals of Applied Statistics*, 1(2):538–559, 2007.

L. Klebanov, C. Jordan, and A. Yakovlev. A new type of stochastic dependence revealed in gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 5(1):7, 2006.

E.S. Lander. Array of hope. *Nature Genetics*, 21(1):3–4, 1999.

M. Langaas, E. Ferkingstad, and B.H. Lindqvist. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society, Series B.*, 67:555–72, 2005.

M.L.T. Lee, F.C. Kuo, GA Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences*, 97(18): 9834–9839, 2000.

J.T. Leek and J.D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

J.T. Leek, E. Monsen, A.R. Dabney, and J.D. Storey. EDGE: Extraction and analysis of differential gene expression. *Bioinformatics*, 22(4):507–508, 2006.

EL Lehmann and J.P. Romano. Generalizations of the familywise error rate. *Annals of Statistics*, 33(3):1138–1154, 2005.

E.L. Lehmann and J.P. Romano. *Testing Statistical Hypotheses*. Springer, 2006.

F. Li and G.D. Stormo. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, 17(11):1067–1076, 2001.

J.C. Liechty, M.W. Liechty, and P. Muller. Bayesian correlation estimation. *Biometrika*, 91(1):1–14, 2004.

D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Norton, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

I. Lonnstedt and T. Speed. Replicated microarray data. *Statistica Sinica*, 12(1): 31–46, 2002.

P.C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Science in India*, 2(1):49–55, 1936.

G. Marsaglia and I. Olkin. Generating correlation matrices. *SIAM Journal on Scientific and Statistical Computing*, 5:470–475, 1984.

C.N. Morris. Comment: Microarrays, empirical Bayes and the two-group model. *Statistical Science*, 23(1):34–40, 2008.

MA Newton, CM Kendziorski, CS Richmond, F.R. Blattner, and KW Tsui. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8(1):37–52, 2001.

I. Olkin. Note on 'The Jacobians of certain matrix transformations useful in multivariate analysis'. *Biometrika*, 40(1-2):43, 1953.

A.B. Owen. Variance of the number of false discoveries. *Journal of the Royal Statistical Society, Series B.*, 67:411–426, 2005.

T.A. Patterson, E.K. Lobenhofer, S.B. Fulmer-Smentek, et al. Performance comparison of one-color and two-color platforms within the Microarray Quality Control (MAQC) project. *Nature Biotechnology*, 24:1140–1150, 2006.

Y. Pawitan, K.R.K. Murthy, S. Michiels, A. Ploner, and O. Journals. Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics*, 21(20): 3865–3872, 2005.

E.F. Petricoin, J.L. Hackett, L.J. Lesko, et al. Medical applications of microarray technologies: a regulatory science perspective. *Nature Genetics*, 32(supp):474–479, 2002.

Houduo Qi and Defeng Sun. A quadratically convergent newton method for computing the nearest correlation matrix. *SIAM Journal on Matrix Analysis and Applications*, 28(2):360–385, 2006.

X. Qiu and A. Yakovlev. Some comments on instability of false discovery rate estimation. *Journal of Bioinformatics and Computational Biology*, 4(5):1057–1068, 2006.

X. Qiu, A.I. Brooks, L. Klebanov, and A. Yakovlev. The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics*, 6:120, 2005a.

X. Qiu, L. Klebanov, and A. Yakovlev. Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology*, 4(1):1157, 2005b.

X. Qiu, Y. Xiao, A. Gordon, and A. Yakovlev. Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics*, 7:50, 2006.

J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32(supp):496–501, 2002.

A. Reiner-Benaim. FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis. *Biometrical journal*, 49(1): 107–126, 2007.

K. Rice and D. Spiegelhalter. Comment: Microarrays, empirical Bayes and the two-groups model. *Statistcal Science*, 23(1):41–44, 2008.

M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235): 467–470, 1995.

D. Singh, P.G. Febbo, K. Ross, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

J.D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B.*, 64(3):479–498, 2002a.

J.D. Storey. *False discovery rates theory and applications to DNA microarrays*. PhD thesis, stanford university, 2002b.

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. *Journal of the Royal Statistical Society Series B*, 66(1):187–205, 2004.

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(2):414–432, 2007.

A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

R. Tibshirani and L. Wasserman. Correlation-sharing for detection of differential gene expression. *Arxiv preprint math.ST/0608061*, 2006.

C.A. Tsai, Y.J. Chen, J.J. Chen, and O. Journals. Testing for differentially expressed genes with microarray data. *Nucleic Acids Research*, 31(9):e52, 2003.

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

M.J. van der Laan, S. Dudoit, and K.S. Pollard. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1):15, 2004.

A.B. van't Wout, G.K. Lehrman, S.A. Mikheeva, et al. Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4+-T-cell lines. *Journal of Virology*, 77(2):1392–1402, 2003.

JD Watson, TA Baker, SP Bell, A. Gann, and R. Losick. *Molecular biology of the gene (5th Edition)*. Benjamin Cummings, 2004.

Y. Woo, J. Affourtit, S. Daigle, et al. A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. *Journal of Biomolecular Techniques*, 15(4):276–284, 2004.

Y. Zheng and M. Pepe. A practical multifaceted approach to selecting differentially expressed genes. *Cancer Informatics*, 2:113–122, 2007.

K.H. Zou and W.J. Hall. On estimating a transformation correlation coefficient. *Journal of Applied Statistics*, 29:745–760, 2002.