Probabilistic Inference for Multiple Testing

Chuanhai Liu and Jun Xie

Department of Statistics,

Purdue University,

West Lafayette, IN 47907.

E-mail: chuanhai, junxie@purdue.edu.

February 22, 2011

Abstract

A probabilistic inferential model is developed for large-scale simultaneous hypothesis testing. For a large set of hypotheses, a sequence of assertions concerning the total number of true alternative hypotheses are proposed. Using a data generating mechanism, the inferential model produces probability triplet (p, q, r) for an assertion conditional on observed data. The probabilities p and q are *for* and *against* the truth of the assertion, whereas $r = 1 - p - q$ is the remaining probability called the probability of "don't know". The inferential model is used for hypotheses of many-normal-means and applied in identifying differentially expressed genes in microarray data analysis. The probabilistic inference offers a new way for hypothesis testing and particularly large-scale multiple testing.

KEY WORDS: False discovery rate; Inferential model; Multiple testing; Random set.

1 Introduction

There have been tremendous research efforts made in last decade on solving large-scale simultaneous hypothesis testing, where one is concerned with a large number n of pairs of competing hypotheses: $\mathcal{H}_0^{(i)}$ versus $\mathcal{H}_a^{(i)}$ for $i = 1, ..., n$. The multiple testing problem is introduced by modern scientific techniques, for example, gene expression microarray in identifying differentially expressed genes from a large number of candidates or even the whole genome. Existing efforts have been made mainly by using the concept of false discovery rate (Benjamini and Hochberg, 1995; Efron *et al.*, 2001; Efron, 2004; Storey, 2002, 2003; and Liang, Liu, and Wang, 2007), which

controls the expected proportion of falsely rejected hypotheses. An alternative way of thinking about this problem is to consider a sequence of assertions:

$$\mathcal{A}_k = \{\text{there are at least } k \ \mathcal{H}_a^{(i)}\text{'s that are true}\}$$

for $k = 1, 2, ..., n$. In our application of identifying significantly expressed genes, we further consider a similar type of assertions as "there are at least j true $\mathcal{H}_a^{(i)}$ in a given interval $[x_1, x_2]$". We will develop probabilistic inference for this type of assertions. We start with a single test for a null hypothesis $\mathcal{H}_0$ versus an alternative hypothesis $\mathcal{H}_a$.

The classic frequency theory of hypothesis testing developed by Neyman, Pearson, and Fisher has been known as the twentieth century's most influential piece of applied mathematics (Berger 2003 and Efron 2008). However, there is a fundamental issue with these existing methods. Fisher (1959) emphasized that p-value, computed from an observed test statistic under the truth of the null hypothesis, provided evidence against $\mathcal{H}_0$. Since the p-value does not have a desirable probability interpretation of whether or not the null hypothesis is true, Fisher (1959) had to argue for the use of p-values by "the force of logic disjunction". In the context of Bayesian hypothesis testing, Bayes factors are often computed to measure evidence in favor one over the other hypothesis. However, like Fisher's p-value, Bayes factors do not have a desirable probability interpretation.

Following Dempster (2008), we view that probabilistic inference for the truth of $\mathcal{H}_0$ or $\mathcal{H}_a$ amounts to producing a probability p for the truth of $\mathcal{H}_0$, a probability q for the truth of $\mathcal{H}_a$, and a residual probability r, called the probability of "don't know", for neither $\mathcal{H}_0$ nor $\mathcal{H}_a$. That is, the triplet (p, q, r) is our uncertainty assessment of $\mathcal{H}_0$ and $\mathcal{H}_a$. Unlike the classic theory of hypothesis testing, this new framework provides

direct statistical evidence for $\mathcal{H}_0$ and $\mathcal{H}_a$. For an analogy with Neyman's hypothesis-testing procedure, with this new framework we could "reject" $\mathcal{H}_0$ by confirming $\mathcal{H}_a$, and "reject" $\mathcal{H}_a$ by confirming $\mathcal{H}_0$. Most important is that the (p, q, r) triplet is calculated from the specification of our uncertainty on unknown model parameters but is not the conditional probability under either the truth of $\mathcal{H}_0$ or the truth of $\mathcal{H}_a$.

We introduce the new framework of probabilistic inference, called inferential models, that produce (p, q, r) for single hypothesis testing in Section 2. The (p, q, r) triplet is calculated based on a data generating mechanism for the observed data. Section 3 considers the many-normal-means problem, where the inferential model is used for multiple testing. Section 4 applies the inferential model of multiple testing in microarray data analysis, to identify differentially expressed genes. Finally, Section 5 concludes with a few remarks.

2 A new framework of probabilistic inference

2.1 A demonstration example

We assume that a set of observed data X is available and that model $f_\theta(X)$ for $X \in \mathcal{X}$ is specified, usually with unknown parameter $\theta \in \Theta$. We use the following example to explain the new framework of probabilistic inference. The key idea is to use an unobserved auxiliary random variable to represent $f_\theta(X)$.

Example 1. Let X be a dichotomous observation with $X \in \mathcal{X} = \{0, 1\}$. Assume a Bernoulli model

$$P_\theta(X = 1) = \theta \qquad \text{and} \qquad P_\theta(X = 0) = 1 - \theta$$

with unknown $\theta \in \Theta = [0, 1]$. The problem is to infer θ from X. We consider a data generating mechanism using an auxiliary random variable $U \sim \text{Unif}\,(0, 1)$:

$$X = \begin{cases} 1, & \text{if } U \leq \theta; \\ 0, & \text{if } U > \theta. \end{cases}$$

This sampling mechanism preserves the model for X given θ. Moreover, it creates a random set for the parameter θ given the observation X

$$\mathcal{S}_X = \begin{cases} U \leq \theta \leq 1, & \text{if } X = 1; \\ 0 \leq \theta < U, & \text{if } X = 0, \end{cases} \qquad (U \sim \text{Unif}\,(0, 1))$$

In other words, we think $\theta \in [U, 1]$ if we observe $X = 1$ and $\theta \in [0, U)$ if $X = 0$, where U is a random variable from $\text{Unif}\,(0, 1)$. *This relationship among the parameter of interest θ, the observation X, and the auxiliary random variable U is critical in our construction of the probabilistic inferential model, where inference about the parameter θ will be derived from prediction of the auxiliary random variable U.*

Given, for example, $X = 1$, we have the random interval $\mathcal{S}_X = [U, 1]$ as the region for θ. Now consider an assertion $\mathcal{A} = \{\theta \leq \theta_0\} \subseteq \Theta$ for a fixed $\theta_0 \in (0, 1)$. There are two possible cases: (i) if $U > \theta_0$, the random set $\mathcal{S}_X = [U, 1]$ for θ provides evidence against the truth of $\mathcal{A}$; (ii) if $U \leq \theta_0$, the random set $\mathcal{S}_X = [U, 1]$ for θ does not have any information about the truth or falsity of $\mathcal{A}$. Note that there is no realization of the random interval that provides evidence for the truth of $\mathcal{A}$, because the random set $[U, 1]$ cannot be fully contained in $\mathcal{A} = \{\theta \leq \theta_0\}$. As a result, the probability triplet (p, q, r) for the assertion $\mathcal{A}$ are calculated in the following

$$p = 0, \quad q = P\{U > \theta_0\} = 1 - \theta_0, \text{ and } r = \theta_0.$$

Generally, assertions about θ can be represented by subsets of Θ. An assertion $\mathcal{A} \subset \Theta$ is said to be true when the true value θ falls into $\mathcal{A}$, and is said to be false when the true value θ falls into $\mathcal{A}^c = \Theta \setminus \mathcal{A}$, the negation of $\mathcal{A}$. For example, in the problem of testing the two competing hypotheses $H_0 : \theta = \theta_0$ and $H_a : \theta \neq \theta_0$, where θ_0 is some known value in Θ, the assertion $\mathcal{A} = \{\theta_0\}$ stands for H_0 and, thereby, the assertion $\mathcal{A}^c = \{\theta : \theta \in \Theta; \theta \neq \theta_0\}$ stands for H_a. The inference problem is then to produce our uncertainty assessment of the truth and falsity of $\mathcal{A}$. That is, the output of our inference is the probability triplet (p, q, r) for $\mathcal{A}$.

2.2 Inferential models

To emphasize the fact that the (p, q, r) output is conditional on the observed data X, we write (p, q, r) as $(p_X(\mathcal{A}), q_X(\mathcal{A}), r_X(\mathcal{A}))$, that is,

$p_X(\mathcal{A})$ — the probability for the truth of $\mathcal{A}$, given X

$q_X(\mathcal{A})$ — the probability against the truth of $\mathcal{A}$, given X

$r_X(\mathcal{A})$ — the probability of "don't know" or neither for nor against the truth of $\mathcal{A}$, given X.

Formally, an inferential model for probabilistic inference about θ is given by a probability model with the sample space consisting of all subsets of Θ. Its probability measure is defined by an auxiliary random variable, for example the uniform variable U in Example 1. More specifically, a random set is constructed for inference about θ using the auxiliary random variable and conditioning on the observed data X. Denote the random set $\mathcal{S}_X$, as in Example 1. The probability for the truth of a given assertation $\mathcal{A}$ (on the parameter θ) is computed as the probability that the random set $\mathcal{S}_X$ is contained in $\mathcal{A}$,

$$p_X(\mathcal{A}) = P(\mathcal{S}_X \subseteq \mathcal{A}).$$

Based on a symmetry argument, the probability against the truth of $\mathcal{A}$ or for the truth of $\mathcal{A}^c$ is computed as the probability that the random set $\mathcal{S}_X$ is contained in $\mathcal{A}^c$,

$$q_X(\mathcal{A}) = P(\mathcal{S}_X \subseteq \mathcal{A}^c).$$

The remaining probability

$$r_X(\mathcal{A}) = 1 - p_X(\mathcal{A}) - q_X(\mathcal{A})$$

is the probability that the random set $\mathcal{S}_X$ intersects with both $\mathcal{A}$ and $\mathcal{A}^c$, in which case we "don't know" the truth or falsity of $\mathcal{A}$.

In order for the probability triplet $(p_X(\mathcal{A}), q_X(\mathcal{A}), r_X(\mathcal{A}))$ to have desirable long-run frequency properties, the concept of credibility is helpful.

Definition 1. *The inferential model is credible for assertion $\mathcal{A}$ if for every α in $(0, 1)$, both*

$$P_\theta(\{X : p_X(\mathcal{A}) \geq \alpha\}) \leq 1 - \alpha \quad \text{and} \quad P_\theta(\{X : q_X(\mathcal{A}) \geq \alpha\}) \leq 1 - \alpha \tag{1}$$

hold respectively for every $\theta \in \mathcal{A}^c = \Theta \setminus A$ and for every $\theta \in \mathcal{A}$. The probabilities in (1) are defined with respect to the random variable X following $f_\theta(X)$.

In other words, credibility requires $p_X(\mathcal{A})$ and $q_X(\mathcal{A})$, as functions of the random variable X, to be stochastically bounded by the uniform distribution over the unit interval $(0, 1)$ in repeated experiments. Thus, the triplet $(p_X(\mathcal{A}), q_X(\mathcal{A}), r_X(\mathcal{A}))$ provides strength of evidence for both $\mathcal{A}$ and $\mathcal{A}^c$ in the long-run frequency probability scale. For those familiar with the Neyman school of thought on hypothesis testing, thresholds for $p_X(\mathcal{A})$ and $q_X(\mathcal{A})$ can be used to confirm the truth and falsity of $\mathcal{A}$.

2.3 An inferential model for a general distribution

Now we generalize the inferential model of the Bernoulli example to any non-parametric distributions. Suppose that we have a sample $X_1, ..., X_n$ from an unknown continuous distribution with cdf $F(x)$, $x \in \mathbb{R}$. Let $X_{(1)} \leq ... \leq X_{(n)}$ denote the order statistics of the sample. Then inference about $F(x)$ at values $x = X_{(1)}, ..., X_{(n)}$ can be made based on the fact that $F(X_{(i)})$, $i = 1, ..., n$, are the unobserved order statistics of a random sample of size n from the uniform distribution Unif $(0, 1)$. Let $\mathbb{S}_n$ denote the space of the order statistics of a uniform sample of size n:

$$\mathbb{S}_n = \{(u_1, ..., u_n) : \ 0 < u_1 < ... < u_n < 1\}.$$

To specify a random set, we define the following function on $\mathbb{S}_n$:

$$g(u) = \sum_{i=1}^{n} [a_i \ln u_i + b_i \ln(1 - u_i)] \qquad (u \in \mathbb{S}_n) \tag{2}$$

where $a_i = 1/(n - i + .7)$ and $b_i = 1/(i - 1 + .7)$ for $i = 1, ..., n$. This function serves as a measurement of how close a sequence of ordered values $0 < u_1 < ... < u_n < 1$ to the individual medians of the ordered uniform random variables $U = (U_{(1)}, ..., U_{(n)})$. Figure 1 shows contours of the function in the space $\mathbb{S}_2 = \{(u_1, u_2) : 0 < u_1 < u_2 < 1\}$. The function achieves the maximum at the marginal medians of $U_{(i)}$'s and decreases towards the boundary of $\mathbb{S}_n$. Define a random set

$$\mathcal{S} = \{u : \ u \in \mathbb{S}_n; \ g(u) \geq g(U)\}$$

where $U = (U_{(1)}, ..., U_{(n)})$ is a vector of sorted n uniform random variables. This random set corresponds to the inner area of the curve $g(U)$ and predicts a region for

the unobserved uniform vector $(F(X_{(1)}), ..., F(X_{(n)}))$.

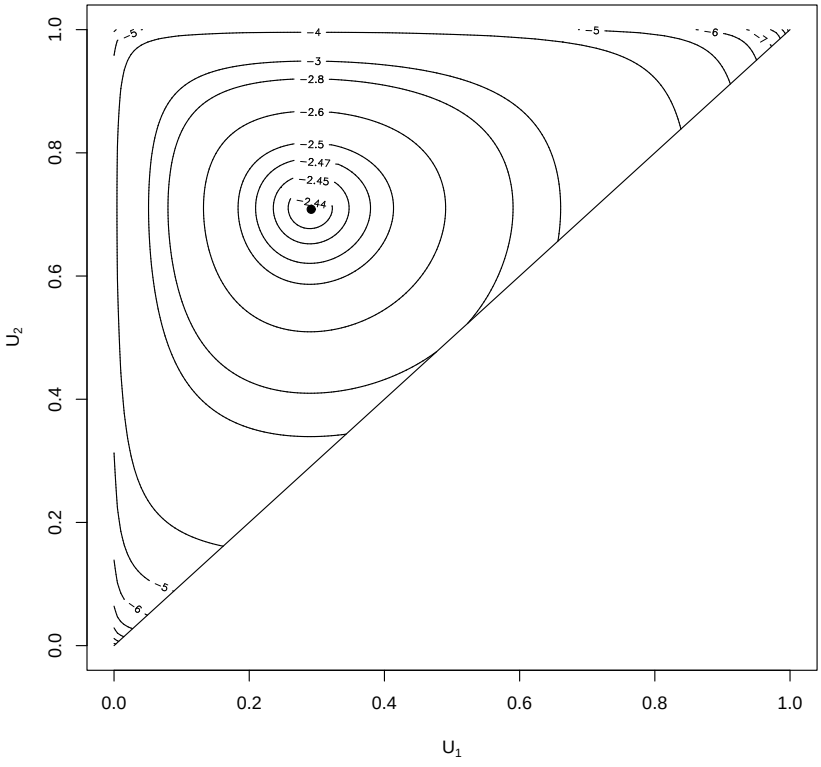

Figure 1: Contours of the g function defined in (2) in the space of two ordered uniform samples $\mathbb{S}_2 = \{(u_1, u_2) : 0 < u_1 < u_2 < 1\}$

For inference about the distribution function $F(x)$, $x \in \mathbb{R}$, we define a random set

$$\mathcal{S}_X = \{F : \ F \in \mathbb{C};\ g(F(X_{(1)}), ..., F(X_{(n)})) \geq g(U)\}, \tag{3}$$

where $U = (U_{(1)}, ..., U_{(n)})$ is a vector of sorted n uniform random variables and $\mathbb{C}$ denotes the space consisting of all continuous cdf's on $\mathbb{R}$. Consider hypotheses $H_0 : F = F_0$ versus $H_a : F \neq F_0$. The inferential model gives $p_X(\mathcal{H}_0) = P_X(\mathcal{S}_X \subseteq \{F_0\}) =$

0,

$$q_X(\mathcal{H}_0) = p_X(\mathcal{H}_a) = P_X(\mathcal{S}_X \subseteq \{F_0\}^c) = P(g(U) \geq g(F_0(X_{(1)}), ..., F_0(X_{(n)}))),$$

and $r_X(\mathcal{H}_0) = 1 - q_X(\mathcal{H}_0)$. Intuitively, for the simple hypothesis $H_0 : F = F_0$, if the observations $X_{(1)}, ..., X_{(n)}$ are really from the null distribution F_0, then $(F_0(X_{(1)}), ..., F_0(X_{(n)}))$ are the order statistics of uniform random variables and $g(F_0(X_{(1)}), ..., F_0(X_{(n)}))$ should have a large value. The event $g(U) \geq g(F_0(X_{(1)}), ..., F_0(X_{(n)}))$ provides evidence against the null hypothesis hence gives the probability $q_X(\mathcal{H}_0)$. Zhang (2010) showed that this inferential model has the desirable frequency property:

Theorem 1. *The inferential model with the random set (3) is credible for any assertion* $\mathcal{A} \subset \mathbb{C}$.

He also demonstrated that when compared in terms of power, this inferential model is more efficient than conventional methods of hypothesis testing (Zhang 2010).

3 The many-normal-means problem

The many-normal-means problem is a benchmark problem for inference about multiple testing. Suppose that the observed data set consists of n data points $X_1, ..., X_n$ from the model:

$$X_i \overset{iid}{\sim} N(\theta_i, 1), \qquad i = 1, ..., n.$$

In the context of multiple testing, a typical assumption is that most of θ_i are zero and we test for a large number of hypotheses $\mathcal{H}_0^{(i)} : \theta_i = 0$ versus $\mathcal{H}_a^{(i)} : \theta_i \neq 0$ for $i = 1, ..., n$. We refer to each non-zero θ_i (and the corresponding X_i) as an outlier and want to identify the outliers presented in the data. As stated in Section 1, this

problem is translated into producing (p, q, r) outputs for a sequence of assertions

$$\mathcal{A}_k = \{\text{the number of true } \mathcal{H}_a^{(i)} \geq k\}, \qquad k = 1, 2, ..., n.$$

An inferential model for the many-normal-means problem can be constructed through the following data generating mechanism,

$$X_i = \theta_i + \Phi^{-1}(U_i), \qquad i = 1, ..., n,$$

where $U_1, ..., U_n$ are random samples from the uniform Unif $(0, 1)$ and $\Phi^{-1}(.)$ stands for the inverse cdf of the standard normal $N(0, 1)$. If $U = (U_1, ..., U_n)$ were observed, the values of θ_i would have been known by calculating

$$\theta_i = X_i - \Phi^{-1}(U_i).$$

We will predict the unobserved $U_1, ..., U_n$ via a random set, which leads to a random set for the parameter $\theta = (\theta_1, ..., \theta_n)$. Recall the random set from a sorted uniform random vector $U = (U_{(1)}, ..., U_{(n)})$, as discussed in Section 2, $S = \{u : u \in \mathbb{S}_n, g(u) \geq g(U)\}$. It derives a random set for inference about θ as

$$\mathcal{S}_X = \{\theta : \theta \in \mathbb{R}^n, g(\Phi(\langle X - \theta \rangle)) \geq g(U)\}, \tag{4}$$

where $\Phi(\langle X - \theta \rangle)$ represents the sorted $(\Phi(X_1 - \theta), ..., \Phi(X_n - \theta))$. An assertion $\mathcal{A}$ about the parameter θ, as a subset in the parameter space, will be evaluated by a probability triplet with $p(\mathcal{A}) = P(\mathcal{S}_X \subseteq \mathcal{A})$, and $q(\mathcal{A}) = P(\mathcal{S}_X \subseteq \mathcal{A}^c)$.

3.1 Inference about the number of outliers

The assertion $\mathcal{A}_k$ means "there are at least k outliers" and thereby its negation is $\mathcal{A}_k^c = \{\|\theta\|_0 \leq k-1\}$, where $\|\theta\|_0$ represents the number of non-zero components of θ. We use the random set (4) to produce a probability triplet about $\mathcal{A}_k$ and $\mathcal{A}_k^c$, with

$$p(\mathcal{A}_k) = q(\mathcal{A}_k^c) = P(\mathcal{S}_X \subseteq \mathcal{A}_k), \quad q(\mathcal{A}_k) = p(\mathcal{A}_k^c) = P(\mathcal{S}_X \subseteq \mathcal{A}_k^c),$$

and

$$r(\mathcal{A}_k) = 1 - p(\mathcal{A}_k) - q(\mathcal{A}_k).$$

To compute $p(\mathcal{A}_k^c) = P(\mathcal{S}_X \subseteq \mathcal{A}_k^c)$, we note that the event $\mathcal{S}_X \subseteq \mathcal{A}_k^c$ is equivalent to that $\|\theta\|_0 \leq k-1$ holds for all $\theta \in \mathcal{S}_X = \{\theta : g(\Phi(\langle X - \theta \rangle)) \geq g(U)\}$. This is an impossible event, because if we take $\theta_i = X_i - \Phi^{-1}(U_i)$ for $i = 1, ..., n$ then we have $\|\theta\|_0 = n$ and $g(\Phi(\langle X - \theta \rangle)) \geq g(U)$. Therefore,

$$p(\mathcal{A}_k^c) = P(\mathcal{S}_X \subseteq \mathcal{A}_k^c) = 0.$$

To compute $q(\mathcal{A}_k^c) = P(\mathcal{S}_X \subseteq \mathcal{A}_k)$, we note that the event $\mathcal{S}_X \subseteq \mathcal{A}_k$ means that $\|\theta\|_0 \geq k$ holds for all θ satisfying $g(\Phi(\langle X - \theta \rangle)) \geq g(U)$. Thus, the event $\mathcal{S}_X \subseteq \mathcal{A}_k$ is equivalent to

$$\max_{\theta : \|\theta\|_0 \leq k-1} g(\Phi(\langle X - \theta \rangle)) < g(U).$$

The constraint $\|\theta\|_0 \leq k-1$ implies that "except for at most $(k-1)$ X_i's, the others form a sample of size $(n-k+1)$ from $N(0,1)$". Therefore, we choose to work on the corresponding $g(.)$ function defined over the $(n-k+1)$-dimensional space instead of that defined over the n-dimensional space. (The resulting inference is more efficient because it effectively marginalizes out the $(k-1)$ potential outliers.) Let $\mathbb{Y}_{n-k+1}$ be

the set of all $\binom{n}{n-k+1}$ combinations of $(n-k+1)$ X_i's. We want to solve the following optimization problem

$$\max_{Y \in \mathbb{Y}_{n-k+1}} g(\Phi(Y_{(1)}), ..., \Phi(Y_{(n-k+1)})).$$

Let $g_* = \max_{Y \in \mathbb{Y}_{n-k+1}} g(\Phi(Y_{(1)}), ..., \Phi(Y_{(n-k+1)}))$. Then $p(\mathcal{A}_k) = P(\mathcal{S}_X \subseteq \mathcal{A}_k)$ can be computed by first finding g_* and then approximating the probability $p(\mathcal{A}_k) = P(g(U) > g_*)$ via Monte Carlo methods, i.e.,

$$p(\mathcal{A}_k) \approx \frac{1}{M} \sum_i I_{\{g(U^{(i)}) > g_*\}}$$

where $U^{(1)}, ...U^{(M)}$ are M samples drawn from the distribution of $n-k+1$ sorted uniforms and $I_{\{g(U^{(i)}) > g_*\}} = 1$ if $g(U^{(i)}) > g_*$ and 0 otherwise.

The problem of maximizing $\max_{Y \in \mathbb{Y}_{n-k+1}} g(\Phi(Y_{(1)}), ..., \Phi(Y_{(n-k+1)}))$ is a so-called NP-hard problem, when all possible combinations $\binom{n}{n-k+1}$ are considered. We propose an efficient algorithm for solving this optimization problem in Appendix.

We provide a simple simulation study to show the performance of the proposed method. We simulate data sets each consisting of a sample of 10,000 from $N(0,1)$ and a sample of 100 from $N(5,1)$. The result, the probability for the truth of the assertion that "there are at least k outliers" for a sequence of $k = 1, 2, ...$ in each of ten simulated data sets, is displayed in Figure 2. We see that the probability of at least k outliers is 1 when k is small and high when $k < 100$. This probability decreases dramatically when k is around 100. After the number of outliers passes 100, the probability becomes to wander towards zero in a slow pace. The probability at this level represents the randomness of true ordered uniform deviates.

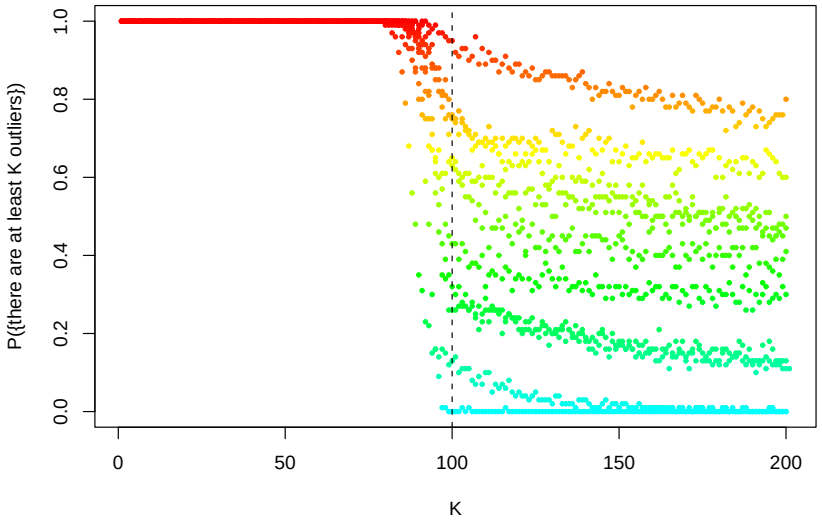

Figure 2: The probability for the truth of the assertion that "there are at least k outliers" in 10 simulated data sets each consisting of a sample of 10,000 from $N(0,1)$ and a sample of 100 from $N(5,1)$.

3.2 Inference for the number of outliers in an interval

Assume that there are k outliers in the observed data set $X_1, ..., X_n$. That is, there are $n-k$ of these n observed values that are known to form a sample from $N(0,1)$. We are interested in the number of outliers in a given interval, say, $[x_1, x_2]$ (e.g, $x_1 = 3$ and $x_2 = \infty$). Formally, we consider the assertion that "there are $J \geq j$ outliers in $[x_1, x_2]$, conditioning on k outliers in the whole set of n observations". To make a probabilistic inference about this assertion, we start with a data generating mechanism for the observed count of the number of X_i's that fall in $[x_1, x_2]$, denoted as C_{x_1,x_2}.

It is known that there are $(n-k)$ X_i's from $N(0,1)$. Consider an auxiliary random variable N_0 as the number of these $n-k$ standard normals that fall into the interval $[x_1, x_2]$. Then N_0 follows a binomial distribution, Binomial$(n-k, \Phi(x_2) - \Phi(x_1))$.

There is a critical relationship among the observed count C_{x_1,x_2}, the quantity of interest J, and the unobserved random variable N_0, that is, $C_{x_1,x_2} = N_0 + J$. If N_0 were observed, we could obtain an inference about J as $C_{x_1,x_2} - N_0$. Since N_0 is unobserved, we use a random set $\{0, 1, ..., N\}$ to predict it, where $N \sim \text{Binomial}(n - k, \Phi(x_2) - \Phi(x_1))$. This leads to a random set for inference about J,

$$\mathcal{S} = \{J : C_{x_1,x_2} - N \leq J \leq C_{x_1,x_2}\}, \qquad N \sim \text{Binomial}(n - k, \Phi(x_2) - \Phi(x_1)).$$

For a probabilistic inference about the assertion $\{J : J \geq j\}$, that is, "there are at least j outliers in $[x_1, x_2]$, conditioned on k outliers in the whole set of n observations", we compute the (p, q, r) output as follows:

$$\begin{aligned} p &= P(\mathcal{S} \subseteq \{J : J \geq j\}) = P(C_{x_1,x_2} - N \geq j) \\ &= P(N \leq C_{x_1,x_2} - j) = \text{pBinomial}(C_{x_1,x_2} - j, n - k, \Phi(x_2) - \Phi(x_1)), \\ q &= P(\mathcal{S} \subseteq \{J : J < j\}) = P(\emptyset) = 0, \\ r &= 1 - p, \end{aligned}$$

where $\text{pBinomial}(., n - k, \Phi(x_2) - \Phi(x_1))$ denotes the cdf of $\text{Binomial}(n - k, \Phi(x_2) - \Phi(x_1))$. For an analogy with the concept of false discovery rate (FDR), one may choose to report

$$\text{FDR}_{x_1,x_2} = \max\left\{0, \frac{(n-k)[\Phi(x_2) - \Phi(x_1)]}{C_{x_1,x_2}}\right\},$$

which gives an expected number of falsly rejected null hypotheses in a rejection interval $[x_1, x_2]$.

4 Application in microarray data analysis

We study an HIV data set included in an R package called *nudge* (Dean, 2006) for detection of differential gene expression. The data consists of cDNA from CD4+ T cell lines at 1 hour after infection with HIV, from a study by van't Wout *et al.* (2003). The data have the following structure:

Data Structure

	Sample 1				Sample 2			
	Dye 1		Dye 2		Dye 1		Dye 2	
Gene	r1	r2	r1	r2	r1	r2	r1	r2
⋮			4608 × 8 Data Values					

where Sample 1 and Sample 2 correspond to the HIV infected and the control samples, Dye 1 and 2 correspond to two dye labeling schemes, and r1 and r2 correspond to two duplicate microarray slides. A standard approach of identifying significance gene expression is to calculate z-scores from the 4,608 individual two-sample t-tests based on four log-intensity values in Sample 1 and four log-intensity values in Sample 2, where $z_i = \Phi^{-1}(F_6(t_i))$ with F_6 as the cumulative distribution function of a standard t variable with 6 degrees of freedom. The null distribution for these z-scores is modeled as either a standard normal or an empirical normal distribution. Then false discovery rate or local false discovery rate can be employed to detect significance genes.

Alternatively, we apply the probabilistic inference model for the multiple testing problem. We first conduct an exploratory data analysis based on the original $4,608 \times 8$ data matrix.

4.1 Exploratory data analysis

Let $R_{g,s,d,r}$ be the intensity in the "raw" data matrix for gene $g \in \{1, 2, ..., n = 4,608\}$, sample $s \in \{1, 2\}$, dye $d \in \{1, 2\}$, and duplicate $r \in \{1, 2\}$. For a variance stabilization transformation better than logarithmic, we use

$$Y_{g,s,d,r} = \ln(R_{g,s,d,r} + 8), \qquad g \in \{1, 2, ..., n\}; s \in 1, 2; d \in \{1, 2\}; r \in \{1, 2\}.$$

For each of the four combinations of s and d, we plot the differences $Y_{g,s,d,r=2} - Y_{g,s,d,r=1}$ versus $\bar{Y}_{g,s,d} = (Y_{g,s,d,r=1} + Y_{g,s,d,r=2})/2$ for all $g = 1, ..., n$. The distribution of the differences $Y_{g,s,d,r=2} - Y_{g,s,d,r=1}$ is approximately symmetric about zero, not depending on $\bar{Y}_{g,s,d}$ (figures not shown).

On the other hand, Figure 3 displays that the differences $(\bar{Y}_{g,s_1,d_1} - \bar{Y}_{g,s_2,d_2})$ depends on the average $(\bar{Y}_{g,s_1,d_1} + \bar{Y}_{g,s_2,d_2})/2$ through a nonlinear function, where $(s_1, d_1) \neq (s_2, d_2)$. In Figure 3, the range of the average $(\bar{Y}_{g,s_1,d_1} + \bar{Y}_{g,s_2,d_2})/2$ is binned into 20 bins and the differences $(\bar{Y}_{g,s_1,d_1} - \bar{Y}_{g,s_2,d_2})$ for all $g = 1, 2, ..., n$ are grouped accordingly into 20 groups. The distribution of the differences $(\bar{Y}_{g,s_1,d_1} - \bar{Y}_{g,s_2,d_2})$ is about symmetric but not around zero. The plots show strong evidence of smooth changes of the medians in the boxplots cross the range of the averages $(\bar{Y}_{g,s_1,d_1} + \bar{Y}_{g,s_2,d_2})/2$, which implies that there is a dye effect within a given sample and there is a sample effect for all genes. Both effects need to be removed if we assume that most genes are not differentially expressed between the two samples and that the two dye labeling schemes should not affect gene expression.

We conduct a loess-regression for the medians in the boxplots in Figure 3. Then we transform the original data as $(Y_{g,s,d,r} - \bar{Y}_{g,s,d}) + loess_{g,s,d}$, where $loess_{g,s,d}$ is the loess fit of the medians. For simplicity, we use the same notation $Y_{g,s,d,r}$ to denote the transformed data. Figure 4 shows that the dependence of $\bar{Y}_{g,s_1,d_2} - \bar{Y}_{g,s_2,d_2}$ on the

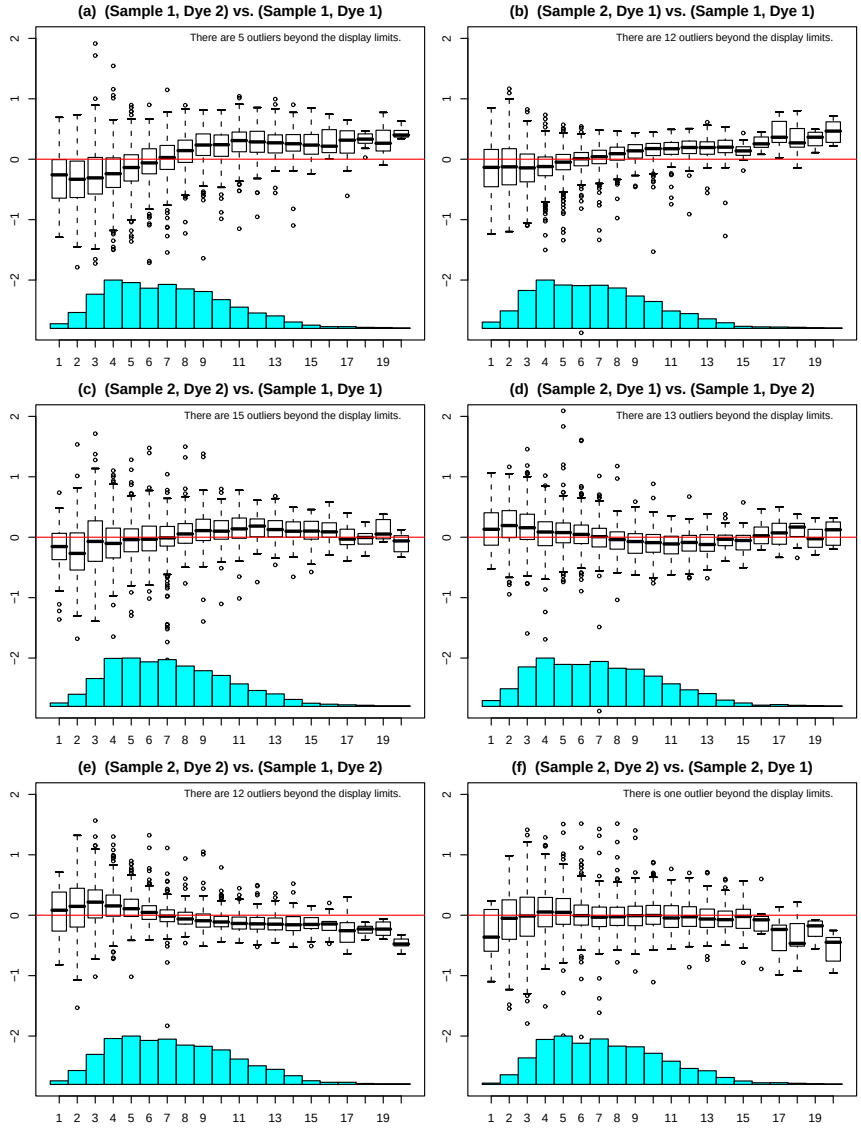

Figure 3: Boxplots of grouped differences $\bar{Y}_{g,s_1,d_2} - \bar{Y}_{g,s_2,d_2}$ for all $g = 1, ..., n$, grouped according to the binned values of $(\bar{Y}_{g,s_1,d_2} + \bar{Y}_{g,s_2,d_2})/2$, where $(s_1, d_1) \neq (s_2, d_2)$. The labels in the x-axis index the bins for $(\bar{Y}_{g,s_1,d_2} + \bar{Y}_{g,s_2,d_2})/2$. The relative frequency of each bin is shown by the histograms in the bottom of each plot.

average expression value $(\bar{Y}_{g,s_1,d_2} + \bar{Y}_{g,s_2,d_2})/2$, $g = 1, ..., n$, has been removed after the transformation.

Let $\bar{Y}_{g,s}$ be the mean of the four transformed values of $\{Y_{g,s,d,r} : d = 1, 2; r = 1, 2\}$ for $s = 1, 2$ and $g = 1, ..., n$. The exploratory data analysis indicates that information on significantly expressed genes are in the differences

$$\delta_g = \bar{Y}_{g,2} - \bar{Y}_{g,1}, \qquad g = 1, ..., n.$$

Since δ_g is computed as the difference of two means of four observations having the same distribution, we expect that δ_g is normally distributed, and has mean zero for most of the n genes. The subsequent effort is to model the n individual variances. The boxplots of grouped δ_g's, according to binned means $\mu_g = (\bar{Y}_{g,2} + \bar{Y}_{g,1})/2$ are shown in Figure 5. It indicates that the variance of δ_g depends on μ_g. To check the local normality along the values of μ_g, we draw the Q-Q normal plots of grouped δ_g's in Figure 6. We conclude that (i) Figure 6 does not show strong evidence against the assumption that δ_g's are normally distributed locally along the values of μ_g and (ii) Figure 5 suggests the means and variances of δ_g's be modeled as smooth functions of μ_g for most of the n genes. We again use loess-regression and compute smooth estimates of means and standard deviations of δ_g's. Finally, we compute the standardized δ_g's, denoted by Z_g, using the estimated mean and standard deviation curves. The histogram of Z_g's in Figure 7 confirms that most of Z_g's follow $N(0, 1)$, supporting our data analysis.

4.2 Identification of significantly expressed genes

Identifying significantly expressed genes in the present case becomes the same problem of many-normal-means discussed in Section 3. We test for 4608 hypotheses that the

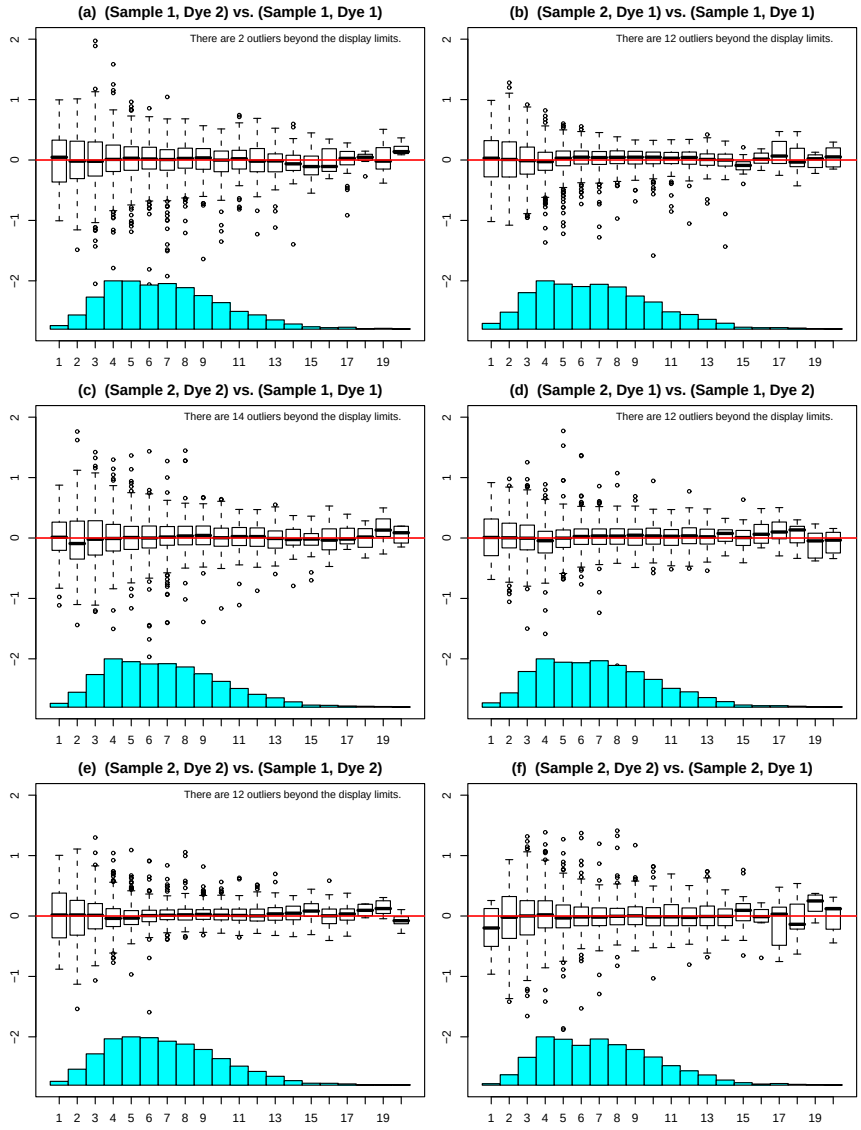

Figure 4: This is the same as Figure 3, but for the transformed data.

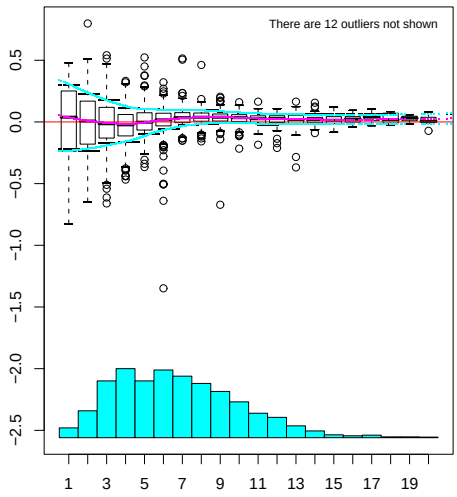

Figure 5: Boxplots of grouped $\delta_g = \bar{Y}_{g,2} - \bar{Y}_{g,1}$, according to the binned values of $(\bar{Y}_{g,2} + \bar{Y}_{g,1})/2$. The labels in the x-axis index the bins. The relative frequency of each bin is shown by the histograms.

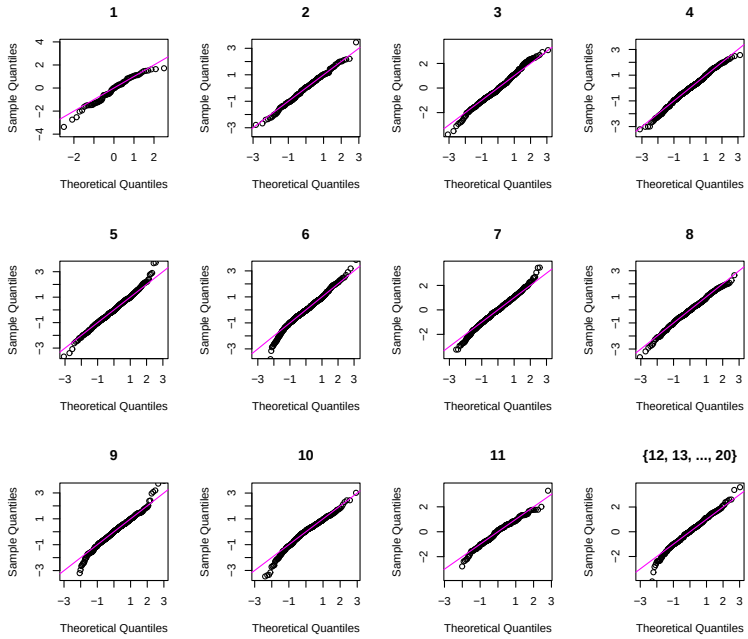

Figure 6: Q-Q normal plots of grouped δ_g's.

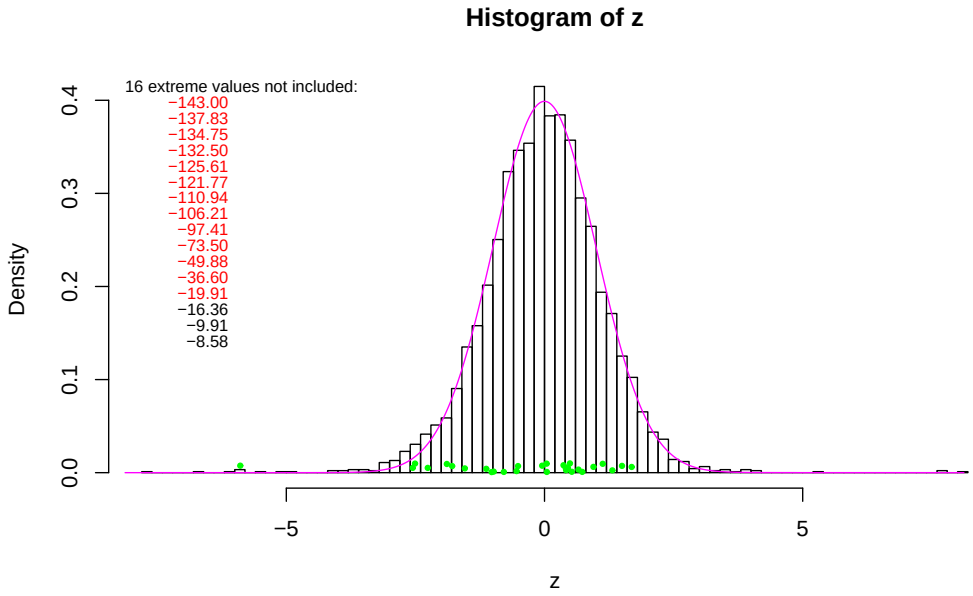

Figure 7: The histogram of the Z-values with the underlying curve of the standard normal density. The Z-values for positive and negative controls are marked by red (text) and green (dots).

mean of Z_g is zero versus the alternative of nonzero. The probability $p_X(\mathcal{A}_k)$ shown in the top panel in Figure 8 is the probability that the number of true alternative hypotheses is at least k, or there are at least k significance genes. This probability is about 1 up to $k = 40$ then drops dramatically and reaches 0 when $k = 80$. The probability curve indicates that the number of significantly expressed genes is in the range of 40 to 60. Probabilistic inference for assertions concerning the total number of significantly expressed genes in a given interval $(-\infty, -3)$ is computed as discussed in Section 3.2. Conditioning on a total of $k = 60$ significance genes, the probability for the assertion that "there are at least J outliers in $(-\infty, -3)$" is displayed in the bottom panel in Figure 8. The probability curve implies that there are about 35-40 significance genes in the given interval.

In this gene expression data set, there are 13 genes known to be differentially expressed (HIV genes) and 29 genes known not to be (non-human genes). These two sets of genes serve as positive and negative controls respectively. The 13 positive control genes have the largest 13 Z-values according to our exploratory analysis and are marked by red in the list of extreme values in Figure 7. The 29 negative control genes are marked by green in Figure 7, where 28 out 29 negative controls have small Z-values (absolute value less than 3) hence are correctly identified as negative.

5 Concluding remarks

We consider a formal way of summarizing uncertainty in statistical inference. We focus on the problem of hypothesis testing, especially multiple testing, by a probabilistic inferential model about an unobserved ordered uniform sample. We show that hypothesis testing problems can be treated as (p, q, r) computations. With the example of microarray data analysis, we have demonstrated the importance of exploratory

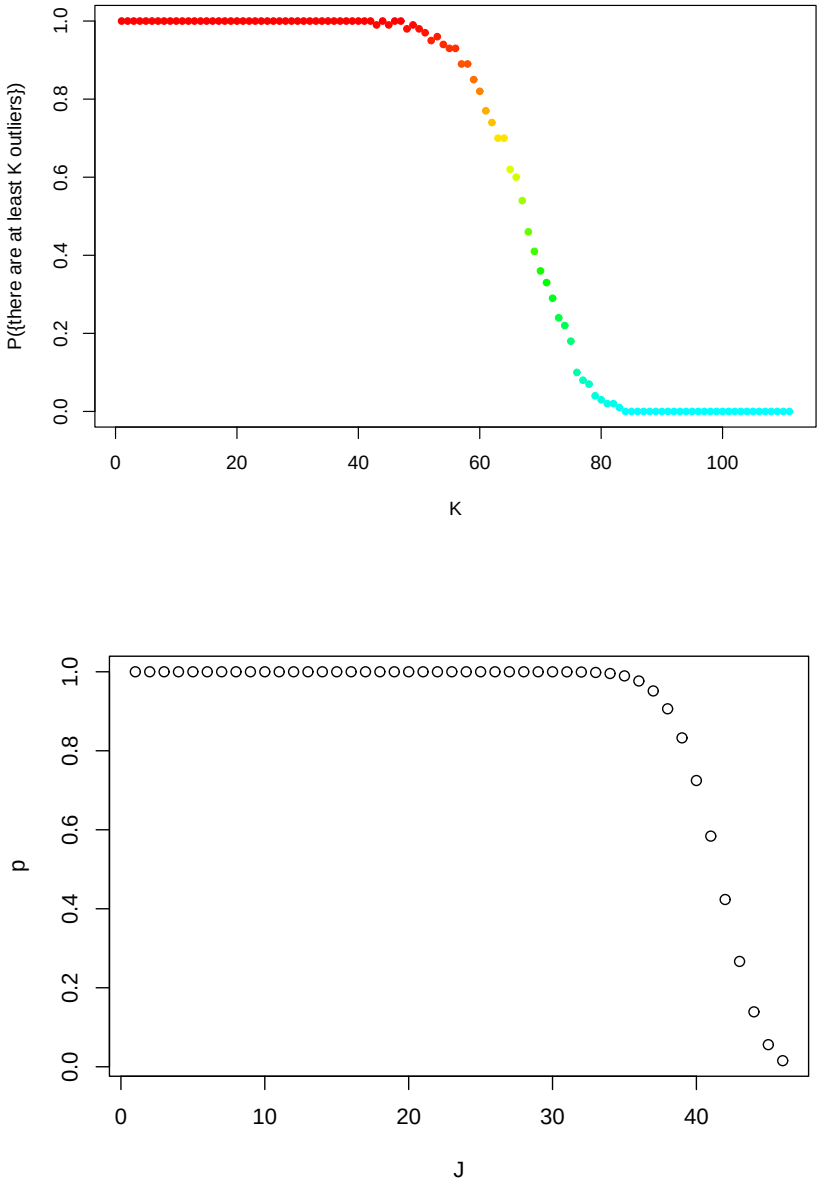

Figure 8: Top panel: the probability $p_X(\mathcal{A}_k)$ for at least k significance genes out of 4608; Bottom panel: the probability for the assertion that "there are at least J significance genes in $(-\infty, -3)$", conditioning on that there exist $k = 60$ significance genes in total.

data analysis as well as the application of probabilitic inference in multiple testing.

The proposed method can be extended to other problems involving simultaneous hypothesis testing, including the many-Poisson-means problem and variable selection in linear regression. In particular, development of such methods for multinomial of a large number of categories is both theoretically challenging and practically useful for single nucleotide polymorphism (SNP) analysis in genome-wide association studies. Those will be our future research works.

Appendix

Suppose that a subset of size $m = n - k + 1$ or $n - k$, $Y_1, ..., Y_m$, is chosen from $X_1, ..., X_n$ by maximizing

$$g(\Phi(Y_{(1)}), ..., \Phi(Y_{(m)})) = \sum_{i=1}^{m} \left[\alpha_i \ln \Phi(Y_{(i)}) + \beta_i \ln(1 - \Phi(Y_{(i)}))\right],$$

where $\beta_i = 1/(i-1+0.7)$ and $\alpha_i = \beta_{m-i+1}$ for $i = 1, ..., m$. The problem of maximizing this objective function over all possible "n-choose-m" combinations $(Y_{(1)}, ..., Y_{(m)})$ from the observed data $(X_{(1)}, ..., X_{(n)})$ is a so-called NP-hard problem. We propose an efficient algorithm by finding the best "match"

$$s(i) : \ (Y_{(i)}, X_{(s(i))}),$$

for each $i = 1, ..., m$, to maximize the i-th term of the objective function, i.e., $\alpha_i \ln \Phi(X_{(s(i))}) + \beta_i \ln(1 - \Phi(X_{(s(i))}))$, over $s(i) = 1, ..., n$. When it is one-to-one, the matching function $s(.)$ produces the desired solution: $Y_{(i)} = X_{(s(i))}$ for $i = 1, ..., m$. Otherwise, this matching function $s(.)$ is updated iteratively in a conditional/constrained optimization fashion toward the target.

An optimal matching algorithm

Denote $U_{(i)} = \Phi(X_{(i)})$ for $i = 1, ..., n$. Rewrite the objective function as

$$g(s) = \sum_{i=1}^{m} \left[\alpha_i \ln U_{(s(i))} + \beta_i \ln(1 - U_{(s(i))})\right]$$

where $s \in \mathbb{C}_n^m$ represents a strictly monotone mapping from $\{1, ..., m\}$ to $\{1, ..., n\}$. The problem is to find s to maximize the objective function $g(s)$. The following algorithm provides the desired solution $\arg\max_s g(s)$ at convergence. Start with an initial mapping by finding the most preferred match $U_{s(i)}$ for each $i = 1, ..., m$:

$$s(i) = \arg \max_{1 \le j \le n} \left[\alpha_i \ln U_{(j)} + \beta_i \ln(1 - U_{(j)})\right].$$

Repeat the following 3 steps until $s(.)$ is one-to-one:

STEP-1. Set $M_j = \{i : s(i) = j\}$ for $j = 1, ..., n$, and take an arbitrary j from $\{j : |M_j| > 1\}$ (to move one in M_j to the next best match in either the left or right side).

STEP-2. Let $Z_j \equiv \{k : k < j; |M_k| = 0\}$. If $Z_j = \emptyset$ set $L = 0$ and $l = s$; otherwise set $L = 1$ and define

$$l(i) = \begin{cases} \max Z_j & \text{if } i = \min\{M_{\max Z_j + 1}\}; \\ s(i-1) & \text{if } \min\{M_{\max Z_j + 1}\} < i \le \min M_j; \\ s(i) & \text{otherwise.} \end{cases}$$

Define R and $r(i)$ in the same fashion as defining $l(i)$, but to shift to the right.

STEP-3. Set $s = l$ if $Lg(l) > Rg(r)$, and set $s = r$ otherwise.

Acknowledgment

This work is supported by the National Science Foundation Grant DMS-1007678.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc.* B **57** 289-300.

Dempster, A. P. (2008). Dempster-Shafer calculus for statisticians, *Internat. J. of Approx. Reason.*, 48, 265–377.

Edlefsen, P. T., Liu, C., and Dempster, A. P. (2009). Estimating limits from Poisson counting data using Dempster-Shafer analysis. *The Annals of Applied Statistics*, **3** 764-790.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96-104.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151-1160.

Fisher, R. A. (1959). *Statistical methods and scientific inference* (2nd ed.). Hafner Publishing Co. New York.

Glaeser, R. M. (2008). Cryo-electron microscopy of biological manostructures. *Physics Today*, 48-54.

Liang, F., Liu, C. and Wang, N. (2007) A sequentail Bayesian procedure for identification of differentially expressed genes. *Statistica Sinica*, **12** 571-597.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc.* B **64** 479-498.

Storey, J. D. (2003). The positive false discovery rate: an Bayesian interpretation and the Q-value. *Ann. Statist.* **31** 2013-2035.

Zhang, J. (2010). *Statistical inference with weak beliefs*, PhD thesis, Purdue University.