International Scholarly Research Network
ISRN Bioinformatics
Volume 2012, Article ID 564715, 18 pages
doi:10.5402/2012/564715

Research Article

Nonlinear Dependence in the Discovery of Differentially Expressed Genes

J. R. Deller Jr.,[1] Hayder Radha,[1] J. Justin McCormick,[2] and Huiyan Wang[3]

[1] *Department of Electrical and Computer Engineering, Michigan State University, 2120 EB, East Lansing, MI 48824, USA*
[2] *Carcinogenesis Laboratory, Department of Molecular Biology and Biochemistry, Michigan State University, 341 FST, East Lansing, MI 48824, USA*
[3] *College of Computer Science and Information Engineering, Zhejiang Gongshang University, 18 Xuezheng Street, Zhejiang Province Hangzhou, 310018, China*

Correspondence should be addressed to J. R. Deller Jr., deller@egr.msu.edu

Received 16 September 2011; Accepted 9 November 2011

Academic Editors: T. Can and S. Panni

Microarray data are used to determine which genes are active in response to a changing cell environment. Genes are "discovered" when they are significantly differentially expressed in the microarray data collected under the differing conditions. In one prevalent approach, all genes are assumed to satisfy a null hypothesis, $\mathbb{H}_0$, of no difference in expression. A *false discovery* (type 1 error) occurs when $\mathbb{H}_0$ is incorrectly rejected. The quality of a detection algorithm is assessed by estimating its *number of false discoveries*, $\mathfrak{F}$. Work involving the second-moment modeling of the z-value histogram (representing gene expression differentials) has shown significantly deleterious effects of intergene expression correlation on the estimate of $\mathfrak{F}$. This paper suggests that nonlinear dependencies could likewise be important. With an applied emphasis, this paper extends the "moment framework" by including third-moment skewness corrections in an estimator of $\mathfrak{F}$. This estimator combines observed correlation (corrected for sampling fluctuations) with the information from easily identifiable null cases. Nonlinear-dependence modeling reduces the estimation error relative to that of linear estimation. Third-moment calculations involve empirical densities of 3×3 covariance matrices estimated using very few samples. The principle of entropy maximization is employed to connect estimated moments to $\mathfrak{F}$ inference. Model results are tested with BRCA and HIV data sets and with carefully constructed simulations.

1. Introduction

This work is motivated by analytical challenges that arise in the use of microarray data to discover genes that are differentially expressed across experimental conditions such as control and treatment. Although the discussion centers around this genomics task, the developed methods are quite general and should be useful in other multiple-testing applications in which there is substantial dependence among test measures, and in which a small sample size may cause significant fluctuations in statistics employed in the testing. The specific aim of this work is to develop a reliable estimator of the *number of false discoveries* (type I errors—denoted $\mathfrak{F}$) in a multiple-testing problem in such an adverse setting.

The classic and contemporary literature in cell biology, and the more recent literature in genomics (both in print and posted on the Internet), is replete with tutorial information at all levels about cell anatomy and physiology, and genomics, as well as the microarray technology deployed in the present application. A good entry point for accessing information about contemporary developments in the genomics field is the web site of the US National Genome Research Institute [1]. The papers by Page et al. [2] and Wang [3] provide relatively current reviews of microarray technology and methods. A brief description of the biological aspects of the genomics application underlying this work is found in Appendix A of this paper.

In the gene-discovery application, each gene is tested against a null hypothesis, $\mathbb{H}_0$, that the gene is *not* differentially expressed across experimental conditions. All genes are initially assumed to satisfy $\mathbb{H}_0$ in this analysis, and $\mathfrak{F}$ is estimated conservatively. This "all-null" presumption is consistent with this application in which $\mathbb{H}_0$ is true for a vast majority of the genes in any experiment. Beyond the gene

detection problem, however, this presumption is realistic in many practical applications of large-scale testing in which the prior probability of null cases, say π_0, is large, and in which the goal is to identify a small set of interesting "nonnull" cases [4]. With $\pi_0 \approx 1$, it is also possible to impose *identifiability* (the strongly justified assumption that a given gene satisfies $\mathbb{H}_0$) on some of the genes, yielding crucial information with which to condition the estimation of $\mathfrak{F}$ [5].

One of the earliest reports of research using the microarray (or "gene chip") appeared in a paper by Schena et al. in *Science* in 1995 [6]. Generally speaking, research that employs the microarray to analyze gene expression data has one (or both) of the following underlying aims: the discovery of gene *coexpression*, or the discovery of gene *differential expression*. To the extent that these problems have been investigated separately, the coexpression problem has frequently been addressed by clustering methods (e.g., [7–13]), whereas differential expression has been studied using variations of classical statistical hypothesis testing (e.g., [5, 14]). Whereas differential-expression/hypothesis-testing research was, and is, concerned with expression in response to differing cell conditions (normal versus pathology, medical treatment versus control, etc.), the early coexpression/clustering research was often focused on phenotypic manifestations of the gene expression.

As the technology has matured, the dichotomy suggested above has blurred with many current applications of the microarray involving "hybrid" research questions into both differential and coexpression. Application areas include discovery and exploration of gene regulatory systems, tissue and tumor classification, biomarker prediction, discovery and reverse engineering of gene expression networks—not to mention the microarray's deployment in the study of protein synthesis, metabolism, evolution, and other areas related to cell biology. Technical approaches to these problems have gone well beyond classical clustering and hypothesis-testing methods. Indeed, in the past few years, *statistical* (i.e., "nonclustering") methods to address the coexpression problem have been reported (e.g., [15, 16]), while the hybrid of the two problems—that of detecting and analyzing *differentially coexpressed* genes—has been researched using an ever-increasing number of methods including clustering with complex and dynamic feature selection methods, image transformation and processing of expression data, biclustering, graph and network theory, hypothesis testing, and other statistical approaches (e.g., [17–28]). In this paper, we return to the focused problem of detecting differential expression across treatment conditions, but it will become clear—as it has to the community working on this problem—that differential expression cannot be studied independently of coexpression.

Early work on classical statistical techniques for microarray-based gene discovery is summarized in the 2002 review paper by Pan [29]. Initially, it was customary to treat gene expression outcomes as realizations of independent random variables (RVs). More recent papers, however—notably, those of Owen [30], Efron [31], and Pawitan et al. [32]—caution researchers of the detrimental effects of correlated gene-expressions on the validity of "discovered" genes. In particular, it was reported that highly correlated tests increase the variance of $\mathfrak{F}$ (or, its normalized counterpart, the *false discovery rate*, $\dot{\mathfrak{F}} \stackrel{\text{def}}{=} \mathfrak{F}/G_*$, where G_* is the number of "discovered" genes), thus making estimates of $\mathfrak{F}$ less reliable. In particular, high variance renders the average, $\widehat{\mu}_{\mathfrak{F}} \approx \mu_{\mathfrak{F}} = \mathcal{E}\{\mathfrak{F}\}$, an unreliable estimator of $\mathfrak{F}$ [30]. Among many causes, intergene correlation is attributable to coexpressed genes [4] and to unmodeled factors that introduce systematic effects across genes [33, 34]. As a result, for most real data, the assumption of independence or weak dependence among gene expressions is unfounded, and methods treating correlation are necessary [35, 36].

Accordingly, there has been significant recent interest in improving statistical gene detection methods in light of this detrimental correlation. For example, Storey et al. [37] present an approach to the notion of sharing information across t scores, which they describe as "borrowing strength across the tests" for a potential increase in statistical power. Tibshirani and Wasserman [38] discuss a quantity called the "correlation-shared" t-statistic and derives theoretical bounds on its performance. Hu et al. [39] examine the covariance structure of the expression data and discover benefits of linking coexpression and differential expression in a distance measure—thus, moving toward the "hybrid" problem described above.

Recent research into the hybrid differential coexpression problem has also yielded results and methods that could ultimately benefit the differential expression problem. Because the differential coexpression research is often concerned with differing phenotypes, rather than with different treatment conditions, two given research efforts involving differential coexpression might seek answers to different sets of genetic questions through expression data. Like the "dual conditions researchers," however, the "phenotype" researchers have encountered their own forms of confounding dependencies, notably the relative gene locations, the expression time sequencing, and phase information (e.g., [40–42]). Papers have been published addressing these issues, including the exposition of new statistical approaches—for example, "CorScor" developed by Dettling et al. [21], the "ECF-statistic" of Lai et al. [22], "the gene-set coexpression analysis" of Choi and Kendziorski [15]—as well as new clustering methods—for example, the web-based expression analyzer of Xiang et al. [43], high-order preclustering method of Wong et al. [44], and the "BioSym" distance measure of Bandyopadhyay and Bhattacharyya [45]. A recent review of clustering methods in genomics appears in the paper by Dalton et al. [46]. A more general examination of the performance of classifiers of microarray expressions appears in the paper by Ancona et al. [47].

The present paper returns to the problem of gene discovery by statistical hypothesis testing, but with the concern for the effects of nonlinear dependencies on (the estimation of) the number of false results. Empirical work below provides cogent evidence that accounting for intergene correlation alone does not sufficiently mitigate the adverse effects of dependency. Recent work by Hu et al. [48] has shown the importance of accounting for nonlinear dependence

Table 1: Notation used for elementary scalar quantities. v, v_1, and v_2 are RVs and $\mathcal{E}$ denotes the expectation.

Mean, average	μ_x or $\mu(x) = \mathcal{E}\{x\}$	Standard deviation	$\sigma_v \stackrel{\text{def}}{=} \sqrt{\mathcal{E}\{(v-\mu_v)^2\}} = \sqrt{\varphi(v,v)}$
Covariance	$\varphi(v_1, v_2) \stackrel{\text{def}}{=} \mathcal{E}\{(v_1 - \mu_{v_1})(v_2 - \mu_{v_2})\}$	Variance	$\sigma_v^2 \stackrel{\text{def}}{=} \mathcal{E}\{(v-\mu_v)^2\} = \varphi(v,v)$
Correlation	$\mathcal{E}\{v_1, v_2\} \stackrel{\text{def}}{=} \varphi(v_1, v_2) + \mu_{v_1}\mu_{v_2}$ (no special symbol reserved)	Correlation coefficient	$\rho(v_1, v_2) \stackrel{\text{def}}{=} \dfrac{\varphi(v_1, v_2)}{\sigma_{v_1}\sigma_{v_2}}$ (normalized covariance)

in imputing missing values in microarray data. Modeling nonlinear dependencies is a challenging problem, and the present work makes only a modest—nevertheless, empirically significant—step into the realm of nonlinear dependence by modeling the third-moment characteristics of the quantity $\mathfrak{F}$. In principle, the proposed extension admits any order moment, but computational constraints limit the present developments. However, even the single step to a third-moment extension under severe sampling fluctuations is very challenging, and, in spite of this modest modeling enhancement, it is a hard-won extension yielding significantly improved estimates for a range of real and simulated examples (see Section 5).

Thus, a central finding of this work is that null statistic histogram approaches can be improved by including third-moment skewness corrections. Advancing the techniques to model higher order dependencies is challenging, but the effort could have a substantial payoff. Errors in gene detection are expensive in financial terms, but the derailing of biomedical research resulting from a false gene discovery could be profoundly costly in many ways. Even modest improvements in genomic techniques are potentially very significant.

2. Notation and Terminology

Because this paper has a practical aim, we will assume, without comment, "friendly" mathematical conditions such as existence of distributions, measurability, and sure convergence of integrals. Even so, the mathematical developments in this paper necessarily involve extensive notation and we strive for consistency and clarity in its use. Quantities are generally formulated as RVs unless stated otherwise. This excludes obviously deterministic quantities like sequence indices, integers defined in the abstract (e.g., the number of microarrays, "M"), and statistical expectations. Precise formulations require that probability distributions ordinarily be denoted formally as, for example, $p_{v_1,v_2}(\xi_1, \xi_2)$ for the joint distribution of RVs v_1 and v_2, but the more common abusive notation "$p(v_1, v_2)$" is more expedient in a few cases. The notation $p(\cdot)$ may denote either a discrete or continuous (i.e., density) distribution, depending, of course, on the RV(s) being modeled. The meaning should be clear in context. We deliberately allow this ambiguity because it avoids some notational awkwardness as discrete distributions are fitted with densities. On the other hand, the notation $\mathbb{P}(A)$ is used to denote a probability assignment to a measurable event A.

Many developments in this paper are centered on second-order statistical concepts. It is important to carefully define terminology used in this regard, since the vocabulary has nuanced differences across disciplines. The elementary notation for scalar RVs in Table 1 is standard and is used conventionally in this paper. A caveat arises in the discussion of related matrices, however. The term "correlation matrix" is used in this paper in a way consistent with its use in many statistical developments, but not in a way that is universal across disciplines. The following definitions are used throughout this paper.

Definitions 1. Consider a random vector $\mathbf{v}^T = [v_1 \cdots v_G]$ with mean vector $\boldsymbol{\mu}_\mathbf{v} \stackrel{\text{def}}{=} \mathcal{E}\{\mathbf{v}\}$. Then, the *covariance matrix* associated with $\mathbf{v}$ is defined as

$$\boldsymbol{\Sigma}_\mathbf{v} \stackrel{\text{def}}{=} \mathcal{E}\left\{\left(\mathbf{v} - \boldsymbol{\mu}_\mathbf{v}\right)\left(\mathbf{v} - \boldsymbol{\mu}_\mathbf{v}\right)^T\right\} \in \mathbb{R}^{G\times G}, \tag{1}$$

in which the (i, j) element is $\varphi(v_i, v_j)$. The *correlation matrix* of $\mathbf{v}$ is

$$\mathbf{R}_\mathbf{v} \stackrel{\text{def}}{=} \mathcal{E}\left\{\mathbf{S}^{-1}\left(\mathbf{v} - \boldsymbol{\mu}_\mathbf{v}\right)\left(\mathbf{v} - \boldsymbol{\mu}_\mathbf{v}\right)^T\mathbf{S}^{-1}\right\}, \tag{2}$$

in which the (i, j) element is $\varphi(v_i, v_j)$ and $\mathbf{S}$ is a diagonal matrix with (i, i) element $= \sigma_{v_i}$, the standard deviation of the ith RV, v_i.

In this paper, *the term "correlation matrix" will refer to the definition in* (2). On the contrary, in much of the engineering literature, the outer product $\mathcal{E}\{\mathbf{v}\mathbf{v}^T\} = \boldsymbol{\Sigma}_\mathbf{v} + \boldsymbol{\mu}_\mathbf{v}\boldsymbol{\mu}_\mathbf{v}^T$ is called the *(spatial) correlation matrix*. In this case the (i, j) element of the matrix is the scalar correlation $\mathcal{E}\{v_i v_j\}$. In our definition the elements are correlation coefficients, which are, in fact, normalized covariances. One significant implication of this fact is that mean values of RVs have no effect on either matrix. This should be kept in mind in the developments to follow.

3. Problem Formulation

G genes are to be studied using M microarray experiments. The expression values are recorded in an $G \times M$ matrix, $\mathbf{X} = [x_{gm}]$. For analytical purposes, the expression quantities x_{gm} are generally RVs. Each of the M microarray experiments takes place under one of two conditions (indexed by $k = 1$ or 2) such as control and treatment. These two subsets of the data are called *treatment groups* in the paper. There are M_k samples (i.e., microarrays) in treatment group k. Based on the evidence in $\mathbf{X}$, we seek to identify a "small" number, $G_* \ll G$, of genes that are significantly differentially expressed across the two treatment groups. One widely used

strategy (e.g., [5, 14]) is to posit that each of the genes, for $g = 1, 2, \ldots G$ satisfies a *null hypothesis*,

$$\mathbb{H}_{0,g}: \text{ Gene } g \text{ is } \textit{not} \text{ differentially expressed in the two treatment groups.} \tag{3}$$

All G genes are tested against this hypothesis using two-sample null statistics $z_1, z_2, \ldots, z_G$ [14]. The magnitudes of z_g scores establish a gene ranking, and the G_* genes with the largest scores are reported as statistically significant discoveries.

Clearly, the list of G_* discovered genes is only meaningful to the extent that $\mathfrak{F}$ is very small. Of course, $\mathfrak{F}$ can only be estimated since the state of any gene (i.e., whether or not it should be "discovered") is unknown. Strong causal relationships among genes give rise to highly correlated z_g scores and greatly complicate the estimate of $\mathfrak{F}$ [30, 31]. Moreover, in spite of their declining cost, microarrays are still a relatively expensive technology. Consequently, the number of microarrays, M, in an experiment is usually smaller than number of genes, G, by as much as four orders of magnitude. Typically, existing microarrays record expression data for at least a few thousand genes. The fact that $M \ll G$ further complicates the problem because the knowledge about the underlying gene-gene correlation structure is critically sparse in the observations. At the same time, the consequences of correlation on differential analysis cannot be overlooked [35]. In fact, the present work will suggest that even nonlinear dependencies must be accounted for in order to properly estimate $\mathfrak{F}$. Theoretical justifications for this contention are given momentarily.

This paper develops a moment-based estimator of $\mathfrak{F}$ by giving the null z histogram a stochastic interpretation. The observed null counts are viewed as realizations of a more fundamental random model shaped by inter-z_g dependence. A small zero-symmetric bin in the space of z statistics is designated as the *center area*, and it is posited that no z_g scores from "nonnull" genes fall in this range. Then, the null count in the center area, say C, is observable, and by conditioning $\mathfrak{F}$ on C, the variance of $\mathfrak{F}$ can be reduced significantly [5, 49]. We relate $\mathfrak{F}$ to C through the discrete joint distribution $p(\mathfrak{F}, C)$. To obtain an approximation of $p(\mathfrak{F}, C)$, we estimate its first three moments, then fit the maximum entropy function (Appendix B). This approach inherently yields an estimate of the conditional distribution $p(\mathfrak{F} \mid C)$. A large number of estimates of the distribution of $\dot{\mathfrak{F}}$ would theoretically be more useful than a point estimate because of the noisy nature of large-scale inferences [30]. Compared to histogram-curve-fitting techniques like empirical null [5], however, the present approach enjoys the attractive feature that covariance is separately estimated, and then explicitly incorporated into the inference.

Efron [31] reports that RVs $\mathfrak{F}$ and C are found to be extremely negatively correlated in a number of real experiments. He provides an explanation for this finding, then employs these insights to develop a Poisson-model-based second-order estimator of $\mathfrak{F}$ which, like the present approach, relies on the center area concept. While Efron's work is extremely important, his own research has gone on to show that purely second-order $\mathfrak{F}$ estimates suffer from over- and under-estimation effects. The second-moment estimates of $\mathfrak{F}$ reported later in the present paper (see Section 5), as well as those in the cited Efron paper, all show these adverse effects. There are three contributory factors: (i) $\mathfrak{F}$ is bounded below by zero, (ii) the mean of $\mathfrak{F}$ is small, and (iii) intergene covariance causes the variance of $\mathfrak{F}$ to inflate. All of these factors suggest that skewness corrections—reflecting nonlinear dependence—are vital.

4. Methods

4.1. Moments of the Joint Distribution $P(\mathfrak{F}, C)$

4.1.1. Count Models. The process begins by transforming test t statistics to z values as $z_g = P_{\mathcal{G}_u}^{-1}\{P_0(t_g)\}$, $g = 1, \ldots, G$, where P_0 is the putative null cumulative distribution function (c.d.f.) of the test statistic, and $P_{\mathcal{G}_u}^{-1}$ is the inverse c.d.f. of the unit normal density, $p_{\mathcal{G}_u} \equiv \mathcal{G}(0, 1)$. The z values, modeled as RVs, provide the analytical convenience of multivariate normal form while describing the joint t-statistic behavior. We formally define the fundamental quantities:

$$\begin{aligned} \mathfrak{F} &\stackrel{\text{def}}{=} \#\left\{z_g : \ z_g \le \delta \cap \mathbb{H}_{0,g} \ \text{ is true}\right\}, \\ C &\stackrel{\text{def}}{=} \#\left\{z_g : \ \left|z_g\right| \le c \cap \mathbb{H}_{0,g} \text{ is true}\right\}, \end{aligned} \tag{4}$$

in which $\#\{\mathcal{S}\}$ denotes the number of elements in the discrete set $\{\mathcal{S}\}$. $\mathcal{Z} \subseteq \mathbb{R}$ is the sample space of z values. The interval $\mathcal{Z}_C \stackrel{\text{def}}{=} \{z \in \mathcal{Z} : \ |z_g| < c\}$ corresponding to count C is called the *center area*, and the semi-infinite interval $\mathcal{Z}_{\mathfrak{F}} \stackrel{\text{def}}{=} \{z \in \mathcal{Z} : \ z \le \delta\}$ associated with count $\mathfrak{F}$ is the *left tail area*. For proper comparison with Efron's results [31], we work with a left tail area; however, the present approach can employ right- or double-sided tail areas equally well.

The premise that very few nonnull z_g scores fall in $\mathcal{Z}_C$ and, hence, that C is practically observable is of prime importance. A similar assumption plays a central role in the literature on estimating the proportion of null genes, as in, for example, papers by Efron [31], Pawitan et al. [32], and Langaas et al. [50]. The *empirical null* approach [5] relies on similar reasoning. We exploit the observability of C to: (i) estimate the moments of $p(\mathfrak{F}, C)$, (ii) use them to infer the distribution (estimate) $\hat{p}(\mathfrak{F}, C)$, and then (iii) report (an estimated) $p(\mathfrak{F} \mid C)$ which in turn could be used to find an estimator of $\mathfrak{F}$ conditioned upon C. Initially, all cases are treated as null. Improvement is possible by estimating π_0 [50, 51].

4.1.2. Assumptions

Assumptions. The following assumptions underlie these developments:

(1) π_0 is large, say $\pi_0 \ge 0.9$ (Efron discusses this bound in [4]).

(2) z_g is a unit normal variate $[\sim \mathcal{G}(0,1)]$ for all $g = 1, 2, \ldots, G$.

(3) The z scores are jointly Gaussian to the third order (but *not* uncorrelated).

(4) Recall that x_{gm} [element (g,m) of the expression matrix $\mathbf{X}$] denotes (the RV model for) the expression of gene g on microarray m. Let $x_{g\bullet}$ denote the marginal RV for x_{gm}, that is, the model for the expression outcomes of gene g. The realizations of $x_{g\bullet}$ are the elements of row g of an observed $\mathbf{X}$. Then, it is assumed that $\varphi(z_g, z_{g'}) = \rho(z_g, z_{g'}) \approx \rho(x_{g\bullet}, x_{g'\bullet})$ for all g, g' (justified below).

4.1.3. The z-Value Histogram. It is convenient and computationally efficient to obtain the moments of $p(\mathfrak{F}, C)$ using the moments of the z-value histogram. We seek central moments because they facilitate working with the maximum entropy distribution (Appendix B). The moment estimation is carried out as follows. $\mathcal{Z}$ is partitioned into B disjoint bins, $\mathcal{Z} = \cup_{b=1}^{B} \mathcal{Z}_b$, where the bth bin has center $z^{[b]}$ and width Δ (constant with b). Then, the *z-histogram bin counts* are

$$Y_b = \#\left\{z_m : \; z_g \in \mathcal{Z}_b\right\} = \sum_{g=1}^{G} I_b\left(z_g\right), \quad \text{for} \;\; b = 1, \ldots, B, \tag{5}$$

in which $I_b(z_g)$ is the indicator function for the event "score z_g falls in bin $\mathcal{Z}_b$."

Consider bin b with count Y_b for some $b \in \{1, \ldots, B\}$. The *mean* of count Y_b is

$$\begin{aligned}
\mu(Y_b) &\stackrel{\text{def}}{=} \mathcal{E}\{Y_b\} = \mathcal{E}\left\{\sum_{g=1}^{G} I_b\left(z_g\right)\right\} \\
&= \sum_{g=1}^{G} \mathbb{P}\left(z_g \in \mathcal{Z}_b\right) = G \int_{z^{[b]}-(\Delta/2)}^{z^{[b]}+(\Delta/2)} p_{\mathcal{G}_u}(\xi)\, d\xi \\
&\approx \hat{\mu}(Y_b) \stackrel{\text{def}}{=} G\Delta p_{\mathcal{G}_u}\left(z^{[b]}\right), \qquad \text{where } p_{\mathcal{G}_u}(\xi) \equiv \mathcal{G}(0,1).
\end{aligned} \tag{6}$$

The *second-order joint central moment covariance* of the pair $(Y_b, Y_{b'})$, where b may equal b', is

$$\begin{aligned}
\varphi(Y_b, Y_{b'}) &\stackrel{\text{def}}{=} \mathcal{E}\{[Y_b - \mu(Y_b)][Y_{b'} - \mu(Y_{b'})]\} \\
&= \mathcal{E}\left\{\sum_{g=1}^{G} I_b\left(z_g\right) \sum_{g'=1}^{G} I_{b'}\left(z_{g'}\right)\right\} - \mu(Y_b)\mu(Y_{b'}) \\
&= \sum_{g \neq g'} \mathbb{P}\left(z_g \in \mathcal{Z}_b, z_{g'} \in \mathcal{Z}_{b'}\right) \\
&\quad + \sum_{g} \mathbb{P}\left(z_g \in \mathcal{Z}_b, z_g \in \mathcal{Z}_{b'}\right) - \mu(Y_b)\mu(Y_{b'}).
\end{aligned} \tag{7}$$

Because of the bivariate normality of z_g and $z_{g'}$, (7) can be approximated by:

$$\begin{aligned}
\hat{\varphi}(Y_b, Y_{b'}) &\stackrel{\text{def}}{=} \sum_{g \neq g'} \Delta^2 p_{\mathcal{G}}\left(z^{[b]}, z^{[b']}; \mathbf{0}, \mathbf{\Sigma}^{[g,g']}\right) \\
&\quad - \hat{\mu}(Y_b)[\hat{\mu}(Y_{b'}) + \delta_{bb'}],
\end{aligned} \tag{8}$$

where $\delta_{bb'}$ is the Kronecker delta, and $p_{\mathcal{G}}(\zeta_1, \zeta_2; \mathbf{0}, \mathbf{\Sigma}^{[g,g']}) \equiv \mathcal{G}(\mathbf{0}, \mathbf{\Sigma}^{[g,g']})$ is the bivariate Gaussian density with mean vector $\mathbf{0} = [0 \cdots 0]^T$, and covariance matrix (equivalent to the correlation matrix in this case)

$$\begin{aligned}
\mathbf{\Sigma}_z^{[g,g']} &\stackrel{\text{def}}{=} \mathcal{E}\left\{\begin{bmatrix} z_g \\ z_{g'} \end{bmatrix}\begin{bmatrix} z_g & z_{g'} \end{bmatrix}\right\} \\
&= \begin{bmatrix} 1 & \rho\left(z_g, z_{g'}\right) \\ \rho\left(z_{g'}, z_g\right) & 1 \end{bmatrix} = \mathbf{R}_z^{[g,g']}.
\end{aligned} \tag{9}$$

That is, defining the vector of arguments $\boldsymbol{\zeta} \stackrel{\text{def}}{=} [\zeta_1 \;\; \zeta_2]^T$,

$$\begin{aligned}
&p_{\mathcal{G}}\left(\boldsymbol{\zeta}; \mathbf{0}, \mathbf{\Sigma}_z^{[g,g']}\right) \\
&\quad = \frac{1}{2\pi\left|\mathbf{\Sigma}_z^{[g,g']}\right|^{1/2}} \\
&\qquad \times \exp\left\{-\frac{1}{2}\boldsymbol{\zeta}^T\left(\mathbf{\Sigma}_z^{[g,g']}\right)^{-1}\boldsymbol{\zeta}\right\} \\
&\quad = \frac{1}{2\pi\sqrt{1-\rho^2\left(z_g, z_{g'}\right)}} \\
&\qquad \times \exp\left\{-\frac{\zeta_1^2 - 2\rho\left(z_g, z_{g'}\right)\zeta_1\zeta_2 + \zeta_2^2}{2\left[1-\rho^2\left(z_g, z_{g'}\right)\right]}\right\},
\end{aligned} \tag{10}$$

where $|\cdot|$ denotes the determinant. In the approximation (8), the density in (10) is evaluated at $\zeta_1 = z^{[b]}$ and $\zeta_2 = z^{[b']}$.

As reflected in the second line of (10), the covariance matrix $\mathbf{\Sigma}_z^{[g,g']}$ is fully specified by a a single-scalar parameter, $\rho(z_g, z_{g'})$ for each g, g' pair. Thus, we can express the Gaussian density in (10) as being parameterized by this autocorrelation coefficient for the given z-score pair, $p_{\mathcal{G}}[\boldsymbol{\zeta}; \mathbf{0}, \rho(z_g, z_{g'})]$. Now, suppose that we can derive an *empirical density* [52], say $q_{\rho_z}(\cdot)$, over the interval $[-1, 1]$, fitted to the discrete set of $\binom{G}{2}$ autocorrelation coefficients, $\{\rho(z_g, z_{g'}),\ 1 \leq g, g' \leq G\}$. This empirical distribution allows the summation over gene indices in (10) to be replaced by a continuous computation. This smoothed computation represents a further approximation of $\varphi(Y_b, Y_{b'})$, which is stated in the form of a lemma below. This result is similar to those of Owen [30, Theorem 1] and Efron [31, Lemma 2].

Lemma 1. *Let $q_{\rho_z}(\xi)$ denote the empirical density of ρ_z, derived from the $\binom{G}{2}$ z-pair sample correlation coefficients, $\rho(z_g, z_{g'})$, $1 \leq g, g' \leq G$. Then, the second joint central*

moment of a histogram count pair $(Y_b, Y_{b'})$*, where* b *may equal* b'*, is approximated by*

$$\varphi(Y_b, Y_{b'}) \approx \hat{\varphi}(Y_b, Y_{b'}) \stackrel{\text{def}}{=} [G!]_2 \Delta^2 Q_{\rho_z}\left(z^{[b]}, z^{[b']}\right) - \hat{\mu}(Y_b)[\hat{\mu}(Y_{b'}) + \delta_{bb'}], \tag{11}$$

where $[G!]_\ell \stackrel{\text{def}}{=} G(G-1)\cdots(G-[\ell-1])$*, for* $1 \le \ell \le G$*, and*

$$Q_{\rho_z}(\zeta_1, \zeta_2) \stackrel{\text{def}}{=} \int_{-1}^{+1} \frac{q_{\rho_z}(\xi)}{2\pi\sqrt{1-\xi^2}} \times \exp\left\{-\frac{\zeta_1^2 - 2\xi\zeta_1\zeta_2 + \zeta_2^2}{2(1-\xi^2)}\right\} d\xi. \tag{12}$$

Further, the *third joint central moment* of a triplet $(Y_b, Y_{b'}, Y_{b''})$, where two or more indices may be equal, is:

$$\begin{aligned}
&\gamma(Y_b, Y_{b'}, Y_{b''}) \\
&\quad \stackrel{\text{def}}{=} \mathcal{E}\{[Y_b - \mu(Y_b)] \\
&\qquad \times [Y_{b'} - \mu(Y_{b''})][Y_{b'} - \mu(Y_{b''})]\} \\
&\quad = \mathcal{E}\{Y_b Y_{b'} Y_{b''}\} - \mu(Y_b)\mu(Y_{b'})\mu(Y_{b''}) \\
&\qquad - [\mu(Y_b)\varphi(Y_{b'}, Y_{b''}) \\
&\qquad\quad + \mu(Y_{b'})\varphi(Y_b, Y_{b''}) \\
&\qquad\quad + \mu(Y_{b''})\varphi(Y_b, Y_{b'})],
\end{aligned} \tag{13}$$

where

$$\begin{aligned}
&\mathcal{E}\{Y_b Y_{b'} Y_{b''}\} \\
&\quad = \mathcal{E}\left\{\sum_{g=1}^{G} I_b\left(z_g\right) \sum_{g'=1}^{G} I_{b'}\left(z_{g'}\right) \sum_{g''=1}^{G} I_{b''}\left(z_{g''}\right)\right\} \\
&\quad = \sum_{g \ne g' \ne g''} \mathbb{P}\left(z_g \in \mathcal{Z}_b, z_{g'} \in \mathcal{Z}_{b'}, z_{g''} \in \mathcal{Z}_{b''}\right) \\
&\qquad + \delta_{bb''} \sum_{g \ne g''} \mathbb{P}\left(z_g \in \mathcal{Z}_b, z_{g''} \in \mathcal{Z}_{b''}\right) \\
&\qquad + \delta_{bb'} \sum_{g \ne g'} \mathbb{P}\left(z_g \in \mathcal{Z}_b, z_{g'} \in \mathcal{Z}_{b'}\right) \\
&\qquad + \delta_{b'b''} \sum_{g' \ne g''} \mathbb{P}\left(z_{g'} \in \mathcal{Z}_{b'}, z_{g''} \in \mathcal{Z}_{b''}\right) \\
&\qquad + \delta_{bb'b''} \sum_{g} \mathbb{P}\left(z_g \in \mathcal{Z}_b\right),
\end{aligned} \tag{14}$$

in which $\delta_{bb'b''}$ is the Kronecker sequence over $\mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$.

The assumed trivariate normality of the z scores implies that the joint distribution for each score triplet is specified by a 3×3 covariance matrix. Let us denote the covariance (equivalent to correlation) matrix for the z-value triplet $(z_g, z_{g'}, z_{g''})$ by

$$\begin{aligned}
\mathbf{R}_z^{[g,g',g'']} &= \mathcal{E}\begin{bmatrix} z_g \\ z_{g'} \\ z_{g''} \end{bmatrix}\begin{bmatrix} z_g & z_{g'} & z_{g''} \end{bmatrix} \\
&= \begin{bmatrix} 1 & \rho\left(z_g, z_{g'}\right) & \rho\left(z_g, z_{g''}\right) \\ \rho\left(z_{g'}, z_g\right) & 1 & \rho\left(z_{g'}, z_{g''}\right) \\ \rho\left(z_{g''}, z_g\right) & \rho\left(z_{g''}, z_{g'}\right) & 1 \end{bmatrix}.
\end{aligned} \tag{15}$$

For each z-score triplet, $\mathbf{R}_z$ is an element of the space—call it $\mathcal{R}^3$—of all symmetric positive-semidefinite matrices in $\mathbb{R}^{3\times 3}$ with element magnitudes no greater than unity. $\mathbf{R}_z$ is continuously distributed over $\mathcal{R}^3$.

Again, we need an empirical way to compute the joint moments of the z scores. Let $q_{\mathbf{R}_z}(\Xi)$ be the empirical density of the $\binom{G}{3}$ correlation matrices, $\mathbf{R}_z$. This density must be inferred from the observed data. Of practical importance is the fact that, although each $\mathbf{R}_z$ is distributed over a subspace of $\mathbb{R}^{3\times 3}$, the domain of $q_{\mathbf{R}_z}$ is effectively a three-dimensional manifold of that space because each covariance matrix is unambiguously determined by its three values above or below its main diagonal (with unity diagonal elements). The argument Ξ may be thought of as a vector of these three elements (over the continuum of allowable values), but we will continue to denote it as a matrix as a reminder of its association with $\mathbf{R}_z$. We have, in conjunction with (6)–(14), the following useful approximation.

Lemma 2. *The third-order joint central moment of a histogram count triplet* $(Y_b, Y_{b'}, Y_{b''})$*, where two or more indices may be equal, is approximated by*

$$\begin{aligned}
&\hat{\gamma}(Y_b, Y_{b'}, Y_{b''}) \\
&\quad \stackrel{\text{def}}{=} [G!]_3\ \Delta^3 Q_{\mathbf{R}_z}\left(z^{[b]}, z^{[b']}, z^{[b'']}\right) \\
&\qquad + [G!]_2\ \Delta^2\left[\delta_{bb''} Q_{\rho_z}\left(z^{[b]}, z^{[b'']}\right)\right. \\
&\qquad\qquad + \delta_{bb'} Q_{\rho_z}\left(z^{[b]}, z^{[b']}\right) \\
&\qquad\qquad \left. + \delta_{b'b''} Q_{\rho_z}\left(z^{[b']}, z^{[b'']}\right)\right] \\
&\qquad + \delta_{bb'b''}\hat{\mu}(Y_b) - \hat{\mu}(Y_b)\hat{\mu}(Y_{b'})\hat{\mu}(Y_{b''}) \\
&\qquad - [\hat{\mu}(Y_b)\hat{\varphi}(Y_{b'}, Y_{b''}) + \hat{\mu}(Y_{b'})\hat{\varphi}(Y_b, Y_{b''}) \\
&\qquad\quad + \hat{\mu}(Y_{b''})\hat{\varphi}(Y_b, Y_{b'})],
\end{aligned} \tag{16}$$

where

$$\begin{aligned}
&Q_{\mathbf{R}_z}(\zeta_1, \zeta_2, \zeta_3) \\
&\quad = \int_{\mathcal{R}^3} \frac{q_{\mathbf{R}_z}(\Xi)}{(2\pi)^{3/2}|\Xi|^{1/2}} \\
&\qquad \times \exp\left\{-\frac{1}{2}\begin{bmatrix}\zeta_1 & \zeta_2 & \zeta_3\end{bmatrix}\Xi^{-1}\begin{bmatrix}\zeta_1 & \zeta_2 & \zeta_3\end{bmatrix}^T\right\} d\Xi,
\end{aligned} \tag{17}$$

$\hat{\mu}$, Q_{ρ_z}, $\hat{\varphi}$, and $[G!]_\ell$ are defined in (6) *and* (11)*, and Δ is the z-histogram bin width.*

To obtain the moments of $p(\mathfrak{F}, C)$ it is simply necessary to combine the moments of the corresponding Y_b counts. For example,

$$\begin{aligned}
\sigma_{\mathfrak{F}}^2 &= \mathcal{E}\left\{(\mathfrak{F}-\mu_{\mathfrak{F}})^2\right\} \approx \sum_{\{b,\, b' :\, \mathcal{Z}_b, \mathcal{Z}_{b'} \subset \mathcal{Z}_C\}} \hat{\varphi}(Y_b, Y_{b'}),\\
\sigma_C^2 &= \mathcal{E}\left\{(C-\mu_C)^2\right\} \approx \sum_{\{b,\, b' :\, \mathcal{Z}_b, \mathcal{Z}_{b'} \subset \mathcal{Z}_C\}} \hat{\varphi}(Y_b, Y_{b'}),\\
\varphi(C,\mathfrak{F}) &= \mathcal{E}\left\{(C-\mu_C)^2(\mathfrak{F}-\mu_{\mathfrak{F}})^2\right\}\\
&\approx \sum_{\{b,\, b' :\, \mathcal{Z}_b \subset \mathcal{Z}_{\mathfrak{F}},\, \mathcal{Z}_{b'} \subset \mathcal{Z}_C\}} \hat{\varphi}(Y_b, Y_{b'}).
\end{aligned} \tag{18}$$

The key quantities in the lemmas for approximating moments are the empirical covariance densities. Obtaining these densities in the presence of severe sampling errors is discussed next.

4.2. Empirical Correlation Densities

4.2.1. Approach. Because of severe sampling fluctuations, the current methods can recover only $q_{\rho_z}(\xi)$ from the data. This density is then used to estimate $q_{\mathbf{R}_z}(\Xi)$. For this purpose, as well as to facilitate the calculations of Lemmas 1 and 2, we seek to parameterize the requisite densities. For most real examples, a single omnibus parameter α is found to be sufficient.

4.2.2. Data Normalization. For all-null false discovery rate calculations, normalization of the per-microarray expression results has been found to be beneficial [31, Remark E]. The columns of the data matrix $\mathbf{X}$ are standardized to mean zero and unity variance. This standardization normalizes output "brightness" among microarrays [53, 54]. It also forces the sum of covariances, and approximately the sum of correlations, to be zero. This permits the fitting of a zero symmetric density to $q_{\rho_z}(\xi)$, which, in turn, has profound consequences for the form of $q_{\mathbf{R}_z}(\Xi)$.

Formally, let $\mathbf{X}^o$ denote the *residual expression matrix*, obtained by subtracting from $\mathbf{X}$ each gene's average response within each treatment group, and let $x^o_{g\bullet}$ denote the marginal random variable modeling the residual expression outcomes for gene g [like $x_{g\bullet}$ of Assumption (4), p. 5]. All further discussion of gene expression values will refer to these normalized residual values.

4.2.3. Obtaining $q_{\rho_z}(\xi)$. The empirical densities q_{ρ_z} and $q_{\mathbf{R}_z}$, as well as others to be introduced below, clearly play a key role in moment estimation above. In each case, the empirical density—a surrogate for the true statistical density of the correlation coefficient(s) being modeled—is a distribution of a correlation function or matrix over a continuum, but it must be inferred from the data samples.

To deduce $q_{\rho_z}(\xi)$, we require $q_{\rho_x}(\xi)$—the empirical density of the $\binom{G}{2}$ *gene expression* correlation coefficients. The mapping between the domains of q_{ρ_z} and q_{ρ_x} is needed, in principle, to calibrate q_{ρ_z}. However, for the usual two-sample t-statistic, assuming independent columns in $\mathbf{X}^o$, $\rho(z_g, z_{g'}) \approx \rho(x_{g\bullet}, x_{g'\bullet})$ [recall Assumption (4), p. 5]; hence, $q_{\rho_z}(\xi) \approx q_{\rho_x}(\xi)$. We make the assumption of equality of these densities below.

Let $\overline{\varphi}(x^o_{g\bullet}, x^o_{g'\bullet})$ denote the *sample covariance* between rows (genes) g and g' of $\mathbf{X}^o$. For convenience, we define the notation

$$\overline{\rho}_{gg'} \stackrel{\text{def}}{=} \overline{\rho}\left(x^o_{g\bullet}, x^o_{g'\bullet}\right) = \frac{\overline{\varphi}\left(x^o_{g\bullet}, x^o_{g'\bullet}\right)}{\sqrt{\overline{\varphi}\left(x^o_{g\bullet}, x^o_{g\bullet}\right)\overline{\varphi}\left(x^o_{g'\bullet}, x^o_{g'\bullet}\right)}}. \tag{19}$$

These are the values to be fit with density $q_{\rho_x}(\xi)$.

To reduce the variability added by sampling errors, we apply the Fisher transform:

$$\overline{\tau}_{gg'} = \frac{1}{2}\log\frac{1+\overline{\rho}_{gg'}}{1-\overline{\rho}_{gg'}}. \tag{20}$$

For bivariate normal samples, the Fisher transform has remarkable normalizing and variance stabilizing properties [55], and each $\overline{\tau}_{gg'}$ is well modeled by the distribution $\overline{\tau}_{gg'} \sim \mathcal{G}(\tau_{gg'}, [G-3]^{-1})$, where $\tau_{gg'}$ is the Fisher-transformed underlying correlation coefficient. Assuming a sampling model

$$\overline{\tau}_{gg'} = \tau_{gg'} + \varepsilon; \qquad \tau_{gg'} \sim p_\tau(\xi), \tag{21}$$

where p_τ is the distribution of the Fisher-transformed underlying correlation coefficients, we can interpret the histogram of Fisher-transformed sample correlations, say the "$\overline{\tau}_{gg'}$-histogram," as a convolution of $p_\tau(\xi)$, the statistical correlation density on the scale resulting from the Fisher transform, and the histogram of sampling errors, say, the "ε-histogram," also on the τ-scale. Then, the underlying $p_\tau(\xi)$ is obtained by deconvolving this density from the convolved pair, $p_{\overline{\tau}}(\xi) = p_\tau(\xi) * p_\varepsilon(\xi)$. For a wide variety of microarray data sets studied in this work (also see [30]), the normal distribution $\mathcal{G}(0, \sigma^2)$ fits nicely to the $\overline{\tau}_{gg'}$ histogram. For bivariate normal samples, where $\varepsilon \sim \mathcal{G}(0, [G-3]^{-1})$, the estimate of $p_\tau(\xi)$, say $\hat{p}_\tau(\xi)$, takes the normal form $\mathcal{G}(0, \sigma^2 - [G-3]^{-1})$. It is this estimate that will serve as the empirical density of the Fisher-transformed correlations, $q_\tau \equiv \hat{p}_\tau$.

Having obtained the underlying $q_\tau(\xi)$, we must, in principle, undo the mapping (20) to obtain $q_{\rho_x}(\xi)$, then deduce q_{ρ_z} from q_{ρ_x}. Recall, however, that we assume that the correlation coefficients of the z and x^o variables are identical [31], so that we may directly seek $q_{\rho_z}(\xi) = q_{\rho_x}(\xi)$ from $q_\tau(\xi)$. A distribution that fits the inverse-transformed $q_\tau(\xi)$ well is

$$q_{\rho_z}(\xi) \propto (1-\xi^2)^a = \left[\frac{1}{2}(\xi+1)\right]^\alpha\left[1-\frac{1}{2}(\xi+1)\right]^\alpha, \qquad |\xi| \le 1, \tag{22}$$

a class of densities in the general *Beta* distribution family given by:

$$p_B(\xi;\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\xi^{\alpha-1}(1-\xi)^{\beta-1}, \quad 0 \le \xi \le 1, \tag{23}$$

where B is the *Beta* function and α and β are nonnegative shape parameters. Comparing (22) and (23) gives a useful probabilistic interpretation of the correlation coefficient, say $\widetilde{\rho}_z$, modeled by the empirical density q_{ρ_z}: We see from (22) that $\widetilde{\rho}_z \sim p_B(0.5\xi + 1; \alpha, \alpha)$. Therefore, $\sigma^2_{\widetilde{\rho}_z}$ will be a factor of four greater than the variance of a *Beta*-distributed random variable with parameters $\alpha = \beta$. That is,

$$\sigma^2_{\widetilde{\rho}_z} = 4\frac{\alpha^2}{(2\alpha)^2(2\alpha+1)} \Longrightarrow \alpha = \frac{1-\sigma^2_{\widetilde{\rho}_z}}{2\sigma^2_{\widetilde{\rho}_z}}. \tag{24}$$

Thus, using $\sigma^2_{\widetilde{\rho}_z}$ as an estimate of $\sigma^2_{\widetilde{\rho}_z}$, we obtain the parameter α, hence, the distribution q_{ρ_z}.

4.2.4. Obtaining $q_{R_x}(\Xi)$. We now pursue $q_{\mathbf{R}_x}(\boldsymbol{\Xi})$ as an extension of $q_{\rho_x}(\xi)$. Like $q_{\mathbf{R}_z}$, the effective domain of $q_{\mathbf{R}_x}$ is of only three dimensions. Also similarly to the scalar density, for the two-sample t-statistic, $q_{\mathbf{R}_z}(\boldsymbol{\Xi}) \approx q_{\mathbf{R}_x}(\boldsymbol{\Xi})$. Hence, we can pursue $q_{\mathbf{R}_z}$ indirectly by finding $q_{\mathbf{R}_x}$.

We seek a joint distribution on the space of all 3×3 correlation matrices such that all the inherent marginal distributions (i.e., the distributions of $\rho(x^o_{g\bullet}, x^o_{g'\bullet})$ for $g \neq g'$) are equivalent to $q_{\rho_x}(\xi)$. Such a result can be obtained from the inverse-Wishart distribution whose marginalization properties are helpful when studying a subset of variables [56]. Suppose that the true statistical covariance matrix of $\mathbf{X}^o$, say,

$$\boldsymbol{\Sigma}_{\mathbf{X}^o} \stackrel{\text{def}}{=} \mathcal{E}\left\{\mathbf{X}^o(\mathbf{X}^o)^T\right\} \in \mathcal{R}^G \subset \mathbb{R}^{G\times G} \quad \left(\text{for simplicity } \boxed{\boldsymbol{\Sigma}^o \stackrel{\text{def}}{=} \boldsymbol{\Sigma}_{\mathbf{X}^o}}\right) \tag{25}$$

follows the inverse-Wishart distribution $\mathcal{W}_G^{-1}(\mathbf{I}, \nu)$, $\nu \ge G$,

$$\boldsymbol{\Sigma}^o \sim p_{\boldsymbol{\Sigma}^o}(\boldsymbol{\Xi} \mid \nu \ge G) \propto |\boldsymbol{\Xi}|^{-0.5(\nu+G+1)} \times \exp\left\{-0.5 \mid \mathrm{tr}\left\{\boldsymbol{\Xi}^{-1}\right\}\right\}, \quad \boldsymbol{\Xi} \in \mathbb{R}^{G\times G}, \tag{26}$$

where ν is the single parameter that characterizes the distribution, and $\mathrm{tr}\{\cdot\}$ indicates the trace. The goal is to relate ν to parameter α of (22) and to determine the distribution of any of the 3×3 covariance submatrices of $\boldsymbol{\Sigma}^o$.

Following the *separation strategy* of Barnard et al. [57], we decompose $\boldsymbol{\Sigma}^o$ into its variances and normalized covariances (i.e., correlation coefficients) as

$$\boldsymbol{\Sigma}^o = \mathbf{S}\mathbf{R}^o\mathbf{S}, \tag{27}$$

where $\mathbf{S} \in \mathbb{R}^{G\times G}$ is the diagonal matrix whose ith diagonal element, s_i, is the standard deviation of the gene i residual expression [recall (2)]. $\mathbf{R}^o \stackrel{\text{def}}{=} \mathbf{R}_{\mathbf{X}_o}$ is the $G \times G$ correlation matrix of the residual expression matrix $\mathbf{X}^o$. Under the transformation $\boldsymbol{\Sigma}^o \rightarrow (\mathbf{S}, \mathbf{R}^o)$, the Jacobian is given by $(2\prod_i s_i)^G$ [58, Theorem 3]. Thus, after marginalization over $\mathbf{S}$:

$$\mathbf{R}^o \sim p_{\mathbf{R}^o}(\boldsymbol{\Xi} \mid \nu) \propto |\boldsymbol{\Xi}|^{-0.5(\nu+G+1)} \prod_{i=1}^{G} \int_0^\infty s_i^{-(\nu+1)} e^{-\xi^{ii}/2s_i^2} ds_i, \tag{28}$$

where ξ^{ii} is the ith diagonal element of $\boldsymbol{\Xi}^{-1}$. The product arises because of independence of the s_i elements. Substituting $\omega_i = \xi^{ii}/2s_i^2$ yields

$$\mathbf{R}^o \sim p_{\mathbf{R}^o}(\boldsymbol{\Xi} \mid \nu) \propto |\boldsymbol{\Xi}|^{-0.5(\nu+G+1)} \left(\prod_i \xi^{ii}\right)^{-0.5\nu} \times \left(\prod_i \int_0^\infty \omega_i^{0.5(\nu-2)} e^{-\omega_i} d\omega_i\right), \tag{29}$$

which leads to an expression for the probability density of the matrix $\mathbf{R}^o$:

$$p_{\mathbf{R}^o}(\boldsymbol{\Xi} \mid \nu) \propto |\boldsymbol{\Xi}|^{0.5(\nu-1)(G-1)-1} \left(\prod_i |(\boldsymbol{\Xi})_{ii}|\right)^{-0.5\nu}, \tag{30}$$

where $(\mathbf{A})_{ii}$ denotes the ith principal submatrix of $\mathbf{A}$, and where we have used the fact that $\xi^{ii} = |(\boldsymbol{\Xi})_{ii}|/|\boldsymbol{\Xi}|$. For $\mathbf{R}^o$ with probability density (30), the marginal density of its arbitrary correlation submatrix also has a useful expression.

Lemma 3. *For a correlation matrix* $\mathbf{R}^o \in \mathcal{R}^G$ *with the probability density* (30)*, the* $\kappa \times \kappa$ *correlation submatrix,* $\mathbf{R}^o_\kappa \in \mathcal{R}^\kappa$*, has the density*

$$p_{\mathbf{R}^o_\kappa}(\boldsymbol{\Xi}_\kappa \mid \nu) \propto |\boldsymbol{\Xi}_\kappa|^{0.5(\nu-G+\kappa-1)(\kappa-1)-1} \times \left(\prod_i |(\boldsymbol{\Xi}_\kappa)_{ii}|\right)^{-0.5(\nu-G+\kappa)}, \quad \boldsymbol{\Xi}_\kappa \in \mathcal{R}^\kappa. \tag{31}$$

Proof. Suppose that the $\kappa\times\kappa$ statistical covariance submatrix $\boldsymbol{\Sigma}^o_\kappa$ undergoes the transformation $\boldsymbol{\Sigma}^o_\kappa \rightarrow (\mathbf{S}_\kappa, \mathbf{R}^o_\kappa)$, where $\mathbf{S}_\kappa$ is the diagonal scaling matrix of appropriate standard deviations [recall (27)]. Then due to the marginalization property of inverse-Wishart, $\mathbf{R}^o_\kappa \sim \mathcal{W}_\kappa^{-1}(\mathbf{I}, \nu - G + \kappa)$. Following steps(26)–(30) for $\boldsymbol{\Sigma}^o_\kappa$ yields (31). □

Substituting $\kappa = 2$ in result (31) yields

$$p_{\mathbf{R}^o_2}(\boldsymbol{\Xi}_2 \mid \nu) \equiv p_{\rho_{12}}(\xi \mid \nu \ge G) \propto (1-\xi^2)^{0.5(\nu-G-1)}, \quad |\xi| \le 1. \tag{32}$$

Note that this density is the function of a scalar argument, the single value of the off-diagonal elements of $\mathbf{R}^o_2$. We have indicated this by use of the subscript "ρ_{12}" in the second term in the expression. The critical property of this result is that it

has the same uniparametric form as (22). By setting $\nu - G = 2\alpha + 1$ we can force the inherent marginal densities of the $\mathbf{R}^o$ entries $(\rho(x^o_{g\bullet}, x^o_{g'\bullet}),\ g \neq g')$ to equal $q_{\rho_x}(\xi)$—the specific aim of this derivation.

Finally, substituting $\kappa = 3$ in result (31), we obtain

$$q_{\mathbf{R}_x}(\Xi) \propto \frac{\left(1 - \xi_{12}^2 - \xi_{23}^2 - \xi_{13}^2 + 2\xi_{12}\,\xi_{23}\,\xi_{13}\right)^{2(\alpha+1)}}{\left[(1-\xi_{12}^2)(1-\xi_{23}^2)(1-\xi_{13}^2)\right]^{\alpha+2}}, \quad (33)$$

in which ξ_{ij} is the (i, j) element of the evaluated matrix (in the abstract) Ξ. The density of a particular covariance matrix in $\mathcal{R}^3$, say $\Xi = \mathbf{R}_x$, involves the use of the three elements in the upper triangle of the matrix, reinforcing earlier assertions that the domain of $q_{\mathbf{R}_x}$ is a manifold of the matrix space. Recall that $\mathbf{R}_x^{[g,g',g'']}$ (assumed equivalent to $\mathbf{R}_z^{[g,g',g'']}$) is the original notation for the 3×3 covariance matrix of a score triplet $(z_g, z_{g'}, z_{g''})$, and, by extension, $(x^o_{g\bullet}, x^o_{g'\bullet}, x^o_{g''\bullet})$. In the present discussion, $\mathbf{R}_x$ assumes the role of a 3×3 submatrix of $\mathbf{R}^o$, namely, $\mathbf{R}^o_3$.

For large G, the assumption of the inverse-Wishart distribution in (26) is not well justified. However, the assumption is used here strictly for its value in deducing $q_{\mathbf{R}_x}(\Xi)$ from $q_{\rho_x}(\xi)$. There is no concern for a model of the entire matrix $\mathbf{R}^o = \mathbf{R}_{\mathbf{X}^o}$. Further, single-parameter distributions on a positive definite matrix space are few. The inverse-Wishart distribution is chosen for its useful marginalization property. Of course, a tenuous assumption is not justified by a useful property if it leads to an unuseful procedure. The practical validation of the assumption is manifest in Section 5. The "Bayesian correlation priors" point of view from the work of Liechty et al. [59] was especially helpful in formulating these ideas. Exploring other ways to obtain $q_{\mathbf{R}_x}$ is a worthwhile but challenging endeavor.

Finally, we have derived $p^o_{\mathbf{R}_\kappa}(\Xi_\kappa|\nu)$ only up to a proportionality constant. However, for $q_{\mathbf{R}_x}(\Xi)$ $(\kappa = 3)$, normalization is straightforward. For $\kappa \geq 4$ it is necessary to resort to Monte Carlo methods which only require densities up to a proportionality constant.

5. Results

5.1. Testing on Real Data Sets. A MATLAB implementation of the approach based on the work above is available at the website http://www.egr.msu.edu/~deller/. The methods were tested on two real data sets, both showing a significant amount of intergene covariance but exactly opposite $\mathfrak{F}$ behaviors. Calculations below are for left-sided tail-area parameter $\delta = -2.5$, and center area parameter $c = 1$ [recall (4)]. Comparisons with Efron's [31] second-order estimator of $\mathfrak{F}$ are made.

The first data set is from the breast cancer (BRCA) study of Hedenfalk et al. [60]. These data record the expression of $G = 3226$ genes on $M = 15$ microarrays with seven samples assigned to BRCA1 mutations and eight to BRCA2. The original research seeks to identify genuine mRNA activity differences between these two categories. In the present paper, the logarithm is applied to the mRNA levels to increase Gaussianity [61].

In a study of the human immunodeficiency virus (HIV), Van 't Wout et al. [62] investigated $G = 7680$ genes over $M = 8$ microarrays with four samples assigned to an HIV infected condition and the remaining four to the control. To produce the test cells, the control cells (CD4 T cell lines) were infected by the HIV $- 1_{\text{BRU}}$ virus. The paper reports raw mRNA levels which, like the BRCA data, are converted to logarithms in the present work.

The present analysis reduces an entire expression matrix to two numbers: the center area null count, C, and the omnibus parameter, α. As is evident in Figure 1, the parametrization described in Section 4 is realistic. For the BRCA data, $\alpha = 17.77$, and for HIV, $\alpha = 3.51$. The Figure 1 caption provides details.

The next step is to compute the moments of $p(\mathfrak{F}, C)$ per Lemmas 1 and 2. These calculations require parameters G, α, c, δ, and Δ. We set $\Delta = 0.1$. These estimated moments are used to find the maximum-entropy (maxent) distribution $\hat{p}(\mathfrak{F}, C)$ (Appendix B). Figure 2, for example, reports the moments and the corresponding maxent distribution for the BRCA data. $\mathfrak{F}$ and C exhibit strong negative correlation of -0.89, a value similar to that in [31, Table 1]. Furthermore, $\mathfrak{F}$ shows significant positive skewness, which causes C to exhibit negative skewness. This is not surprising as $\mathfrak{F}$ is bounded below by zero, and yet has small mean but inflated variance. The third-moment provides an additional level of detail about the joint behavior of $\mathfrak{F}$ and C.

During the maxent numerical optimization, a 100×500 equispaced mesh was found sufficient for the BRCA data; however, for the HIV data the resolution had to be increased to 400×2000. This is because, in addition to the larger G, the HIV $\mathbf{X}$ also exhibits more covariance. The BRCA optimization required 30 iterations, while HIV took ~70 iterations, to converge to an estimated distribution.

Figure 3 reports the estimated $p(\mathfrak{F} \mid \mathcal{O})$, where $\mathcal{O}$ refers to the observed data. Second- and third-moment estimates are shown separately. In the framework of statistical inference such a distribution is the ultimate goal, but this result could later be used for other purposes like computing point estimates and associated confidence intervals.

If the mean of the estimated $\hat{p}(\mathfrak{F} \mid \mathcal{O})$ is used as a point estimate of $\mathfrak{F}$, then for the BRCA data, third-moment calculations suggest 104 false discoveries versus 79 for the second-moment while the usual $\hat{\mu}_{\mathfrak{F}} \approx \mathcal{E}\{\mathfrak{F}\}$ suggests only 20 false discoveries. These numbers must be put in perspective by noting that the actual z_g count falling in the left-sided tail-area is 116. For the HIV data, the third-moment analysis found eight false discoveries compared to 19 for second-moment, while the mean estimator $\hat{\mu}_{\mathfrak{F}} \approx \mathcal{E}\{\mathfrak{F}\}$ produces 48. This time the z_g count in the left-sided tail area is 46. Clearly, extensive analysis of intergene dependence can lead to very different conclusions from the same data, relative to those of the mean estimate of false discoveries (and, in turn, the procedures built around it).

The second-order estimator designed by Efron [31] found 77 false discoveries for BRCA. Efron compares that to the results of nonparametric analysis in the same paper

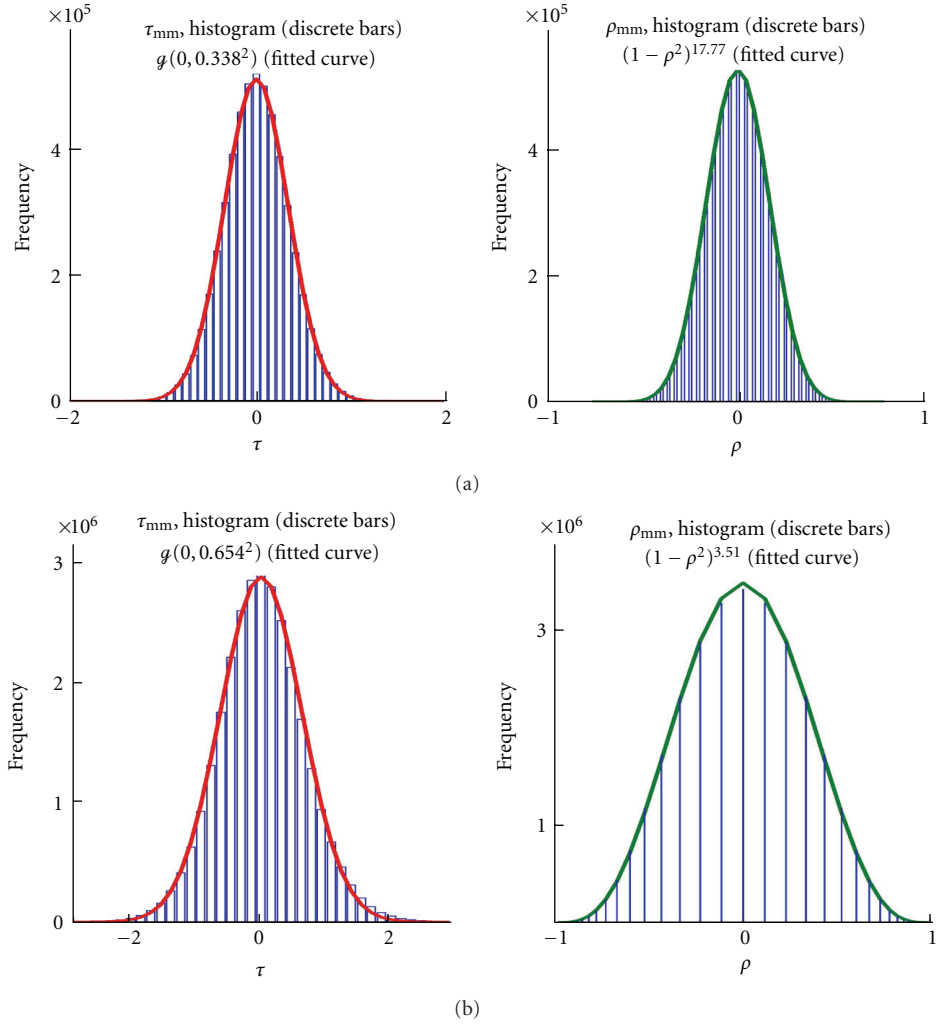


FIGURE 1: Effect of sampling fluctuations on the empirical covariance density. (a) Upper curves: BRCA data. (b) Lower curves: HIV data. For each subfigure: left panel is the histogram of sample covariances after applying the Fisher transformation (20) and a normal distribution (heavy curve) fitted to it; right panel is the histogram of denoised covariances and a modified beta distribution fitted to it (heavy curve). These distributions summarize the cumulative effect of $\binom{G}{2}$ gene-gene covariances in a single parameter (α) distribution.

and concludes underestimation, but the issue is not further pursued. Our findings show that nonlinear dependence (as reflected in the present case by moments higher than second) is potentially very important in characterizing the null z histogram.

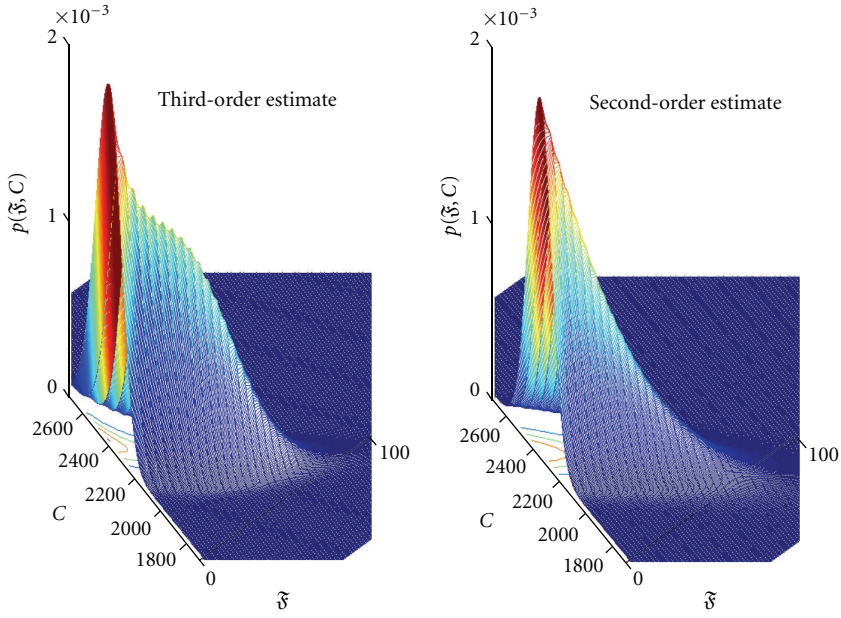
We note in passing that the availability of $\hat{p}(\mathfrak{F} \mid \mathcal{O})$ permits the application of the bound $\mathbb{P}(\mathfrak{F}/G_* \mid \mathcal{O} \geq \gamma) \leq \lambda$ as a control measure, as recommended by Lehmann and Romano [63]. It is not a trivial matter to choose γ and λ such that a fair comparison with other error measures is possible; however, for illustrative purposes, we set $\gamma = 0.15$ and $\lambda = 0.5$. With this constraint, the present approach reports 174 discoveries (108 for second-moment) for the HIV **X**. This compares favorably with the results of Efron [31], where the Benjamini-Hochberg procedure, with false discover rate control level 0.10 and an empirical null from [5], yields 180 discoveries. Without covariance modeling, the Benjamini-Hochberg procedure reports only 20 discoveries.

5.2. Testing on Simulated Data. Insight is gained by testing the approach on simulated data for which the "correct answer" is known. In the studies below, all cases are null (no treatment, residuals only). The goal is to see how well the realized left-sided tail-area count can be estimated from the center count. To maintain realism, we simulate raw mRNA levels. The testing scenario is a "two-group study," so en route to z-values we take the usual two-sample t-statistic.

Let the mRNA expression level, x_{gm}, of gene g measured by microarray m, be distributed as the *Gamma density*: For $m = 1, \ldots, M$,

$$x_{gm} \sim p_\Gamma(\xi; \kappa, \theta) = \xi^{\kappa-1} \frac{e^{-\xi/\theta}}{\theta^\kappa \Gamma(\kappa)}, \quad \xi \geq 0,\ \kappa, \theta > 0, \tag{34}$$

where $\Gamma(\kappa) = \int_0^\infty v^{\kappa-1} e^{-v} dv$ is the Gamma function. κ and θ are called the *shape parameter* and the *scale parameter* of the distribution, respectively. This distribution is similar to the Gamma-Gamma model used by Newton et al. [64].



Moment of $p(\mathfrak{F}, C)$	Estimated value
$\mu_{\mathfrak{F}}$	19.9
μ_C	2203.0
$\mathcal{E}\{(\mathfrak{F} - \mu_{\mathfrak{F}})^2\}$	289.7
$\mathcal{E}\{(C - \mu_C)^2\}$	33113.3
$\mathcal{E}\{(\mathfrak{F} - \mu_{\mathfrak{F}})(C - \mu_C)\}$	−2732.8
$\mathcal{E}\{(\mathfrak{F} - \mu_{\mathfrak{F}})^3\}$	10299.8
$\mathcal{E}\{(C - \mu_C)^3\}$	−1399069.6
$\mathcal{E}\{(\mathfrak{F} - \mu_{\mathfrak{F}})^2(C - \mu_C)\}$	−66210.9
$\mathcal{E}\{(\mathfrak{F} - \mu_{\mathfrak{F}})(C - \mu_C)^2\}$	388225.9

Figure 2: BRCA example: estimated $p(\mathfrak{F}, C)$ moments and estimated distributions using maxent, $\hat{p}(\mathfrak{F}, C)$. The distribution estimate on the left uses third-moment information in the maxent optimization, while the right estimate uses only second moments. The third-order estimate exhibits finer details than its second-order counterpart, and a contour that cannot be modeled using a quadratic distribution.

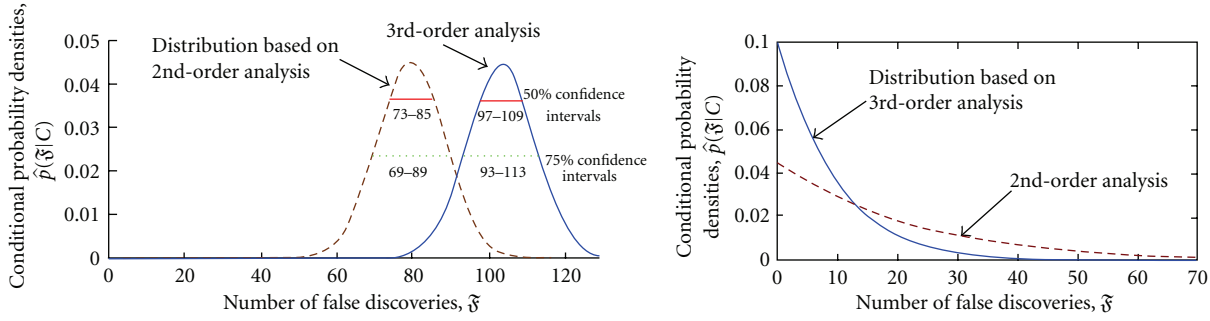


Figure 3: Estimated conditional distributions of the number of false discoveries $\hat{p}(\mathfrak{F} \mid C)$. Panel (a) BRCA data. Panel (b) HIV data. To show the effect of skewness corrections the third-moment $\mathfrak{F}$ distribution (solid curve) is compared to its second-moment counterpart (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 curves are also labeled with 50% (solid line) and 75% (dotted line) confidence intervals.

In (34) the shape parameter κ is common to all genes, while the scale parameters $\{\theta_g\}_{g=1}^{G}$ characterize varying mRNA levels from gene to gene, but are assumed i.i.d. as

$$\theta_g \overset{\text{i.i.d.}}{\sim} p_\Gamma(\xi;\kappa_0,\theta_0), \quad \text{for } g = 1,\ldots,G. \tag{35}$$

The intuition that genes with larger underlying mRNA levels would have higher variance is supported by model (34) since the mean of the gth gene is $\kappa\theta_g$ and variance is $\kappa\theta_g^2$.

The parameter set $(\kappa,\kappa_0,\theta_0)$ in (34) and (35) can be chosen on the basis of the overall gene expression histogram of real microarray data. Results for three such parameter sets: $(1, 0.6, 500)$, $(2, 0.39, 384)$, and $(3, 0.33, 300)$, are presented. These numbers were chosen to preserve the total sample variance, and the particular values are based on the HIV data of Van 't Wout et al. [62] which were collected using Affymetrix microarrays. In particular, $\kappa = 1$ models x_{gm} variables that are exponentially distributed, $\kappa = 2$ models a unimodal distribution with heavy tails and a noticeable departure from Gaussianity. Case $\kappa = 3$ represents an approximation to a Gaussian distribution, but with slightly heavier tails.

Substantial row-wise covariance was added via the Gaussian copula technique: A $(G \times M)$ matrix, say $\mathbf{Z}$, of i.i.d. unit normal RVs was used to produce a correlated matrix, $\mathbf{Z}^c$, via the mapping

$$\mathbf{Z}^c = \mathbf{L}^T\mathbf{Z}, \quad \text{where } \mathbf{Z} \text{ has elements } z_{gm} \overset{\text{i.i.d.}}{\sim} \mathcal{G}(0,1), \tag{36}$$

in which $\mathbf{L}$ is the lower-triangular Cholesky factor (e.g., [65]) of $\mathbf{R}_{z,G} + \varepsilon\mathbf{I}_G$, the correlation matrix of the actual expression matrix $\mathbf{X}$ from the BRCA study, plus a small diagonal load to prevent singularity due to the fact that $M < G$. Several other dense matrices, $\mathbf{R}$, generated through a different method [66], yielded similar results. This process imposes the covariance of the real BRCA data on the simulated substrate of independent Gaussian variables. The resulting elements z^c_{gm} were mapped to P values, $P_{\mathcal{G}_u}(z^c_{gm})$, then further transformed to simulated expression variables, x_{gm}, through the inverse Gamma c.d.f. as in (34). The result is the simulated expression matrix $\mathbf{X} = [x_{gm}]_{G\times M}$.

Figures 4 and 5 compare second- and third-moment estimates for $\delta = -2.0$ and $\delta = -2.5$, respectively. In both cases, $c = 1$ for the center area (see Section 6). For each $(\kappa,\kappa_0,\theta_0)$, 800 matrices $\mathbf{X}$ were processed. On each $\mathbf{X}$ the approach was applied in its entirety and no additional knowledge was assumed. The *a posteriori* mean was used as the final estimate. The usual mean estimator $\hat{\mu}_{\mathfrak{F}} \approx \mathcal{E}\{\mathfrak{F}\}$ consistently reported 20 for $\delta = -2.5$ and 73 for $\delta = -2.0$, regardless of the particular $\mathbf{X}$.

Strikingly, for all three parameter sets $(\kappa,\kappa_0,\theta_0)$, the third-moment skewness corrections make the estimation process more accurate. For some of the scenarios third-moment estimates saturate somewhat, but the effect is minor compared to that in the second-order approaches. To the extent that these parameter sets cover a wide range of realistic gene expression data, the practical utility of the proposed approach is evident.

6. Discussion

Advances in DNA microarray technology, improved standardization procedures, and a careful execution of laboratory protocols collectively lead to testing situations with marginally correct but strongly correlated null hypotheses. If correlation is the result of intrinsic gene-gene interactions, no experimental design can circumvent it. Correlation can cause the realized $\hat{\mathfrak{F}}$ to vary significantly from case to case [63], and the control of $\mathcal{E}\{\hat{\mathfrak{F}}\}$ via the usual $\mu_{\mathfrak{F}} = \mathcal{E}\{\mathfrak{F}\}$ may no longer represent the basic facts. The moment theory of the null statistic histogram can be used to deduce an estimator of $\mu_{\mathfrak{F}}$ which explicitly combines identifiability and covariance. Though we have explored these ideas in the differential analysis context above, the findings are quite general.

It is reasonable to question the necessity of the heavy mathematical machinery of the foregoing sections since it is possible to simulate a number of sets of z-scores $\{z_g\}_{g=1}^{G}$ from the distribution $\mathbf{z} \sim \mathcal{G}(\mathbf{0}, \mathbf{\Sigma}_z^{G\times G})$, then estimate the moments. However, due to sampling fluctuations, the underlying covariance matrix $\mathbf{\Sigma}_z^{G\times G}$ is ordinarily unattainable; however, pursuing quantities like $q_{\rho_z}(\xi)$ and $q_{\mathbf{R}_z}(\xi)$ is still possible. Also, as G gets larger ($\sim$25,000 for recent microarray studies) computational demands, as well as the large number of z-score sets required, become prohibitive.

Permutation calculations, as in [31, Section 4], offer an alternative way to estimate the moments. They too can run into computational difficulties, especially when the test statistic itself is computationally intensive. Further difficulty arises when samples are few. For a two-group study like HIV (four samples each condition), the data provide only 70 unique permutations.

When a direct extraction of inter-hypotheses covariance is not feasible, single omnibus parameter models remain useful in that the investigator can still use judgment to intelligently incorporate some form of covariance effect by setting a value of the parameter α.

The distribution of interest $p(\mathfrak{F}, C)$ resides over support domain $\mathcal{D}$ as shown in Figure 6, and the maxent algorithm is adept at handling such complicated support regions. At a more fundamental level, through maxent, we seek to minimize the amount of unintentional prior information brought into the inference.

Apart from the numerical parameters Δ (bin width) and the mesh resolution in maxent, the only open choice of parameterization in the present method is c, the center area boundary. The selection of $c = 1$ in this paper is based on the first eigenvector analysis of Efron [31] which suggests that (within certain approximations) the interval $[-1, 1]$ has completely opposite count behavior from the rest of the $\mathcal{Z}$ space.

One surprising result of this and similar studies is that more inter-z_g covariance does not translate into more extreme covariance between variables $\mathfrak{F}$ and C. In the BRCA data, for example, the coefficient between $\mathfrak{F}$ and C is -0.89, while for the HIV data the covariance drops to -0.75. Further research to gain insight into this behavior would be very useful.

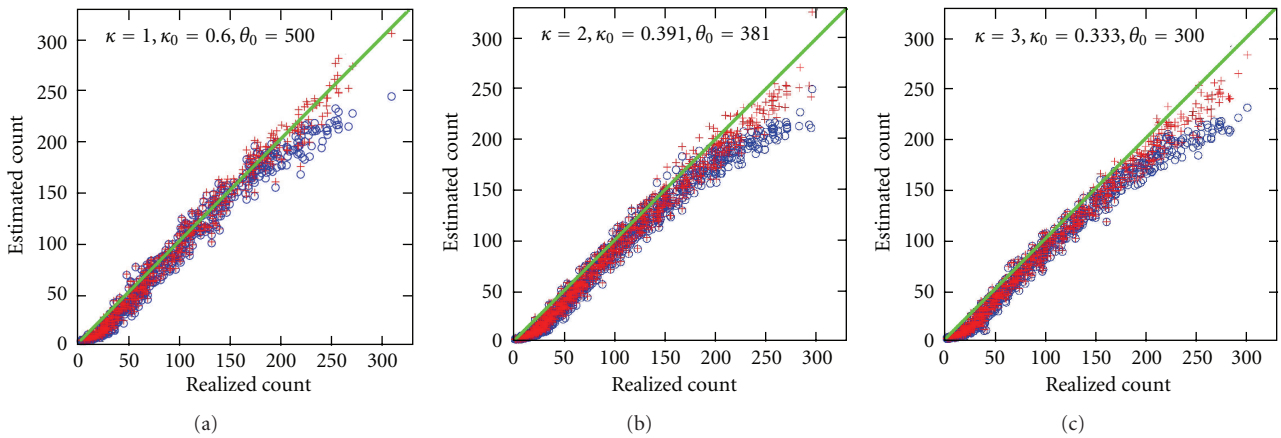


FIGURE 4: Simulation experiments comparing conditional estimates: third-moment estimates (+ marker) and second-moment (∘ marker). This figure corresponds to the left-sided tail area with $\delta = -2.0$ [see (4)]. Substantial rowwise covariance is present. The abscissa is the realized count while the ordinate is the estimated count. The significance of parameters $(\kappa, \kappa_0, \theta_0)$ is discussed the text. Third-moment skewness corrections tend to make the estimation process more accurate.

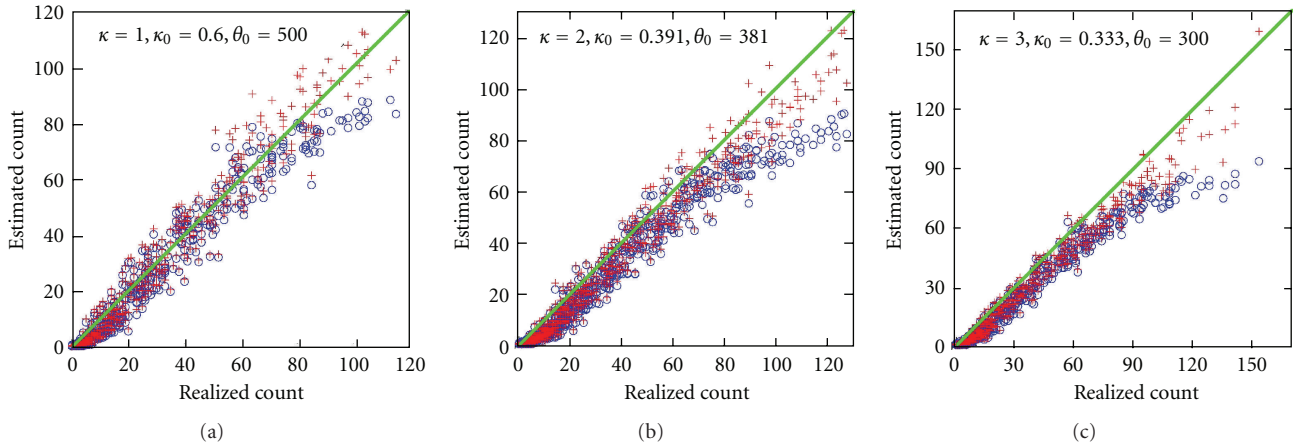


FIGURE 5: Left-sided tail area with $\delta = -2.5$. Otherwise Figure 4 caption is applicable.

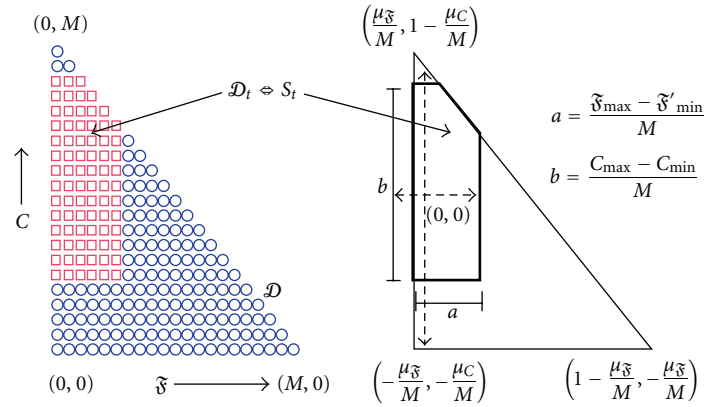


FIGURE 6: Discrete support $\mathcal{D}_t$ (□ markers) versus continuous support $\mathcal{S}_t$ (solid boundary). $\mathcal{S}_t$ is normalized to improve numerical stability.

Finally, while covariance was viewed in the present paper as a "destructive factor" in the attempt to estimate $\mathfrak{F}$, inter-z_g covariance can, in fact, be *exploited* to increase power by finding a superior ranking of potential discoveries. The recent multiple testing literature has begun to address this possibility.

Ethical Approval

Human subject data used in this study are publicly available and anonymous and are, therefore, exempted from continuing Internal Review Board scrutiny according to US Health and Human Services Policy 45 CFR 36, Subpart A, §46.101 (2.b.4).

Appendix

A. Genes and Microarrays

To understand the core statistical methods developed in this paper, it is only necessary to know some very general facts about the biological application. The "blueprint" for building the components of every cell in an organism (except some viruses) is encoded in macromolecules collectively called *deoxyribonucleic acid* or "DNA." Each DNA molecule consists of a complementary pair of long sequences (polymers) of small molecular elements known as *nucleotides*. The characteristic "double helix" configuration of DNA molecules results from chemical forces as the nucleotides on the complementary polymers strongly bind to one another. Each nucleotide is built around one of the four nitrogen bases guanine, adenine, thymine and cytosine, commonly known by their initial letters G, A, T, and C. It is the sequence of these bases that encodes the information for producing proteins that, in turn, determine the structures and functions of cells. In the human genome (the complete set of life-sustaining instructions in the genetic material of an organism), the DNA consists of about three billion nucleotides organized into 23 sets of structures called *chromosomes*.

Genes are certain identifiable sections of the DNA—varying in length from a few hundred to a few million bases—that encode a particular trait or "hereditary unit" of biological information by specifying and regulating the types of proteins produced. Almost every gene is present in the DNA of every cell of a given organism, but only a relatively few genes are active in any given cell type. Genes comprise only about 1% of the DNA material in the human geneome. The reasons for the existence of the remaining 99% of the DNA remain the subject of much conjecture, hypothesis, and research. It is currently estimated that there are on the order of 25,000 genes in the human genome, 99% of which are identical among all humans. The slight variances in the other 1% of the genes (gene *alleles*) determine differences among individuals (e.g., eye, hair, and skin color, blood type, etc.). Apart from normal variations (alleles), mutations in genes can lead to the formation of abnormal proteins with beneficial, neutral, or negative consequences for cell function and replication.

A gene is defined by its sequence of nucleotide bases, which, in turn, can be expressed as a sequence whose elements come from the set of four letters G, A, T, and C as described above—thus reducing the information carried by the gene to a simple code like "ATCGCT…". A three letter code (i.e., a three base set like "AGA"), called a *codon*, ultimately specifies one of the 20 *amino acids* that are the building blocks of proteins. There are 64 possible codons, and 61 of these are used to indicate 20 amino acids, so there is redundant representation of the amino acids in the codons. The remaining three codons are used for regulating the protein synthesis.

The manifestation of genes as proteins is called *gene expression*. The degree to which a gene is "active" in a given collection of cells (e.g., a basal cell (skin) tumor) in a given set of conditions (e.g., untreated *versus* treated with radiation or chemotherapy) can be ascertained by measuring the levels of certain molecules [messenger RNA (mRNA) or complementary DNA (cDNA)] related to to proteins manufactured by the gene. A microarray consists of thousands of binding sites ("probes"), each populated with DNA fragments that can be associated with particular genes. The extent to which a gene is expressed in a particular preparation determines the extent to which the related microarray site "lights up" as the phosphorescent mRNA or cDNA in the preparation binds to its site. A single microarray experiment can be used to simultaneously quantify the expression levels of thousands of genes in a particular tissue preparation.

The applied purpose of the statistical modeling work in this paper is to develop methods for determining from microarray data which genes are expressed at significantly different levels when a cell type is exposed to different conditions. Ultimately, this information can be used to understand normal and abnormal cell function and replication, to target genes for medical therapies, and to develop drugs for treating myriad diseases and systemic disorders at the cellular level.

B. Estimating $P(\mathfrak{F}, C)$ by Maximum Entropy Optimization (MAXENT)

B.1. MAXENT Distribution. $p(\mathfrak{F}, C)$ is a discrete distribution with support domain

$$\mathcal{D} = \{(i, j) : \; i, j \in \mathbb{Z};\; 0 \le i \le G,\; 0 \le j \le G,\; 0 \le i + j \le G\}. \tag{B.1}$$

Any moment-based inference involves computation over $\mathcal{D}$ whose increasing cardinality ($\propto G^2$) makes processing difficult for a large G. However, computation can be reduced substantially by truncating the domain $\mathcal{D}$ to the set

$$\begin{aligned} \mathcal{D}_t = \{(i, j) : {} & \mathfrak{F}_{\min} \le i \le \mathfrak{F}_{\max}, \\ & C_{\min} \le j \le C_{\max},\; 0 \le i + j \le G\}, \end{aligned} \tag{B.2}$$

where,

$$\begin{aligned}\mathfrak{F}_{\min} &= \max(\lfloor \mu_{\mathfrak{F}} - \ell\sigma_{\mathfrak{F}} \rfloor, 0),\\ \mathfrak{F}_{\max} &= \min(\lceil \mu_{\mathfrak{F}} + \ell\sigma_{\mathfrak{F}} \rceil, G),\\ C_{\min} &= \max(\lfloor \mu_C - \ell\sigma_C \rfloor, 0),\\ C_{\max} &= \min(\lceil \mu_C + \ell\sigma_C \rceil, G).\end{aligned} \tag{B.3}$$

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

Computation can be reduced further by recognizing that distributions imposed on the basis of a small number of moment constraints often enjoy a high level of regularity so that a sparser mesh should be adequate. The computation-accuracy trade-off becomes much easier to analyze if the problem is posed as one of learning a density function over a continuous domain, say

$$\begin{aligned}\mathcal{S}_t = \Big\{(v, w) :\ & v \in \Re(\mathfrak{F}),\ w \in \Re(C),\\ & -h_\mu \leq v + w \leq 1 - h_\mu\Big\},\end{aligned} \tag{B.4}$$

where, $h_\mu \stackrel{\text{def}}{=} G/(\mu_{\mathfrak{F}} + \mu_C)$,

$$\begin{aligned}\Re(\mathfrak{F}) &\stackrel{\text{def}}{=} \text{range of } \mathfrak{F}\\ &= \text{the interval } \left[\frac{G}{\mathfrak{F}_{\min} - \mu_{\mathfrak{F}}}, \frac{G}{\mathfrak{F}_{\max} - \mu_{\mathfrak{F}}}\right],\end{aligned} \tag{B.5}$$

and similarly for $\Re(C)$ (see Figure 6). Note that the continuous domain is scaled by the factor G to improve numerical stability. The moment constraints must be scaled accordingly. Let $\mathcal{P}_c$ denote the space of feasible distributions $p(\xi_1, \xi_2)$. Then, for all $p \in \mathcal{P}_c$:

$$\int_{\mathcal{S}_t} \xi_1^i \xi_2^j p(\xi_1, \xi_2) \mathrm{d}\xi_1 \mathrm{d}\xi_2 = \frac{\mathcal{E}\left\{(\mathfrak{F} - \mu_{\mathfrak{F}})^i (C - \mu_C)^j\right\}}{G^{i+j}} \stackrel{\text{def}}{=} \eta^{ij}, \tag{B.6}$$

with $0 \leq i + j \leq J$. In (B.6), $(i, j) = (0, 0)$ corresponds to the constraint $\int_{\mathcal{S}_t} p(\xi_1, \xi_2)\mathrm{d}\xi_1\mathrm{d}\xi_2 = 1$, while $(i, j) = (1, 0)$ together with $(i, j) = (0, 1)$ imply that every $p \in \mathcal{P}_c$ has mean $[0 \quad 0]^T$. We are concerned with problems for which $J \leq 3$. For convenience, we have defined the notation η^{ij} as a shorthand for the *scaled* (i, j) joint central moment of $\mathfrak{F}$ and C.

The selection of a unique $p(\xi_1, \xi_2)$ is based on the principle of entropy maximization (MAXENT) which seeks a $p \in \mathcal{P}_c$ with maximum information entropy [67]. The information entropy essentially measures the spread of the distribution, and hence, maxent can be seen as a criterion, which, within one's knowledge constraints, maximizes the representation of unknown information ("ignorance")—arguably, a suitable approach for statistical inference.

MAXENT seeks the following solution in $\mathcal{P}_c$:

$$p^*(\xi_1, \xi_2) = \max_{p \in \mathcal{P}_c} \left\{ - \int_{\mathcal{S}_t} p(\xi_1, \xi_2) \ln\{p(\xi_1, \xi_2)\} \mathrm{d}\xi_1 \mathrm{d}\xi_2 \right\}. \tag{B.7}$$

The solution takes the following exponential form:

$$p_\lambda(\xi_1, \xi_2) = \frac{\exp\left\{\sum_{1 \leq i+j \leq J} \lambda_{ij}\ \xi_1^i\ \xi_2^j\right\}}{\int_{\mathcal{S}_t} \exp\left\{\sum_{1 \leq i+j \leq J} \lambda_{ij}\, \xi_1^i\, \xi_2^j\right\} \mathrm{d}\xi_1 \mathrm{d}\xi_2}, \tag{B.8}$$

in which the λ_{ij} are Lagrange multipliers. The derivation of the exponential form (B.8) and a procedure to determine optimal multipliers λ_{ij} are given in the following subsection.

B.2. MAXENT Solution Details. The information entropy functional is concave [68], and the constraints in (B.6) are linear in $p(\xi_1, \xi_2)$. Thus, the problem in (B.7) is a convex program that be solved in a Lagrangian dual framework, where one works with an unconstrained upper bound that is easy to optimize (e.g., [69]). More importantly, in the present case, the framework allows the conversion of the original infinite-dimension problem of functional variation into a finite-dimension problem with as few variables as the number of constraints.

Lemma 4. *The dual, $\Psi(\boldsymbol{\lambda})$, of the concave optimization problem* (B.7) *is given by:*

$$\Psi(\boldsymbol{\lambda}) = \ln\left[\int_{\mathcal{S}_t} \exp\left\{\sum_{1 \leq i+j \leq J} \lambda_{ij} \xi_1^i \xi_2^j\right\} d\xi_1 d\xi_2\right] - \sum_{2 \leq i+j \leq J} \lambda_{ij} \eta^{ij}, \tag{B.9}$$

where $\boldsymbol{\lambda}$ is the set of Lagrange multipliers, $\{\lambda_{ij}\}_{i,j}$; λ_{ij} is the multiplier corresponding to the (i, j)th constraint; and i, j, J, and η^{ij} are defined in (B.6).

Proof. By the definition of the Lagrangian dual function

$$\begin{aligned}\Psi(\boldsymbol{\lambda}) = \sup_{p \in \mathcal{P}_c} \Bigg\{ & - \int_{\mathcal{S}_t} p(\xi_1, \xi_2) \ln\{p(\xi_1, \xi_2)\} \mathrm{d}\xi_1 \mathrm{d}\xi_2\\ & + \sum_{i+j \leq J} \lambda_{ij} \left(\int_{\mathcal{S}_t} \xi_1^i \xi_2^j p(\xi_1, \xi_2) \mathrm{d}\xi_1 \mathrm{d}\xi_2 - \eta^{ij} \right) \Bigg\}.\end{aligned} \tag{B.10}$$

Taking the functional variation of the bracketed term in (B.10) with respect to the unknown density $p(\xi_1, \xi_2)$, and using the fact that $\int_{\mathcal{S}_t} p(\xi_1, \xi_2)\, \mathrm{d}\xi_1\, \mathrm{d}\xi_2 = 1$, we obtain the maximizer of (B.10)

$$p_\lambda(\xi_1, \xi_2) = \frac{\exp\left\{\sum_{1 \leq i+j \leq J} \lambda_{ij} \xi_1^i \xi_2^j\right\}}{\int_{\mathcal{S}_t} \exp\left\{\sum_{1 \leq i+j \leq J} \lambda_{ij} \xi_1^i \xi_2^j\right\}\, \mathrm{d}\xi_1\, \mathrm{d}\xi_2}. \tag{B.11}$$

Inserting (B.11) into (B.10) yields

$$\begin{aligned}\Psi(\boldsymbol{\lambda}) = &\int_{\mathcal{S}_t} p(\xi_1', \xi_2') \\ &\times \ln\left\{\int_{\mathcal{S}_t} \exp\left(\sum_{1\le i+j\le J} \lambda_{ij}\xi_1^i\xi_2^j\right)\mathrm{d}\xi_1\mathrm{d}\xi_2\right\}\mathrm{d}\xi_{1'}\mathrm{d}\xi_{2'} \\ &- \sum_{1\le i+j\le J} \lambda_{ij}\eta^{ij} \\ = &\ln\left\{\int_{\mathcal{S}_t} \exp\left(\sum_{1\le i+j\le J} \lambda_{ij}\xi_1^i\xi_2^j\right)\mathrm{d}\xi_1\mathrm{d}\xi_2\right\} \\ &- \sum_{2\le i+j\le J} \lambda_{ij}\eta^{ij},\end{aligned} \tag{B.12}$$

where we have used the facts $\int_{\mathcal{S}_t} p(\xi_1, \xi_2)\mathrm{d}\xi_1\,\mathrm{d}\xi_2 - \eta^{00} = 0$, $\eta^{10} = 0$, and $\eta^{01} = 0$ from (B.6). □

It is easy to verify that the Hessian, denoted $\mathcal{H}\{\cdot\}$, of (B.9) is positive definite and hence $\Psi(\boldsymbol{\lambda})$ is convex. Suppose that $\boldsymbol{\lambda}^*$ is the minimum of $\Psi(\boldsymbol{\lambda})$. Then the corresponding *primal solution* $p_{\lambda^*}(\xi_1, \xi_2)$—obtained via (B.11)—indeed maximizes (B.7). To verify this, let $p_o(\xi_1, \xi_2)$ be the maximizer of (B.7). Then, from (B.10), $\Psi(\boldsymbol{\lambda}) \ge \mathcal{H}\{p_o(\xi_1, \xi_2)\}$ for all $\boldsymbol{\lambda}$. Now from general optimization theory, the functional variation of the Lagrangian with respect to $p(\xi_1, \xi_2)$ evaluated at $p_o(\xi_1, \xi_2)$ must be zero, which implies that $p_o(\xi_1, \xi_2)$ can be written in the form (B.11) for some $\boldsymbol{\lambda}_o$. But, then, $\Psi(\boldsymbol{\lambda}^*) \le \Psi(\boldsymbol{\lambda}_o) \Rightarrow \Psi(\boldsymbol{\lambda}^*) \le \mathcal{H}\{p_o(\xi_1, \xi_2)\}$. Consequently, $\Psi(\boldsymbol{\lambda}^*) = \mathcal{H}\{p_o(\xi_1, \xi_2)\}$, so that $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}_o$. Hence $p_{\lambda^*}(\xi_1, \xi_2)$ is the maximizer of (B.7).

Newton's method (e.g., [70, Section 4.6]) specifies that if a multivariable function $\Psi(\boldsymbol{\lambda})$ is twice differentiable and the initial value $\boldsymbol{\lambda}_0$ is chosen close enough to the optimal $\boldsymbol{\lambda}^*$, then the sequence over indices $t = 0, 1, 2, \ldots$,

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \gamma\,\frac{\Delta\Psi(\boldsymbol{\lambda}_t)}{\mathcal{H}\{\Psi(\boldsymbol{\lambda}_t)\}} \tag{B.13}$$

converges to $\boldsymbol{\lambda}^*$. In (B.13), $\Delta\Psi(\boldsymbol{\lambda}_t)$ denotes the gradient of $\Psi(\boldsymbol{\lambda})$ evaluated at $\boldsymbol{\lambda}_t$. The parameter $\gamma > 0$ allows a finer control of step sizes to avoid numerical instabilities. Intuitively, at the tth iteration, $\Psi(\boldsymbol{\lambda})$ is replaced by its second-order Taylor expansion around $\boldsymbol{\lambda}_t$ and then minimized exactly, which produces the minimum $\boldsymbol{\lambda}_{t+1}$. At the $(t+1)$st iteration, $\boldsymbol{\lambda}_{t+1}$ becomes the point of expansion and the method continues until the desired convergence level is achieved.

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

$$\begin{aligned}&\frac{\partial\Psi\left(\check{\boldsymbol{\lambda}}\right)}{\partial\lambda_{ij}} \\ &= \int_{\mathcal{S}_t} \xi_1^i\xi_2^j \\ &\quad\times\left[\frac{\exp\left\{\sum_{1\le i+j\le J}\check{\lambda}_{ij}\xi_1^i\xi_2^j\right\}}{\int_{\mathcal{S}_t}\exp\left\{\sum_{1\le i+j\le J}\check{\lambda}_{ij}\xi_1^i\xi_2^j\right\}\,\mathrm{d}\xi_1\mathrm{d}\xi_2}\right]\times\mathrm{d}\xi_1\,\mathrm{d}\xi_2 - \eta^{ij} \\ &= \int_{\mathcal{S}_t}\xi_1^i\xi_2^j\check{p}(\xi_1,\xi_2)\mathrm{d}\xi_1\mathrm{d}\xi_2 - \eta^{ij} = \check{\eta}^{ij} - \eta^{ij},\end{aligned} \tag{B.14}$$

where $\check{\eta}^{ij}$ denotes the (i, j) central moment of the distribution of (B.11) parameterized by $\check{\boldsymbol{\lambda}}$. Similarly, the elements of the Hessian are given by:

$$\begin{aligned}\frac{\partial^2\Psi\left(\check{\boldsymbol{\lambda}}\right)}{\partial\lambda_{i'j'}\partial\lambda_{ij}} &= \int_{\mathcal{S}_t}\xi_1^{i+i'}\xi_2{}^{j+j'}\,\check{p}(\xi_1,\xi_2)\,\mathrm{d}\xi_1\mathrm{d}\xi_2 \\ &\quad - \int_{\mathcal{S}_t}\xi_1^{i'}\xi_2^{j'}\,\check{p}(\xi_1,\xi_2)\,\mathrm{d}\xi_1\mathrm{d}\xi_2 \\ &\quad\times\int_{\mathcal{S}_t}\xi_1^i\xi_2^j\check{p}(\xi_1,\xi_2)\,\mathrm{d}\xi_1\mathrm{d}\xi_2 \\ &= \check{\eta}^{(i+i')(j+j')} - \check{\eta}^{i'j'}\check{\eta}^{ij}.\end{aligned} \tag{B.15}$$

The gradient calculations that occur as a part of the Hessian essentially involve integration over the planar domain $\mathcal{S}_t$. Many advanced techniques for implementing numerical integration on a quadrangle like $\mathcal{S}_t$ are available in the literature; however, an equispaced rectangular mesh is found to be sufficient for the present purpose. We initiate the sequence (B.13) with $\boldsymbol{\lambda} = 0$ which implies a uniform distribution over $\mathcal{S}_t$.

Acknowledgments

This work was supported in part by the Quantitative Biology Initiative at Michigan State University. H. Wang received support from the National Science Foundation of China (Grant no. 61002003) and from the Zhejiang Provincial Natural Science Foundation of China (Grant no. Z1111051). The authors are grateful to the personnel in the MSU High Performance Computing Center for assistance in implementing the extensive simulations, and to Dr. K.H. Desai for many contributions to this work.

References

[1] National Genome Research Institute, U.S. National Institutes of Health, http://www.genome.gov/.

[2] G. P. Page, S. O. Zakharkin, K. Kim, T. Mehta, L. Chen, and K. Zhang, "Microarray analysis," *Methods in Molecular Biology*, vol. 404, pp. 409–430, 2007.

[3] J. Wang, "Computational biology of genome expression and regulation—a review of microarray bioinformatics," *Journal of*

Environmental Pathology, Toxicology and Oncology, vol. 27, no. 3, pp. 157–179, 2008.

[4] B. Efron, "Size, power and false discovery rates," *Annals of Statistics*, vol. 35, no. 4, pp. 1351–1377, 2007.

[5] B. Efron, "Large-scale simultaneous hypothesis testing: the choice of a null hypothesis," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 96–104, 2004.

[6] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, no. 5235, pp. 467–470, 1995.

[7] S. Datta and S. Datta, "Evaluation of clustering algorithms for gene expression data," *BMC Bioinformatics*, vol. 7, no. 4, article S17, 2006.

[8] F. D. Gibbons and F. P. Roth, "Judging the quality of gene expression-based clustering methods using gene annotation," *Genome Research*, vol. 12, no. 10, pp. 1574–1581, 2002.

[9] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, 2005.

[10] M. Sémon and L. Duret, "Evolutionary origin and maintenance of coexpressed gene clusters in mammals," *Molecular Biology and Evolution*, vol. 23, no. 9, pp. 1715–1723, 2006.

[11] K. C. Li, "Genome-wide coexpression dynamics: theory and application," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 26, pp. 16875–16880, 2002.

[12] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, "Spectral biclustering of microarray data: coclustering genes and conditions," *Genome Research*, vol. 13, no. 4, pp. 703–716, 2003.

[13] C. A. Tsai, T. C. Lee, I. C. Ho, U. C. Yang, C. H. Chen, and J. J. Chen, "Multi-class clustering and prediction in the analysis of microarray data," *Mathematical Biosciences*, vol. 193, no. 1, pp. 79–100, 2005.

[14] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[15] Y. Choi and C. Kendziorski, "Statistical methods for gene set co-expression analysis," *Bioinformatics*, vol. 25, no. 21, pp. 2780–2786, 2009.

[16] W. Barry, A. Nobel, and F. Wright, "A statistical framework for testing functional categories in microarray data," *Annals of Applied Statistics*, vol. 2, pp. 286–315, 2008.

[17] T. Peters, D. W. Bulger, T.-H. Loi, J. Y. H. Yang, and D. Ma, "Two-step cross-entropy feature selection for microarrays-power through complementarity," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 4, pp. 1148–1151, 2011.

[18] F. Yang and K. Z. Mao, "Robust feature selection for microarray data based on multicriterion fusion," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 4, pp. 1080–1092, 2011.

[19] G. Tiño, H. Zhao, and H. Yan, "Searching for coexpressed genes in three-color cDNA microarray data using a probabilistic model-based hough transform," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 4, pp. 1093–1107, 2011.

[20] J. Ruan, A. K. Dean, and W. Zhang, "A general co-expression network-based approach to gene expression analysis: comparison and applications," *BMC Systems Biology*, vol. 4, article 8, 2010.

[21] M. Dettling, E. Gabrielson, and G. Parmigiani, "Searching for differentially expressed gene combinations," *Genome Biology*, vol. 6, no. 10, article R88, 2005.

[22] Y. Lai, B. Wu, L. Chen, and H. Zhao, "A statistical method for identifying differential gene-gene co-expression patterns," *Bioinformatics*, vol. 20, no. 17, pp. 3146–3155, 2004.

[23] Y. K. Ng, W. Wu, and L. Zhang, "Positive correlation between gene coexpression and positional clustering in the zebrafish genome," *BMC Genomics*, vol. 10, article 42, 2009.

[24] I. Bernthaler, A. Mühlberger, R. Fechete, P. Perco, A. Lukas, and B. Mayer, "Interpreting microarray experiments via Co-expressed Gene Groups Analysis (CGGA)," in *Proceedings of the 9th International Conference on Discovery Science*, vol. 4265 of *Lecture Notes in Computer Science*, pp. 316–320, Springer, 2006.

[25] A. Reverter and E. K. F. Chan, "Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks," *Bioinformatics*, vol. 24, no. 21, pp. 2491–2497, 2008.

[26] A. B. Tchagang and A. H. Tewfik, "DNA microarray data analysis: a novel biclustering algorithm approach," *Eurasip Journal on Applied Signal Processing*, vol. 2006, Article ID 59809, 12 pages, 2006.

[27] R. Martinez, N. Pasquier, C. Pasquier, and L. Lopez-Perez, "A dependency graph approach for the analysis of differential gene expression profiles," *Molecular BioSystems*, vol. 5, no. 12, pp. 1720–1731, 2009.

[28] C. A. Tsai and J. J. Chen, "Multivariate analysis of variance test for gene set analysis," *Bioinformatics*, vol. 25, no. 7, pp. 897–903, 2009.

[29] W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 18, no. 4, pp. 546–554, 2002.

[30] A. B. Owen, "Variance of the number of false discoveries," *Journal of the Royal Statistical Society. Series B*, vol. 67, no. 3, pp. 411–426, 2005.

[31] B. Efron, "Correlation and large-scale simultaneous significance testing," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 93–103, 2007.

[32] Y. Pawitan, K. R. K. Murthy, S. Michiels, and A. Ploner, "Bias in the estimation of false discovery rate in microarray studies," *Bioinformatics*, vol. 21, no. 20, pp. 3865–3872, 2005.

[33] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS Genetics*, vol. 3, no. 9, article e161, 2007.

[34] S. A. Degrelle, C. Hennequet-Antier, H. Chiapello et al., "Amplification biases: possible differences among deviating gene expressions," *BMC Genomics*, vol. 9, article 46, 2008.

[35] 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*, vol. 4, no. 1, article 34, 2005.

[36] X. Qiu and A. Yakovlev, "Some comments of instability of false discovery rate estimation," *Journal of Bioinformatics and Computational Biology*, vol. 4, no. 5, pp. 1057–1068, 2006.

[37] 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*, vol. 8, no. 2, pp. 414–432, 2007.

[38] R. Tibshirani and L. Wasserman, "Correlation-sharing for detection of differential gene expression," preprint, http://arxiv.org/abs/math/0608061.

[39] R. Hu, X. Qiu, and G. Glazko, "A new gene selection procedure based on the covariance distance," *Bioinformatics*, vol. 26, no. 3, pp. 348–354, 2010.

[40] Q. Cui, B. Liu, T. Jiang, and S. Ma, "Characterizing the dynamic connectivity between genes by variable parameter regression and Kalman filtering based on temporal gene expression data," *Bioinformatics*, vol. 21, no. 8, pp. 1538–1541, 2005.

[41] V. Martyanov and R. H. Gross, "Identifying functional relationships within sets of co-expressed genes by combining upstream regulatory motif analysis and gene expression information," *BMC Genomics*, vol. 11, no. 2, article S8, 2010.

[42] R. Tewhey, V. Bansal, A. Torkamani, E. J. Topol, and N. J. Schork, "The importance of phase information for human genomics," *Nature Reviews Genetics*, vol. 12, no. 3, pp. 215–223, 2011.

[43] Z. Xiang, Z. S. Qin, and Y. He, "CRCView: a web server for analyzing and visualizing microarray gene expression data using model-based clustering," *Bioinformatics*, vol. 23, no. 14, pp. 1843–1845, 2007.

[44] A. K. C. Wong, W. H. Au, and K. C. C. Chan, "Discovering high-order patterns of gene expression levels," *Journal of Computational Biology*, vol. 15, no. 6, pp. 625–637, 2008.

[45] S. Bandyopadhyay and M. Bhattacharyya, "A biologically inspired measure for co-expression analysis," *IEEE Transactions Computational Biology and Bioinformatics*, vol. 8, pp. 929–942, 2011.

[46] L. Dalton, V. Ballarin, and M. Brun, "Clustering algorithms: on learning, validation, performance, and applications to genomics," *Current Genomics*, vol. 10, no. 6, pp. 430–445, 2009.

[47] N. Ancona, R. Maglietta, A. Piepoli et al., "On the statistical assessment of classifiers using DNA microarray data," *BMC Bioinformatics*, vol. 7, article 387, 2006.

[48] T. Hu, H. Peng, and W. Sun, "Incorporating nonlinear relationships in microarray missing value imputation," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 3, pp. 723–731, 2011.

[49] H. M. Hsueh, C. A. Tsai, and J. J. Chen, "Incorporating the number of true null hypotheses to improve power in multiple testing: application to gene microarray data," *Journal of Statistical Computation and Simulation*, vol. 77, no. 9, pp. 757–767, 2007.

[50] M. Langaas, B. H. Lindqvist, and E. Ferkingstad, "Estimating the proportion of true null hypotheses, with application to DNA microarray data," *Journal of the Royal Statistical Society. Series B*, vol. 67, no. 4, pp. 555–572, 2005.

[51] 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*, vol. 66, no. 1, pp. 187–205, 2004.

[52] M. Waterman and D. Whiteman, "Estimation of probability densities by empirical density functions," *Journal of Mathematical Education in Science and Technology*, vol. 9, no. 2, pp. 127–137, 1978.

[53] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.

[54] X. Qiu, A. I. Brooks, L. Klebanov, and A. Yakovlev, "The effects of normalization on the correlation structure of microarray data," *BMC Bioinformatics*, vol. 6, article 120, 2005.

[55] H. Hotelling, "New light on the correlation coefficient and its transforms," *Journal of the Royal Statistical Society*, vol. 15, no. 2, pp. 193–232, 1953.

[56] J. Wishart, "The generalised product moment distribution in samples from a normal multivariate population," *Biometrika*, vol. 20, pp. 32–52, 1928.

[57] J. Barnard, R. McCulloch, and X. L. Meng, "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage," *Statistica Sinica*, vol. 10, no. 4, pp. 1281–1311, 2000.

[58] I. Olkin, "Note on 'The Jacobians of certain matrix transformations useful in multivariate analysis'," *Biometrika*, vol. 40, no. 1-2, p. 43, 1953.

[59] J. Liechty, M. Liechty, and P. Muller, "Bayesian correlation estimation," *Biometrika*, vol. 91, no. 1, pp. 1–14, 2004.

[60] I. Hedenfalk, D. Duggan, Y. Chen et al., "Gene-expression profiles in hereditary breast cancer," *New England Journal of Medicine*, vol. 344, no. 8, pp. 539–548, 2001.

[61] C. A. Tsai, Y. J. Chen, and J. J. Chen, "Testing for differentially expressed genes with microarray data," *Nucleic Acids Research*, vol. 31, no. 9, article e52, 2003.

[62] 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*, vol. 77, no. 2, pp. 1392–1402, 2003.

[63] E. L. Lehmann and J. P. Romano, "Generalizations of the familywise error rate," *Annals of Statistics*, vol. 33, no. 3, pp. 1138–1154, 2005.

[64] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui, "On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data," *Journal of Computational Biology*, vol. 8, no. 1, pp. 37–52, 2001.

[65] G. Golub and C. Van Loan, *Matrix Computations*, Johns-Hopkins University Press, 3rd edition, 1996.

[66] H. Qi and D. Sun, "A quadratically convergent Newton method for computing the nearest correlation matrix," *SIAM Journal on Matrix Analysis and Applications*, vol. 28, no. 2, pp. 360–385, 2006.

[67] E. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[68] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, New York, NY, USA, 1991.

[69] I. Gelfand and S. Fomin, *Calculus of Variations [English translation]*, translated by R. A. Silverman, Dover Press, New York, NY, USA, 2000.

[70] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Upper Saddle River, NJ, USA, 4th edition, 2002.