Correlation and Large-Scale Simultaneous Significance Testing

Jinxi Liu

November 1, 2017

Introduction

This article concerns the effect of correlation on multiple testing procedures, particularly false discovery rate techniques.

Introduction

The experiments report two-sample t-statistics t_i comparing expression levels under two different conditions for N genes, $N = 3226$ for the breast cancer study, and $N = 7680$ for the HIV experiment; t_i tests the null hypothesis that gene i has the same expression distribution under both conditions.

The $t_i's$ have been converted to z-values for easy analysis.

$$Z_i = \Phi^{-1}(G_0(t_i)), \qquad i = 1, 2, ..., N, \tag{1}$$

where G_0 is a putative null cdf for the t-values. G_0 was taken to be a standard Student t cdf with appropriate degrees of freedom for the HIV study, whereas a permutation method provided G_0 for the breast cancer experiment (also nearly a Student t cdf).

Assuming that G_0 is the correct null distribution for t_i, transformation (1) yields for the null cases,

$$Z_i \sim N(0, 1) \tag{2}$$

called the theoretical null.

Introduction

Microarray experiments involving genome-wide scans usually presuppose most of the genes to be null, the goal being to identify a small subset of interesting nonnull genes for future study, so we expect $N(0,1)$ to fit the center of the z-value histogram.

This is not the case in Figure (1), where $N(0,1)$ is too narrow for the breast cancer histogram and too wide for the HIV data.

Introduction

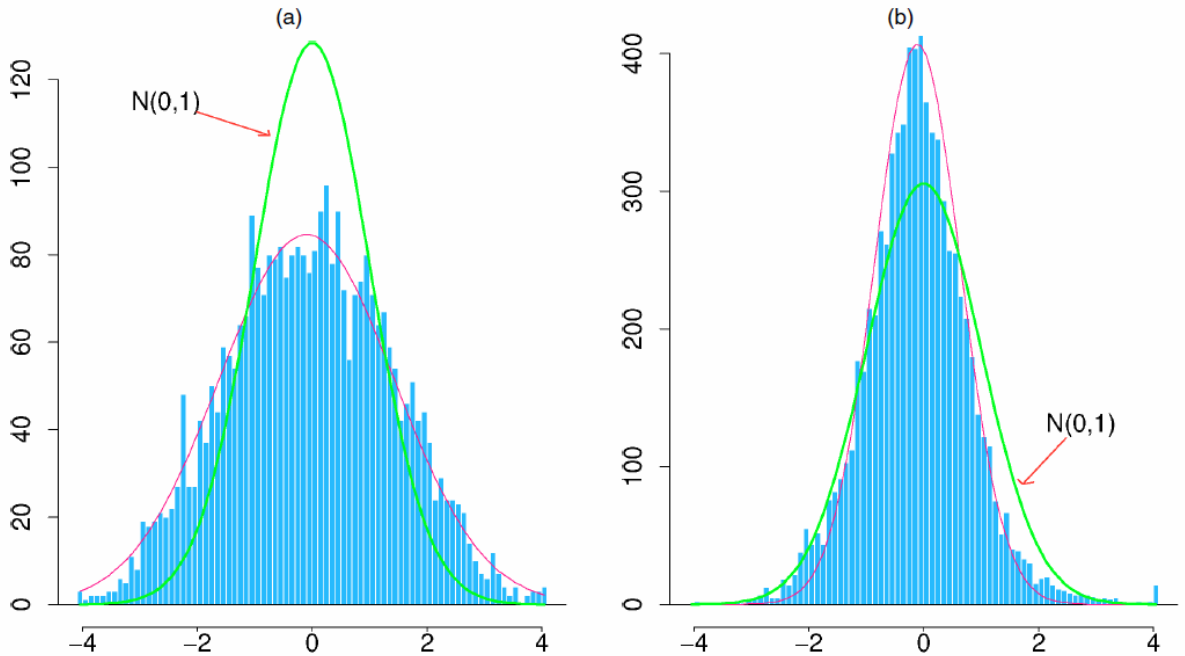

Figure 1: Histograms of z-Values From Two Microarray Experiments. (a) Breast cancer study, 3,226 genes. (b) HIV study, 7,680 genes. Heavy curves indicate N(0, 1) theoretical null densities; Light curves indicate empirical null densities fit to central z-values, as done by Efron (2004). The theoretical null distributions are too narrow in (a) and too wide in (b). Both effects can be caused by correlations among the null z-values.

Introduction

Correlation can cause effects like those shown in Figure 1, considerably widening or narrowing the distribution of the null z-values.

These effects have a substantial impact on simultaneous significance testing and must be accounted for in deciding which cases should be reported as nonnull.

Broadly speaking, a wide central histogram like that for the breast cancer data implies more null z-values in the tails, so that significance levels judged according to the theoretical $N(0, 1)$ null are too liberal. Conversely, the theoretical null is too conservative for the HIV data.

Introduction

A surprisingly simple result emerges in which the main effect of all the pairwise correlations (several million of them for Fig. 1âĂŹs examples) is summarized by a single dispersion variate, A:

a positive value of A widens the central peak of the z-value histogram, even assuming that the theoretical null (2) is individually correct for all the cases, whereas negative A narrows it, as in Figure 1(b).

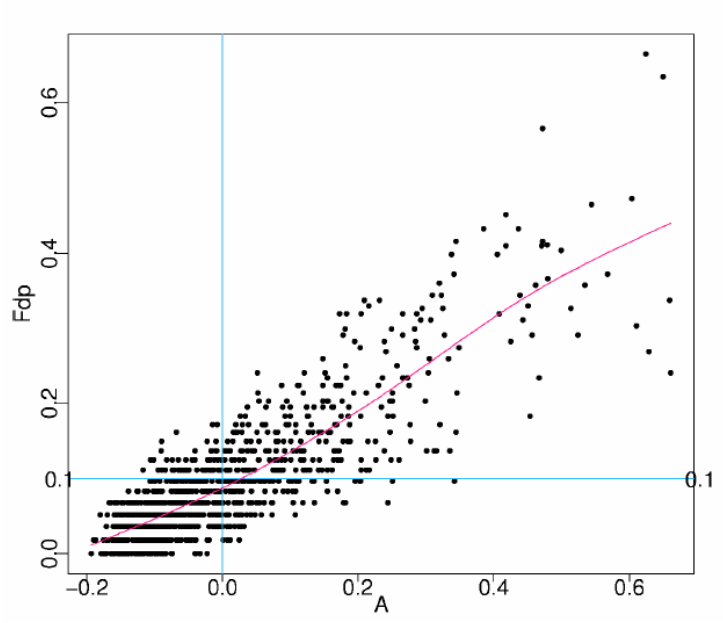

Figure 2: BenjaminiâĂŞHochberg FDR Controlling Procedure, $q = .10$,Run for $1,000$ Simulation Trials; True False Discovery Proportion, FDP, Plotted versus Dispersion Variate A. Overall FDP averaged .096, close to q, but with a strong dependence on A, as shown by smooth regression curve.

Introduction

We see a strong dependence on the dispersion variate A; FDP averaged .34 for the upper 5% of A realizations but only .03 for the lower 5%.

A can be estimated from the width of the central peak of the z-value histogram; narrow peaks imply negative $A's$ and wide peaks imply positive $A's$.

In terms of Figure 2, this enables the statistician to condition the FDR estimate on its approximate location along the A axis.

Introduction

Correlation effects on the null distribution

For following calculations, it is assumed that all cases are null,

$$Z_i \sim N(0,1), \qquad for\ i = 1, 2, ..., N \tag{3}$$

It is helpful to work with histogram counts rather than with the vector of z-values itself. Each histogram in Figure 1 has its z-axis partitioned into $K = 82$ bins of width $\Delta = .1$, running from âĹŠ4.1*to*4.1.

Correlation effects on the null distribution

We denote the count vector by y,

$$y_k = \#\{z_i \ in \ kth \ bin\}, \qquad k = 1, 2, ..., K \tag{4}$$

The histogram counts y_k arise from a partition of Z into K bins,

$$Z = \bigcup_{k=1}^{K} z_k, \tag{5}$$

each bin being of width Δ, with center point "$z[k]$"

Correlation effects on the null distribution

The following definitions lead to useful representations for the mean and covariance of y.

$$\pi_k(i) = Pr\{z_i \in Z_k\}, \qquad \pi_{k\cdot} = \frac{\sum_{i=1}^{N} \pi_k(i)}{N} \tag{6}$$

$$\gamma_{kl}(i,j) = Pr\{z_i \in Z_k \text{ and } z_j \in Z_l\}, \qquad \gamma_{kl\cdot} = \frac{\sum_{i \neq j} \gamma_{kl}(i,j)}{N(N-1)}. \tag{7}$$

Correlation effects on the null distribution

Because of assumption (3), all of the $\pi_k(i)$ are determined by $\varphi(z)$, the standard normal density, with Taylor approximation around centerpoint $z[k]$,

$$\pi_{k\cdot} = \pi_k(i) \doteq \Delta \cdot \varphi(z[k]) \tag{8}$$

The expectation vector, $\mathbf{v} = (v_1, v_2, ..., v_K)'$ of y is determined by (3),

$$\mathbf{v} = N\boldsymbol{\pi}_{\cdot} \doteq (..., N\Delta\varphi(z[k]), ...)'. \tag{9}$$

Correlation effects on the null distribution

Lemma 1

$$cov(\boldsymbol{y}) = C_0 + C_1 \tag{10}$$

$$C_0 = diag(\boldsymbol{v}) - \boldsymbol{v}\boldsymbol{v}'/N = N[diag(\boldsymbol{\pi}.) - \boldsymbol{\pi}.\boldsymbol{\pi}.'] \tag{11}$$

$$C_1 = (1 - \frac{1}{N})diag(\boldsymbol{v})\boldsymbol{\delta}diag(\boldsymbol{v}), \quad with\ \delta_{kl} = \gamma_{kl.}/\pi_{k.}\pi_{l.} - 1 \tag{12}$$

Correlation effects on the null distribution

If z_i and z_j are independent, then$\gamma_{kl}(i,j) = \pi_k(i)\pi_j(l)$. Independence of all z-values implies that all of the elements of matrix $\boldsymbol{\delta}$ in (12) are 0, leaving $cov(y) = C_0$.

Conversely, the amount of correlation between the z-values determines the size of $\boldsymbol{\delta}$ and the increase of $cov(y)$ above C_0.

Correlation effects on the null distribution

To approximate $\boldsymbol{\delta}$, we add the assumption of bivariate normality for any pair of z-values, $cov(z_i, z_j) \equiv \rho_{ij}$.

Let $g(\rho)$ indicate the empirical density of the $N(N-1)$ correlations ρ_{ij}

Lemma 2 Under the bivariate normal approximation, the matrix $\boldsymbol{\delta}$ has elements

$$\delta_{kl} \doteq \int_{-1}^{1}\int_{-1}^{1}[\frac{1}{\sqrt{1-\rho^2}}exp(\frac{\rho}{2(1-\rho^2)} \times \{2z[k]z[l]- \rho(z[k]^2 + z[l]^2)\}) - 1]g(\rho)d\rho \tag{13}$$

Correlation effects on the null distribution

Application of Lemmas 1 and 2 reqires estimation of the correlation density $g(\rho)$, which we can obtain from observed correlations between the rows of X.

$$\textit{breastcancer}: \; g(\rho) \dot{\sim} N(0, .153^2) \tag{14}$$

$$\textit{HIV}: \; g(\rho) \dot{\sim} N(0, .42^2) \tag{15}$$

Approximation (14) indicates a substantial amount of global correlation among genes in the breast cancer study, and even more correlation for the HIV study.

First Eigenvector

Lemmas 1 and 2 decompose the covariance matrix of the count vector y into an independence term C_0 and an additional term C_1 that accounts for correlation among the z-values.

This section presents a simple approximation to C_1 in terms of its first eigenvector,

First Eigenvector

Lemma 3 Suppose that $g(\rho)$, the correlation density in (13) has mean 0 and standard deviation

$$\alpha = \left[\int_{-1}^{1} \rho^2 g(\rho) d\rho\right]^{-1/2} \tag{16}$$

Then the matrix $\boldsymbol{\delta}$ in (12) is approximated by the outer product

$$\boldsymbol{\delta} \doteq \alpha^2 \boldsymbol{q}\boldsymbol{q}', \text{ where } q_k = \frac{z[k]^2 - 1}{\sqrt{2}} \tag{17}$$

First Eigenvector

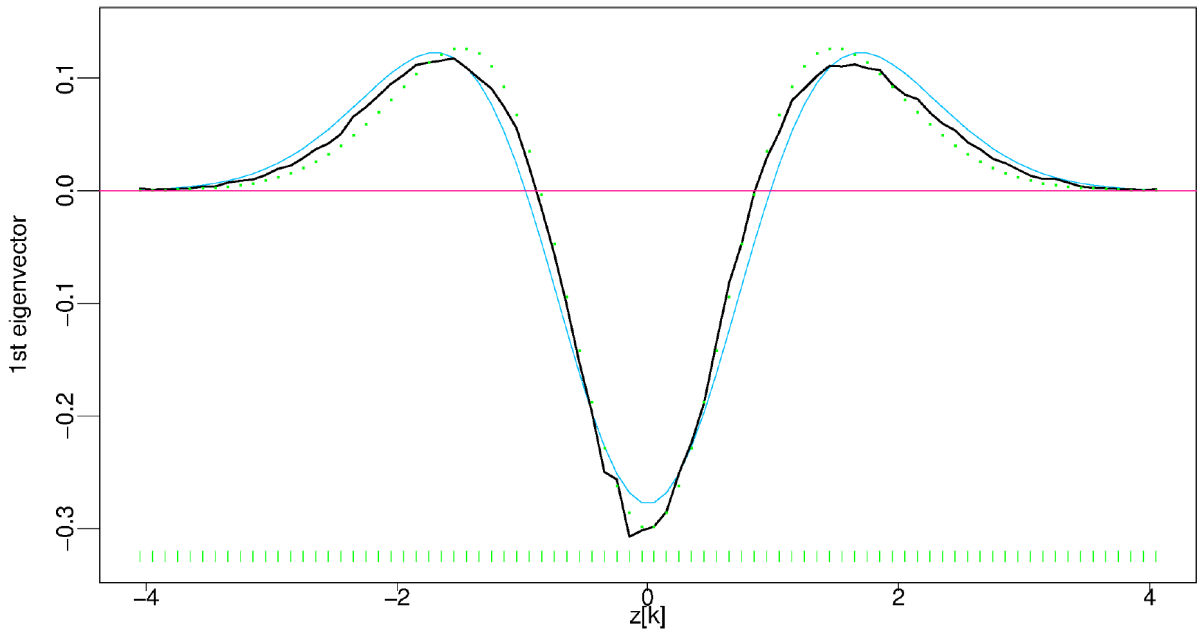

Figure 3: First Eigenvectors of Three Different Estimates of $cov(y)$: Normal-Theory Estimate for Breast Cancer Data (smooth curve); Normal-Theory Estimate for HIV Data (dots), and Permutation Estimate for Breast Cancer Data (jagged curve). The striking 'wing-shaped' form is proportional to the second derivative of the standard normal density.

First Eigenvector

Combining the three lemmas yields a useful approximation for the null covariance matrix of the count vector $\boldsymbol{y}$ under the bivariate normal assumptions.

Theorem If $g(\rho)$ has mean 0 and standard deviation α, then

$$cov(\boldsymbol{y}) \doteq [diag(\boldsymbol{v}) - \boldsymbol{v}\boldsymbol{v}'/N] + \left(1 - \frac{1}{N}\right)(\alpha \boldsymbol{W})(\alpha \boldsymbol{W})', \qquad (18)$$

Here $\boldsymbol{v} = E\{\boldsymbol{y}\}$, $\boldsymbol{W}$ has components

$$W_k = N\Delta\varphi(z[k])\frac{z[k]^2 - 1}{\sqrt{2}} = N\Delta w(z[k]) \qquad (19)$$

The Theorem summarizes the effect of $z's$ entire correlation structure in a single parameter α.

Poisson process

Poisson process considerations lead to a somewhat rough but evocative interpretation of the Theorem. Let $\boldsymbol{y} \sim Po(\boldsymbol{u})$ indicate a vector of independent Poisson variates, $y_k \overset{ind}{\sim} Po(u_k)$ for $k = 1, 2, ..., K$, whereas $\boldsymbol{y} \sim (\boldsymbol{v}, \Gamma)$ denotes that vector $\boldsymbol{v}$ has mean $\boldsymbol{v}$ and covariance Γ.

Poisson process

Consider the number of cases N to be a Poisson variate, say

$$N \sim Po(N_0), \tag{20}$$

with $N_0 = 3226$ in the breast cancer study. This simplifies (18) slightly, to

$$cov(\boldsymbol{y}) \doteq diag(\boldsymbol{v}) + (\alpha \boldsymbol{W})(\alpha \boldsymbol{W})'. \tag{21}$$

If the z-values are independent, then (20) makes the counts, y_k,independent Poisson variates,

$$\boldsymbol{y} \sim Po(\boldsymbol{v}),\ \textit{agreeing with}\ (21)\ \textit{at}\ \alpha = 0. \tag{22}$$

Poisson process

A hierarchical model generalizes (22) to incorporate dependence. We assume that $\boldsymbol{y}$ depends on a mean vector $\boldsymbol{u}$, itself random,

$$\boldsymbol{y}|\boldsymbol{u} \sim Po(\boldsymbol{u}), \ \textit{where } \boldsymbol{u} \sim (\boldsymbol{v}, \Gamma), \tag{23}$$

so that the components of $\boldsymbol{y}$ are conditionally independent given $\boldsymbol{u}$ but marginally dependent, with mean and covariance

$$\boldsymbol{y} \sim (\boldsymbol{v}, diag(\boldsymbol{v}) + \Gamma). \tag{24}$$

To match (21), we set

$$\Gamma = (\alpha \boldsymbol{W})(\alpha \boldsymbol{W})'. \tag{25}$$

Poisson process

Formulas (24) and (25) suggest a hierarchical Poisson structure for the count vector $\boldsymbol{y}$,

$$\boldsymbol{y} \sim Po(\boldsymbol{u}), \textit{where } \boldsymbol{u} = \boldsymbol{v} + A\boldsymbol{W}, \textit{with } A \sim (0, \alpha^2) \qquad (26)$$

If $\alpha = 0$, then this reduces to the independence case (22); otherwise, the Poisson intensity vector $\boldsymbol{v}$ is modified by the addition of an independent random multiple A of $\boldsymbol{W}$ with standard deviation α.

Large-scale Significance Testing

Suppose for a moment that we knew which $z_i's$ among the full set of z-values correspond to null cases. For a given choice of x, define

$$Y(x) = \#\{null\ z_i \geq x\} \quad T(x) = \#\{z_i \geq x\} \tag{27}$$

$$FDR(x) = E\{Y(x)\}/T(x). \tag{28}$$

FDR(x) is an empirical Bayes estimate of the a posteriori probability that case i is null given $z_i \geq x$, amounting to a version of Storey's "q-value".

The hierarchical structure gives conditional expectation

$$E\{\boldsymbol{y}|A\} = \boldsymbol{v} + A\boldsymbol{W} \tag{29}$$

Letting the bin width $\Delta \to 0$ produces a conditional version of (29)

$$E\{Y(x)|A\} = N\bar{\Phi}(x)\left[1 + A\frac{x\varphi(x)}{\sqrt{2}\bar{\Phi}(x)}\right] \tag{30}$$

Conditional FDR

$$FDR(Y(x)|A) = FDR_0(x)\left[1 + A\frac{x\varphi(x)}{\sqrt{2}\bar{\Phi}(x)}\right], \tag{31}$$

where $FDR_0(x)$ is the standard unconditional estimate $N\bar{\Phi}(x)/T(x)$ based on the theoretical $N(0,1)$ null.

A can be estimated from the central spread of the histogram of z-values and then used to condition inferences, as in figure 2.

For $x_0 > 0$, let

$$Y_0 = \#\{z_i \in [-x0, x0]\}, \tag{32}$$

and define

$$P_0 = 2\Phi(x_0) - 1 \quad \textit{and} \quad Q_0 = \sqrt{2} x_0 \varphi(x0) \tag{33}$$

Then (30) give $E\{Y_0|A\} \doteq N[P_0 - AQ_0]$, so

$$\hat{A} = \frac{P_0 - \hat{P}_0}{Q_0}, \ \textit{where} \ \hat{P}_0 = Y_0/N \tag{34}$$

We could substitute $\hat{A}$ from (34) into (31) to obtain a conditional FDR estimate,

$$FDR(x|\hat{A}) = FDR_0(x)\left[1 + \hat{A}\frac{x\varphi(x)}{\sqrt{2}\bar{\Phi}(x)}\right], \tag{35}$$

In situations like that of Figure 2, the goal would be to accurately assess our position on the A axis, to better estimate FDP rather than estimating the unconditional average of FDP.

Large-scale Significance Testing

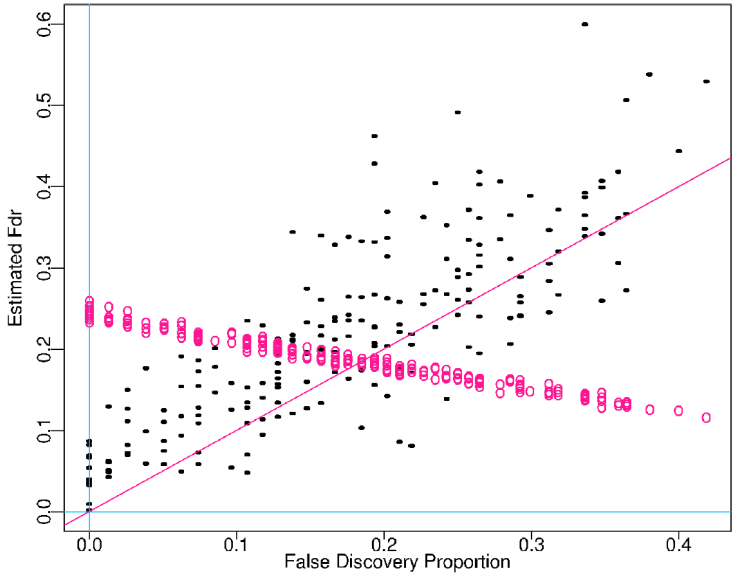

Figure 4: Simulation Experiment Comparing Conditional FDR Estimates (solid points), With Unconditional Estimates (open circles); N = 3,000, p0 = .95. Null counts are generated as in (26), $\alpha = .15$; nonnull counts from $z \sim N(2.5, 1.25)$. The horizontal axis is the actual FDP, FDP(x), x = 2.5, for each of the 200 trials. The unconditional estimate based on the theoretical null distribution declines as actual FDP increases.

Strikingly, the unconditional estimate goes in the wrong direction, declining as the actual *FDP* increases. This yields misleading inferences at both ends of the *FDP* scale. The conditional *FDR* estimate correctly tracks FDP, although with a considerable amount of noise.