International Journal of Statistics and Probability; Vol. 2, No. 2; 2013
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

On Multiple Hypothesis Testing Maximizing the Average Power

John Nixon[1]

[1] Agriculture and Agri-Food Canada, Canada

Correspondence: John Nixon, Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon SK S7N 0X2, Canada. Tel: 1-306-956-2844. E-mail: John.Nixon@agr.gc.ca

Received: March 3, 2013 Accepted: April 16, 2013 Online Published: April 27, 2013

doi:10.5539/ijsp.v2n2p112 URL: http://dx.doi.org/10.5539/ijsp.v2n2p112

Abstract

A general theory is described for making decisions as to which of two modelled hypotheses (that can depend on unknown parameters) best fits each of a set of data sets, such that the average power is maximized. Statistical independence between the large number of data sets of the same type is assumed. Therefore error rates can be expressed as proportions and the continuous approach to the data model is used. The framework of decision theory is used and the equivalence between different criteria for optimization is demonstrated. General procedures are shown to satisfy this criterion in the cases when each hypothesis has a finite number of unknown parameters, and when the alternative hypothesis is vacuous. If the null hypothesis is determined by a known distribution of a test statistic, this reduces to using the density of p-values of this test statistic as the final test statistic to rank the data into significance order. For two scenarios, one of three density estimation methods based on the kernel density estimate gave a result almost equivalent in power to the likelihood ratio test that uses full knowledge of the null and alternative models, and compared favourably with the optimal discovery procedure (ODP) and its iterated variant. For genetic expression data from microarrays and more recently RNA-Seq experiments where the data for different genes are not generally independent, it is suggested to use this technique with the p-values from methods such as Surrogate Variable Analysis that removes much of the effects of dependence.

Keywords: multiple hypothesis testing, two groups model, maximum average power, MAP test, multiple testing procedures

1. Introduction

There is continuing interest in multiple hypothesis testing procedures resulting from the desire to maximize efficiency in the simultaneous testing of large numbers of data sets such as microarray data and more recently RNA-Seq data (Wang, Gerstein, & Snyder, 2009) that after a lot of pre-processing (Givan, Bottoms, & Spollen, 2012) give simultaneously the expression levels of a very large number of genes from RNA samples. Such experiments are capable of extracting information about how much each gene is involved (i.e. transcribed and potentially translated into a functional protein) in different organisms and tissues and under different experimental conditions, including for example many applications in crop and human disease. Genes of interest for further study show unusual expression patterns indicating a possible relevance to the study. Probability models for these patterns underlie the statistical analysis used in the search for such genes. Much recent theoretical work involves correction for effects not explicitly modelled that cause correlation among the data for individual tests (Leek & Storey, 2007; Leek & Storey, 2008; Lunceford et al., 2011; Chakraborty, Datta, Somnath, & Datta, Susmita, 2012), while the simpler problem of handling independent tests (Storey, 2007; Hwang & Liu 2010), does not seem to have been fully explored in its practical implementation when one or both hypotheses have unknown (hyper)parameters (Nixon, 2012). This may be because the perception of the need to deal with dependence makes such a study almost irrelevant. However it has been shown theoretically that data sets after the extraction of the surrogate variables are independent (Leek & Storey, 2008) undermining this perception and in fact demonstrating that a thorough analysis of the simpler case where the separate data sets are independent is actually of fundamental importance.

In this paper multiple testing procedures are analysed assuming no statistical dependence amongst the data for the individual tests to develop optimal procedures of this type. It is also assumed that the dataset is very large and can be modelled as continuously infinite so that the distributions of test statistics and p-values will be assumed to be continuous i.e. have probability densities, and all sums are represented as integrals, so for example the data set D is defined by the number m and density $t(x)$ of points in x space i.e. the data space for a single data set. This makes the notation relatively simple. The view is also taken that a multiple hypothesis testing problem can be regarded

as a model to be fitted to the entire data set and multiple hypothesis testing procedures (MTP's) are model fitting procedures (MFP's) applied to a model that describes multiple hypothesis testing (MHT). Such procedures will be called solutions of the model.

In Section 2 the probability models to be used are introduced and in Section 3 hypothesis testing as implied by the above models i.e. in the "multiple" scenario is developed. Section 4 discusses multiple hypothesis testing from the point of view of decision theory and the assumption that the number of tests (m) is large, introduces some notation, demonstrates the equivalence of different optimisation criteria, and shows that there is always a trade-off between certain expected error rates. Section 5 explains the relationship between Storey's (2007) ODP in general and tests maximizing the average power (MAP tests). In Section 6 arguments are given establishing multiple testing procedures appropriate for different probability models. The procedures themselves are described step by step in Section 7 and the important special case when there is only information about the null hypothesis available, leads to the surprising conclusion in Section 8 that the density of p-values from these tests is the most efficient final test statistic. The remainder of the paper discusses the implementation of this and the results showing that the density of p-values can generate a very efficient testing procedure when the alternative model is only known empirically.

2. Probability Models and Data Structures

As in all the general models in this paper, the symbols denoting variables will be assumed to have as many dimensions as needed in any specific case, for example the variable x will represent the collection of all measured variables including any covariates.

Apart from simply adding more parameters to a model for a data set to make it potentially fit better, there is a basic way in which complex models can be built from simpler ones and is related to complex data sets that have a structure that is essentially a multiple replication of the structure of a simpler data set. If the simpler data set has been modelled with several parameters then the entire data set should be modelled by the joint distribution of these parameters (whether they be continuous or discrete). If this model itself involves parameters they are sometimes known as hyperparameters, and the hyperparameters describe the entire data set. It is here implicit that the separate data sets would be independently generated from this model according to the joint distribution of the parameters. If however this is not so, modelling on another level is involved that is beyond the scope of this paper.

The simplest example of these ideas is as follows. Suppose there are two probability models for some phenomenon A and B, each having parameters to be fitted, a model C could specify that there is a probability π_0 that model A holds, and a probability $1 - \pi_0$ that model B holds, where π_0 is the extra parameter to be fitted in model C that describes the distribution of the binary variable in the entire dataset. Model C would be appropriate for a data set that consisted of many instances of data to which either model A or model B could be applied. In this case, models A and B would be said to be submodels of the composite model C. An MFP for model C should include an estimate of the partition of the data set into two parts each of which fits one of the submodels A and B i.e. π_0 as well as the parameters in A and B. The two-groups model is specified by the density

$$\pi_0 f(x) + (1 - \pi_0)g(x), \tag{1}$$

which defines the term "multiple hypothesis testing" as used in this paper. The rest of this paper will be concerned only with analysis of models of this sort. The following forms of this model will be considered:

1) The densities $f(x)$ and $g(x)$ are uniquely defined.

2) The densities $f(x)$ and/or $g(x)$ have known functional forms with unknown parameters or are unspecified (not both f and g obviously). Here f and g are here not dependent on the test instance i.e. i.

3) The densities $f(\cdot)$ and $g(\cdot)$ are dependent on the test being carried out, and are then uniquely defined so they are denoted by $f_i(x)$ and $g_i(x)$ respectively for $1 \leq i \leq m$.

For each of these cases the assignment of the data to the two hypotheses may be (A) known or (B) unknown.

In case A, an analysis can only determine the order in which the assignment of the data to the hypotheses may be in error i.e. which data points x seem to be most unlikely to have come from the model they actually came from, and order the data most unlikely first. In the practical case B, not only is the partition of the data between the hypotheses estimated but the parameter π_0, which is the proportion of null hypotheses (that come from $f(x)$ or $f_i(x)$) in the data, is an unknown that must be estimated from the data. For the models where the assignment to the hypotheses is known, the number of null hypotheses ($m\pi_0$) and alternative hypotheses ($m - m\pi_0$) are also known.

Clearly not all of these models may be useful practically particularly models with known the assignment of the data to the hypotheses, but they are instructive when developing statistical procedures because the relevant methods used for deciding whether the models fit the data and if so estimating their free parameters are closely related.

The models of type (3) reduce to models of type (1) when all the densities are the same, and models of type (2) may be considered to reduce to models of type (1) when the number of free parameters is zero. Further, models of type B reduce to models of type A when the assignment of the data to the hypotheses is specified.

3. Hypothesis Testing

3.1 Single Hypothesis Testing

In the simplest hypothesis testing situation there are two normalised densities $f(x)$ and $g(x)$ representing two hypotheses ("simple" hypotheses are represented by fixed probability density functions that have no parameters as in the case here), the null hypothesis H_0 and the alternative hypothesis H_1 respectively, and an observation x. The object is to determine which model is most appropriate in the light of the data x on the assumption that one or other of these models is the origin of the observation. Analysis shows that there is always a trade-off between the two types of errors: type I, choosing $g(\cdot)$ when $f(\cdot)$ is correct and type II, choosing $f(\cdot)$ when $g(\cdot)$ is correct. Minimizing the probability of errors of type II for a given value of the probability of an error of type I (or vice versa) occurs when the decision is made based on the ratio of the probability densities as the Neyman Pearson (NP) lemma states (Dudewicz & Mishra, 1988). The critical value of the ratio is determined by which point on the trade-off curve is chosen. In the case when one or both hypotheses have parameters to be fitted (composite hypotheses), this can be done by maximum likelihood before invoking the NP lemma.

3.2 Multiple Hypothesis Testing

Now suppose that m such observations x_i have been made. This situation is quite different, the multiple testing scenario, where many statistical tests have to be carried out, one for each data point. Here there are models on two levels, instances of the simple model describing single data points x_i, and another model describing the distribution of parameter values in this simple model (that could specify fixed values or be another distribution with known or unknown parameter values). The two-groups model with simple hypotheses has two submodels that can be represented by an overall model for the data distribution $t(x)$ of the form given by Equation (1) where π_0 is the probability of the null hypothesis i.e. x is drawn from the distribution $f(x)$, and $\pi_1 = 1 - \pi_0$ is the probability of the alternative hypothesis, which is given by the distribution $g(x)$. In the general case to be considered with composite hypotheses, the functions f and g will be dependent on parameters θ and ϕ (that can be multidimensional) to be estimated from the data i.e. the overall model is

$$t(x) = \pi_0 f(x, \theta) + (1 - \pi_0) g(x, \phi). \tag{2}$$

Therefore one can write

$$t(x) = \frac{\text{Number of data points in } dx}{m \times dx} \tag{3}$$

where the total number of points is $\#\{D\} = m$

$$\int dx t(x) = 1.$$

If $w(x)$ is the test statistic to be derived, (the smaller $w(x)$ is the more significant the data x is) the condition that the p-values are uniformly distributed for the set of null hypotheses is consistent with defining the p-values by

$$p(x) = \int_{w(x') \le w(x)} dx' f(x', \hat{\theta}). \tag{4}$$

To see this note that Equation (4) implies that $p(x_1) = p(x_2)$ for any two points x_1 and x_2 such that $w(x_1) = w(x_2)$, and if $w(x)$ increases so does $p(x)$. Therefore a relationship of the form $p(x) = h(w(x))$ holds where $h(.)$ is a monotone increasing function. Thus the condition determining the range of integration $w(x') \le w(x)$ is equivalent to $h(w(x')) \le h(w(x))$ and to $p(x') \le p(x)$. So Equation (4) can be written as

$$p(x) = \int_{p(x') \le p(x)} dx' f(x', \hat{\theta}). \tag{5}$$

By the definition of H_0, $P_{H_0}(s < p(x) \leq s + ds) = \int_{s<p(x')\leq s+ds} dx' f(x', \hat{\theta})$, and by (5) this reduces to ds. Therefore the distribution of $p(x)$ under the null hypothesis H_0 is uniform as required.

The p-value to be associated with a single data point x will in general depend on the entire data set via $\hat{\pi_0}$ and $\hat{\theta}$. For simplicity of notation this has been left implicit on the left hand sides of Equations (4) and (5) and this will be done throughout the paper for p-values, q-values and test statistics.

4. The Decision Theoretical View of Multiple Testing Procedures (MTP's) and the Equivalence of Optimization Criteria

The results of this section assume that the hypotheses are simple, or the parameter values (θ and ϕ) can be considered as fixed while considering the relationship between the error frequencies n_{01} and n_{10} defined below. This is related to the optimal multiple testing procedure which first fixes values of both these parameters (see Section 7).

In Table 1, π_0 and π_1 should be considered fixed as different tests are considered giving different values of n_{ij}. There are therefore 2 degrees of freedom in this table and each such table is defined by the point (n_{01}, n_{10}) representing the two error frequencies (Figure 1).

Table 1. General notation for the expected relative frequencies i.e. probabilities of different outcomes of the multiple testing procedure

	Test for null hypothesis		
True situation	Accept	Reject	Total
Null	n_{00}	n_{01}	π_0
Alternative	n_{10}	n_{11}	π_1
Total	n_0	n_1	1

After carrying out the tests with $w(\cdot)$ as the test statistic and with the data point x just significant, the four outcomes in Table 1 have the definite probabilities:

$$n_{00} = \pi_0 \int_{w(x')>w(x)} dx' f(x', \theta) \qquad n_{01} = \pi_0 \int_{w(x')\leq w(x)} dx' f(x', \theta)$$
$$n_{10} = \pi_1 \int_{w(x')>w(x)} dx' g(x', \phi) \qquad n_{11} = \pi_1 \int_{w(x')\leq w(x)} dx' g(x', \phi) \tag{6}$$

These could be estimated by simulation for example by simulating m data points in each of d data sets, where d is a suitable large number and using $w(x)$ as the cut-off for the statistic $w(\cdot)$. In the situation in which m is very large, the value of d will make practically no difference to the probabilities. In each simulated data set each point is null with probability π_0, (then an instance of the distribution with density $f(\cdot)$ is simulated), and alternative with probability $\pi_1 = 1 - \pi_0$ (then an instance of the distribution with density $g(\cdot)$ is simulated). Therefore as x takes on each possible value in turn, the points (n_{01}, n_{10}) trace out a curve on Figure 1 determined by the test statistic $w(\cdot)$ which characterises the MTP.

The performance of any MTP could be compared with the following trivial procedure that does not depend on the data: Accept H_0 (the null hypothesis $f(\cdot)$) with probability λ and reject it with probability $1 - \lambda$. Then $n_{10} = \pi_1\lambda$ and $n_{01} = \pi_0(1 - \lambda)$. Elimination of λ gives $n_{10}\pi_0 + n_{01}\pi_1 = \pi_0\pi_1$ which is the line joining $(0, \pi_1)$ and $(\pi_0, 0)$ shown in Figure 1. Therefore any MTP should have a risk point (n_{01}, n_{10}) that is not above this line otherwise each of the errors n_{01} and n_{10} can be reduced keeping the other fixed by using the trivial MTP with an appropriate value of λ. Further, for any two MTP's a and b with different risk points below this line, it is possible to define an MTP (MTPλ) as MTPa with probability λ and MTPb with probability $1 - \lambda$, and the risk point P, for MTPλ is the fraction $1 - \lambda$ along the line from MTPa to MTPb. Therefore the optimal MTP with the same value of n_{01} as P must have n_{10} less than or equal to n_{10} for P. This argument shows that optimal MTP's must fall on a convex curve (i.e. with second derivative $\frac{d^2 n_{10}}{dn_{01}^2} \geq 0$) in (n_{01}, n_{10}) space representing a trade-off between the two error rates.

If $n_{01} = 0$ i.e. every null hypothesis instance in the data is not rejected, then because the statistical test depends on a probability being less than a threshold, the only way this can be certain is if for every test the null hypothesis is not rejected. This implies that $n_{11} = 0$ so $n_{10} = \pi_1$. Likewise if $n_{10} = 0$, for every alternative hypothesis instance the null hypothesis is rejected. Then for all tests, H_0 is rejected i.e. $n_{00} = 0$ and $n_{01} = \pi_0$ so the points $(0, \pi_1)$ and

$(\pi_0, 0)$ are always on the optimal curve. Two error frequencies $\alpha = \frac{n_{01}}{\pi_0}$, which is the fraction of null hypothesis instances in the data that are rejected in error is the type I error rate, and $\beta = \frac{n_{10}}{\pi_1}$, which is the fraction of alternative hypothesis instances in the data that are not rejected is the type II error rate. These two error rates play a role analogous to the case of a single statistical test. Two further error rates are the 'false discovery rate' (FDR) (Storey & Tibshirani, 2003) given by

$$q = \frac{n_{01}}{n_{01} + n_{11}} = \frac{n_{01}}{n_{01} - n_{10} + \pi_1}, \tag{7}$$

which is the fraction of rejections of the null hypothesis that are in error, and the 'missed discovery rate'(MDR) (Storey, 2007) is the fraction of acceptances of the null hypothesis that are in error, and is given by

$$s = \frac{n_{10}}{n_{10} - n_{01} + \pi_0}. \tag{8}$$

The error rates q and s are introduced here because criteria involving them were suggested for defining optimization of MTP's. Storey (2007) used the criterion that MTP's should be optimized such that the expected number of true positives (ETP = $n_{11}m$) should be maximized for each fixed expected number of false positives (EFP = $n_{01}m$). This is referred to as maximizing the average power (MAP) and is the same as maximizing n_{11} for fixed n_{01} i.e. minimizing n_{10} for fixed n_{01} i.e. minimizing β for fixed α. The second criterion suggested (the MDR criterion) was that the MDR must be minimized for each fixed value of FDR, i.e. s must be minimized for fixed q. A third criterion (criterion II) was suggested (Hwang & Liu, 2010) in which EFN/(EFN+ETP) i.e. $\frac{n_{10}}{\pi_1}$ is minimized for fixed FDR i.e. q, where EFN is the expected number of false negatives = $n_{10}m$. Finally, the criterion used in this paper for deriving the optimal MTP is to minimize q for fixed n_1. Because $n_{10} = n_{01} + \pi_1 - n_1$, this criterion is also easily illustrated in Figure 1 as are the others. In fact it turns out that all these four criteria are equivalent i.e. equivalent to the MAP criterion.

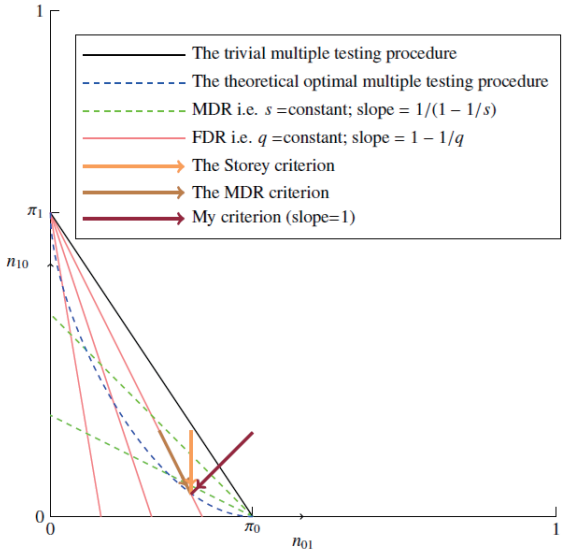


Figure 1. Different optimization criteria for multiple testing procedures that turn out to be equivalent. In this illustration, $\pi_0 = 0.4$ and $\pi_1 = 0.6$

Lines of constant q and s can be considered in the space of points (n_{01}, n_{10}) which can be expressed by rearranging (7) and (8) respectively to be solved for n_{10}:

$$n_{10} = n_{01}\,(1 - 1/q) + \pi_1 \tag{9}$$

$$n_{10} = (\pi_0 - n_{01})\left(\frac{s}{1-s}\right). \tag{10}$$

Using (9) to eliminate n_{10} in (8) gives $s = \frac{n_{01}(1-1/q)+\pi_1}{1-n_{01}/q}$ from which $\frac{ds}{dn_{01}}|_q = \frac{q-\pi_0}{q(1-n_{01}/q)^2}$ follows. Then $\frac{ds}{dn_{01}}|_q = 0 \Leftrightarrow q = \pi_0 \Leftrightarrow n_{10} = n_{01}\,(1 - 1/\pi_0) + \pi_1$ which is the line passing through $(0, \pi_1)$ and $(\pi_0, 0)$ and by examining the sign of $z = \frac{ds}{dn_{01}}|_q$ near $(0, 0)$ where $q \to 0$ it follows that $z < 0$ for all points below that line (which is the only region

of interest). From (9) it follows that lines of constant q all pass through $(0, \pi_1)$ and have slope $1 - 1/q < 0$, and from (10) the lines of constant s all pass through $(\pi_0, 0)$ and have slope $s/(s-1)$ so for fixed q by (9), minimizing n_{10} is equivalent to maximizing n_{01} is equivalent to minimizing s. Therefore criterion II is equivalent to the MDR criterion and the MDR criterion of optimal MTP's is that for fixed q, the point with the largest n_{01} i.e. with the smallest n_{10} must be chosen.

The convexity property of the optimal curve in Figure 1 allows an inequality to be derived as follows. Using $\frac{dq}{ds}$ and $\frac{dn_{10}}{dn_{01}}$ to denote the derivatives along the theoretical optimal curve representing the optimal MTP in Figure 1, and the subscript P to denote an arbitrary point on this curve

$$\left.\frac{dn_{10}}{dn_{01}}\right|_P - \frac{1}{n_{01}(P)} \int_0^{n_{01}(P)} dn_{01} \frac{dn_{10}}{dn_{01}} = \frac{1}{n_{01}(P)} \int_0^{n_{01}(P)} dn_{01} \left(\left.\frac{dn_{10}}{dn_{01}}\right|_P - \frac{dn_{10}}{dn_{01}} \right). \tag{11}$$

The term in parentheses is

$$\int_{n_{01}}^{n_{01}(P)} dn_{01} \frac{d^2 n_{01}}{dn_{01}^2} \geq 0 \tag{12}$$

provided $n_{01} < n_{01}(P)$ because of the convexity property of the theoretical optimal curve. Therefore

$$\left.\frac{dn_{10}}{dn_{01}}\right|_P \geq \frac{1}{n_{01}(P)} \int_0^{n_{01}(P)} dn_{01} \frac{dn_{10}}{dn_{01}} = \frac{n_{10}(P) - \pi_1}{n_{01}(P)} \tag{13}$$

This shows that the optimal curve (that must pass through $(0, \pi_1)$) always has a slope of smaller magnitude at a point than the line of constant q that intersects it at that point. Therefore the Storey (2007) criterion is equivalent to the MDR criterion for optimal MTP's. These criteria are also equivalent to the criterion that q must be minimized for all tests with the same value of n_1 because the slope of the lines of constant n_1 is 1 and so these lines intersect the optimal curve just once. Likewise

$$\left.\frac{dn_{10}}{dn_{01}}\right|_P - \frac{1}{\pi_0 - n_{01}(P)} \int_{n_{01}(P)}^{\pi_0} dn_{01} \frac{dn_{10}}{dn_{01}} \leq 0 \tag{14}$$

so

$$\left.\frac{dn_{10}}{dn_{01}}\right|_P \leq -\frac{n_{10}(P)}{\pi_0 - n_{01}(P)} \leq 0 \tag{15}$$

which shows that optimal curve always has a slope of higher magnitude (it is < 0) at a point than the curve of constant s that intersects it at that point. In order to relate the inequalities (13) and (15) to the possible range of values of dq/ds on the optimal curve, the relationship between these derivatives is needed which is as follows:

$$\frac{dq}{ds} = \frac{dq}{dn_{01}} \Big/ \frac{ds}{dn_{01}} = \left(\frac{\partial q}{\partial n_{01}} + \frac{\partial q}{\partial n_{10}} . \frac{dn_{10}}{dn_{01}} \right) \Big/ \left(\frac{\partial s}{\partial n_{01}} + \frac{\partial s}{\partial n_{10}} . \frac{dn_{10}}{dn_{01}} \right) \tag{16}$$

After inserting the partial derivatives of q and s with respect to n_{01} and n_{10} and using the definitions in Table 1 this becomes

$$\frac{dq}{ds} = \left(\frac{n_0}{n_1} \right)^2 . \left(\frac{\pi_1 - n_{10} + n_{01} \frac{dn_{10}}{dn_{01}}}{n_{10} + (\pi_0 - n_{01}) \frac{dn_{10}}{dn_{01}}} \right). \tag{17}$$

The function $x \to \frac{A+Bx}{C+Dx} = \frac{B}{D} + \frac{AD-BC}{D(C+Dx)}$ is monotone decreasing if x does not pass through the singular point $-\frac{C}{D}$ and $(AD - BC)$ is greater than 0. To apply this result note that $A = \pi_1 - n_{10} \geq 0$, $B = n_{01} \geq 0$, $C = n_{10} \geq 0$, $D = \pi_0 - n_{01} \geq 0$ and $AD - BC = \pi_0 \pi_1 - n_{10} \pi_0 - n_{01} \pi_1 \geq 0$ in the lower left triangular region of Figure 1, and the upper bound (15) for $\frac{dn_{10}}{dn_{01}}$ coincides with the singular point $-C/D = \frac{-n_{10}}{\pi_0 - n_{01}}$. Therefore the RHS of (17) is decreasing with $\frac{dn_{10}}{dn_{01}}$ in its allowed range

$$\frac{n_{01} - \pi_1}{n_{01}} \leq \frac{dn_{10}}{dn_{01}} < \frac{-n_{10}}{\pi_0 - n_{01}}, \tag{18}$$

and this range is mapped to

$$\left(\frac{n_0}{n_1} \right)^2 \left(\frac{\pi_1 - n_{10} + n_{01} \left(\frac{n_{10} - \pi_1}{n_{01}} \right)}{n_{10} + (\pi_0 - n_{01}) \left(\frac{n_{10} - \pi_1}{n_{01}} \right)} \right) \geq \frac{dq}{ds} \tag{19}$$

i.e.

$$\frac{dq}{ds} \leq 0, \tag{20}$$

which shows that there is in general a trade-off between FDR and MDR as indicated in Figure 1. There is also a trade-off between α and β as is indicated in Equation (15). For the hypothetical tests mentioned by Hwang and Liu (2010) $\pi_0 = 2/21$ and $\pi_1 = 19/21$ and test 1 has $n_{01} = 2/21$ and $n_{10} = 1/21$ and test 2 has $n_{01} = 1/21$ and $n_{10} = 10/21$. Therefore $AD - BC < 0$ in each case, so the risk points for both tests are outside the region of interest i.e. above the line joining $(0, 19/21)$ and $(2/21, 0)$. Therefore both tests result from test procedures that can be bettered by a randomizing test procedure that does not depend on the data, and consequently no conclusions regarding optimization criteria should be drawn from them.

5. Storey's ODP and MAP Tests

Storey (2007) proposed one of the criteria for optimizing an MTP which were shown above to be equivalent to each other. He also proved that this optimization is satisfied by the "optimal discovery procedure" (ODP) which is an MTP for a class of models based on the NP lemma. The ODP maximizes the average power (MAP) for the two groups model Equation (1) involving any two hypotheses that are represented by known probability functions $f(\cdot)$ and $g(\cdot)$ of the data. Tests based on this principle are often referred to as MAP tests. However as Hwang and Liu (2010) pointed out, Storey (2007) did not model the variation in the variances, only the variation in the means in connection with the application of the ODP to microarray data. When Hwang and Liu (2010) did this they derived MAP tests that are more robust, and approximate MAP tests appropriate for the microarray analysis problem, in particular the useful approximation that is numerically almost exact, the Fss test that can be computed very fast. Their work illustrates the general principle that most powerful statistical procedures result from correctly modelling the data. These MAP tests involve hypotheses that are fixed because they involve fixed probability distributions of the parameters appropriate to each test and can therefore be expressed as integrals. The next section describes extensions of these arguments to models involving undetermined parameters (sometimes referred to as hyper-parameters if they refer to the whole model rather than parameters that could be estimated from the individual data sets denoted by x here) other than π_0. Established techniques for solving this kind of models involve estimation of the parameters by maximum likelihood (Storey, 2007), or the use of iterative techniques (Nixon, 2012).

6. Arguments Establishing MTP's for Different Models

Storey (2007) made a fundamental advance in the theory of multiple hypothesis testing by showing that the optimal solution for models of types 3A and 3B (Section 2) is essentially unique. These are generalisations of the Neyman Pearson (NP) lemma whose proofs rely essentially on this lemma.

Specifically, he showed for model 3A that to use the MAP criterion i.e. in order to maximize n_{11} for fixed n_{01} the test statistic can be chosen as $\sum_{i=m_0+1}^{m} g_i(x) \Big/ \sum_{i=1}^{m_0} f_i(x)$, where without loss of generality the null hypotheses are data points 1 to m_0, and the alternative hypotheses are data points $m_0 + 1$ to m. Storey (2007) also showed, in the case that the assignment of the hypotheses to the data is not known i.e. the model is of type 3B by a very similar argument, that under the same optimization, the test statistic can be chosen as $\sum_{i=1}^{m} g_i(x) \Big/ \sum_{i=1}^{m} f_i(x)$. The word "can" is used instead of "must" because any other statistic giving the same order of the test statistics hence the same order of the p-values can be used, where the p-values are calculated according to Equation (4) with $w(x) = \sum_{i=1}^{m} f_i(x) \Big/ \sum_{i=1}^{m} g_i(x)$. (Note that the statistic $w(x)$ is smaller the more significant it is, which is the reverse convention to the one used by Storey (2007) in his test statistics. Therefore the reciprocal of his statistics may be taken if the values are to be interpreted in the reverse order i.e. smallest values are most significant.) These statistics are referred to as the ODP's for their respective models. For models of types 1A and 1B, Storey's (2007) test statistic becomes

$$\sum_{i=m_0+1}^{m} g(x) \Big/ \sum_{i=1}^{m_0} f(x) = (m - m_0)g(x)/m_0 f(x),$$

which is equivalent to $\frac{g(x)}{f(x)}$. These results are essentially the same as the NP lemma.

Proceeding systematically to develop MTP's note that for model 1A, the two subsets of the data corresponding to each hypothesis can be tested against their corresponding submodel to find out how good the fits are. If they are satisfactory, the test statistic and p-values can be calculated for each data point, giving the order of significance of the data. This only serves to show which data points that are actually null hypotheses look most likely to have been alternatives and vice versa, and to order the data accordingly. This is obviously of limited practical value.

For model 1B, the statistic $\frac{g(x)}{f(x)}$ and p-values could be calculated for each data point x. After a p-value cut-off λ has been selected, the data can be partitioned into two parts leading to an estimate of π_0 and two goodness-of-fit (GOF) tests one for each of the probability densities. If a value of λ can be found that satisfies both these GOF tests then that estimate of π_0 can be selected. This procedure does have the consequence that the p-value distributions of the data subsets from which $f(\cdot)$ and $g(\cdot)$ are estimated each have a sharp cut-off, and also the result is dependent on the unknown parameter λ. Therefore in this case, with the assignment of the data to the hypotheses unknown, it seems more appropriate to just compare the overall density which is $\hat{\pi_0} f(x) + (1 - \hat{\pi_0}) g(x)$ with the observed distribution of the entire data set D, to determine the best estimate of π_0 and how well this model fits.

If the GOF test shows a fit for some value of $\hat{\pi_0}$ then that estimate of π_0 can be taken, but if there is more than one plausible region of $\hat{\pi_0}$ it would probably indicate that the model is not satisfactory. If the fit is satisfactory, the obvious question now is to estimate which data D_0 come from $f(x)$ and which data D_1 come from $g(x)$, where $D = D_0 \cup D_1$. The obvious estimate of D_0 is the first $\hat{\pi_0}$ fraction of the m data taken in decreasing order of p-value.

Because the cut-off $\hat{\pi_0}$ depends on the data set as a whole, the statistical test to be carried out for each data point clearly has an interpretation that can depend on properties of the whole data set, i.e. the interpretation of the statistical test for data point i could have been different if the same data x_i was part of a different data set. This should not be seen as a paradox, it is simply a consequence of fitting a model with an unknown parameter (in this case π_0) that describes the whole data set.

The only difference between models of type (2) and models of type (1) is that in the former one or both hypotheses are arbitrary or have parameters and therefore have to be estimated from the data. For model 2A, again both GOF tests have to be done and are used to fit the parameters or estimate the distribution of the data non-parametrically. In the latter case this might involve data smoothing techniques (see for example Simonoff, 1996). If the fits are satisfactory, the test statistic can be written as $\frac{\hat{g}(x)}{\hat{f}(x)}$, which can be used to sort the two data subsets into order of significance and generate p-values. This is of limited practical value as in model 1A.

Model 2B is the most practical case because the assignment of the hypotheses to the data is not known as well as the precise functional forms of one or both hypotheses. This case obviously presents the most challenges. Again a GOF test could be done to try to fit $\hat{\pi_0} \hat{f}(x) + (1 - \hat{\pi_0}) \hat{g}(x)$ to the observed distribution of the data, to obtain estimates $\hat{\pi_0}$, $\hat{f}(x)$ and $\hat{g}(x)$. If it is successful, the test statistic $\frac{\hat{g}(x)}{\hat{f}(x)}$ can be used to sort the data into order of significance as for model 1B. There is clearly a problem here because this test should determine which data should contribute to the estimation of $f(\cdot)$, and which should contribute to the estimation of $g(\cdot)$, both of which are involved in the test statistic, hence there is circularity here that suggests using an iterative algorithm.

In contrast with the models of types (1) and (3), there is no theoretical reason to suppose that such a converged solution obtained by an iterative algorithm for models of type 2B is optimal in any sense. This is a consequence of the estimation required to obtain the density functions of the models. A similar conclusion was found in Nixon (2012) where the calculation was done by letting the weight w_i with which data x_i contributes to the null hypothesis be given by $w_i = \left\{ \begin{array}{l} 0 \text{ if } p(x_i) \leq \lambda \\ 1 \text{ if } p(x_i) > \lambda \end{array} \right\}$ where $p(x_i)$ is the p-value for data point i in the previous cycle, and iterating the procedure to convergence, then repeating this for a range of values of the free parameter λ. It was found that the power expressed by the fraction of the data declared significant n_1 for a given q-value $q = \frac{n_{01}}{n_1}$ varies with parameters (λ, m_0) and it was not clear how this can be optimized. This calculation gives rise to artificial cut-off's in the p-value distributions of the data subsets suggesting that there is a better solution based on the optimization criteria discussed above. An obvious approach is to try to minimize $q = E\left\{\frac{\#\{w(x)=0 \ \& \ x \text{ is from } f(\cdot)\}}{\#\{w(x)=0\}}\right\} = \frac{n_{01}}{n_1}$, which is the probability of errors amongst those x for which the null hypothesis is rejected, conditional on (i) a fixed value $n_1 = E\left\{\frac{\#\{w(x)=0\}}{m}\right\}$ and (ii) the GOF test being satisfactory. For the purpose of this optimization, the function w: $D \rightarrow \{0, 1\}$ can be varied. Here the function $w(\cdot)$ has the interpretation as follows: $w_i \equiv w(x_i) = 0$ means the null hypothesis for data x_i is rejected, and $w(x_i) = 1$ means the null hypothesis for data x_i is not rejected. However while implementing the approach above for a hypothetical continuous theoretically infinite ($m \rightarrow \infty$) data set, clearly it makes no sense that w is an arbitrary function of x. Instead the transition point where w changes from 0 to 1 should be determined by a continuous function of x, or for greater flexibility w should be continuously varying with x. Therefore $w(\cdot)$ will be allowed to take any values in the interval $[0, 1]$.

The function $w(x)$ is to be thought of as the weight or probability associated with data x in the estimation of f (the null hypothesis) and $1 - w(x)$ is the weight or probability associated with x in the estimation of g (the alternative hypothesis) i.e. $w(x)$ is the probability that data point x is from the null hypothesis. This ensures that all the data

are represented equally weighted, which could be adjusted if necessary as Storey (2007) suggested, if for example some data were known to be more reliable than others. The function $w(x)$ is also a measure of significance of the data x, i.e. the smaller $w(x)$ is the more significant the data x is.

With this change to $w(\cdot)$, for an arbitrary data point x, q must now be rewritten as $q(x) = E\left\{\frac{\#\{y|w(y)<w(x)\ \&\ y \text{ is from } f(\cdot)\}}{\#\{y|w(y)<w(x)\}}\right\}$ and $n_1(x) = E\left\{\frac{\#\{y|w(y)<w(x)\}}{m}\right\}$. Both $q = \frac{n_{01}}{n_1}$ and $n_1 = n_{01} - n_{10} + \pi_1$ are determined in terms of the cut-off point x, i.e. q is determined as a function of n_1; this could be plotted on Figure 1 after elimination of n_1 (π_1 is considered as a known constant). Now $w(\cdot)$ should be varied to as to simultaneously minimize q for each fixed value of n_1 if possible. For simplicity this will initially be applied to models of type 1B under the assumption that this model, as a whole, fits the data. It will turn out to be an alternative argument equivalent to that given by Storey (2007), but is instructive in developing the final MTP for model 2B.

In the following lines s represents $w(x)$ for an arbitrary data point x in the hypotheses listed in order of significance with the most significant first. The variable y represents another arbitrary data point and q, now expressed as a function of s, is

$$q^*(s) = E\left\{\frac{\#\{y|w(y) < s\ \&\ y \text{ is from } f(\cdot)\}}{\#\{y|w(y) < s\}}\right\}. \tag{21}$$

This can be written as

$$q^*(s) = \frac{\text{Prob}\{y \text{ satisfies } w(y) < s\ \&\ y \text{ is from } f(\cdot)|y \text{ is from distribution } t(\cdot)\}}{\text{Prob}\{y \text{ satisfies } w(y) < s|y \text{ is from distribution } t(\cdot)\}}. \tag{22}$$

In order to write this in terms of integrals, the following result will be used. The probability that Q holds if (1) the probability of Q depends on x and (2) x is drawn randomly from the probability distribution with density $t(x)$ is

$$\begin{array}{rcl} \sum_i dx_i \frac{\text{Prob}\,(Q\ \&\ x\in dx_i)}{dx_i} & = & \sum_i dx_i \frac{\text{Prob}(Q|x\in dx_i)\text{Prob}(x\in dx_i)}{dx_i} \\ & = & \int dx \text{Prob}\,(Q|x) \times t(x). \end{array} \tag{23}$$

Now the denominator of (22) can be written as

$$\begin{array}{rcl} \text{Prob}\{y \text{ satisfies } w(y) < s\} & = & \int dyt(y)\text{Prob}(w(y) < s) \\ & = & \int dyt(y) \begin{cases} 1 & w(y) < s \\ 0 & \text{otherwise} \end{cases} = \int_{w(y)<s} dyt(y), \end{array} \tag{24}$$

and the numerator of (22) after a little tricky manipulation becomes

$$\begin{aligned} & \int dyt(y)\text{Prob}\{w(y) < s\ \&\ y \text{ is from } f(\cdot)\} \\ &= \int dyt(y)P(w(y) < s)P(y \sim f(\cdot)|w(y) < s) \\ &= \int_{w(y)<s} dyt(y)P(y \sim f(\cdot)|w(y) < s) \\ &= \int_{w(y)<s} dyt(y)P(y \sim f(\cdot)) \\ &= \int_{w(y)<s} dyt(y)\gamma(y) \end{aligned} \tag{25}$$

In these equations, the fact that y is randomly selected from $t(\cdot)$ does not have to be spelled out on every line, and restriction of the integral to $w(y) < s$, which is known under the integral sign i.e. its probability is 0 or 1, implies that this does not have to be given as a condition under the integral sign. Here $\gamma(y)$ is the probability that the given value y is from $f(\cdot)$, and model 1B as a whole fits the data (that come from the distribution with density $t(x)$) i.e. $\gamma(x) \equiv \frac{\pi_0 f(x)}{t(x)} = \frac{\pi_0 f(x)}{\pi_0 f(x)+(1-\pi_0)g(x)} = \left(1 + \frac{g(x)}{f(x)}\left(\pi_0^{-1} - 1\right)\right)^{-1}$. From this it follows that $\gamma(x)$ increases monotonically with $\frac{f(x)}{g(x)}$ and

$$q^*(s) = \frac{\int_{w(y)<s} dyt(y)\gamma(y)}{\int_{w(y)<s} dyt(y)}. \tag{26}$$

Now $q^*(s)$ has to be minimized simultaneously for each value of s between 0 and 1 by choosing the function $w(\cdot)$. If $s = 0$ the formula for q becomes 0/0. Suppose s is a small quantity δs then $q^*(\delta s) = \frac{\int_{w(y)<\delta s} dyt(y)\gamma(y)}{\int_{w(y)<\delta s} dyt(y)}$. Both the numerator and denominator refer to the same small region of data space and the value of the quotient is the average of the fixed function $\gamma(y)$ over the region $\{y\colon 0 \le w(y) < \delta s\}$. To make this as small as possible requires that $w(\cdot)$ be such that the smallest value of γ occurs in this region where w is smallest i.e. the region where f/g is smallest. This argument can be extended to other values of s but another simple result is needed first.

Suppose $\frac{A+x}{B+y}$ must be minimized by varying x and y, such that $x > 0$ and $y > 0$ are both much smaller than A and B, which are both fixed and positive. The line $\frac{A+x}{B+y} = c$ can be written as $y = \frac{1}{c}(x + A) - B$. Therefore these lines for different values of c intersect at $x = -A$ and $y = -B$ and have slope $\frac{1}{c}$. Minimizing c means maximizing the slope, so this implies maximizing y/x i.e. minimizing x/y. This fact is obvious from the geometry and is regardless of the precise allowed region of x and y.

Now consider minimizing

$$q^*(s + \delta s) = \frac{\int_{w(x)<s+\delta s} dxt(x)\gamma(x)}{\int_{w(x)<s+\delta s} dxt(x)} = \frac{\int_{w(x)<s} dxt(x)\gamma(x) + \int_{s\le w(x)<s+\delta s} dxt(x)\gamma(x)}{\int_{w(x)<s} dxt(x) + \int_{s\le w(x)<s+\delta s} dxt(x)} \tag{27}$$

assuming that $q^*(r)$ for $0 \le r \le s$ has already been minimized and a region of smallest γ i.e. $\gamma(y) < h$ corresponds to a region of smallest w i.e. $0 \le w(y) < s$. This requires minimizing $\int_{s\le w(y)<s+\delta s} dyt(y)\gamma(y) \Big/ \int_{s\le w(y)<s+\delta s} dyt(y)$, which is the average of $\gamma(\cdot)$ over the region $\{y\colon s \le w(y) < s + \delta s\}$. To minimize this subject to the above constraint needs $w(\cdot)$ to be chosen such that $\gamma(\cdot)$ is next smallest in the region $s \le w(y) < s + \delta s$. Now a region of smallest $\gamma(\cdot)$ is $\{y\colon 0 \le w(y) < s+\delta s\}$. This argument is peculiar because it resembles induction, but the variables are continuous. The conclusion is that the function $w(\cdot)$, that minimises q for each value of s such that $0 \le s \le 1$, is itself such that $w(x)$ is dependent on $\gamma(x)$ only and $w(y)$ is a monotone increasing function of $\gamma(y)$. Therefore to take the data in the order most significant first, the data should be taken in the order given by increasing $w(x)$ i.e. increasing γ, i.e. increasing f/g. This argument shows that Storey's analysis procedure for solving models of type 1B based on the NP lemma (Storey, 2007) is equivalent to the analysis based on simultaneously minimizing q for each fixed n_1 by varying $w(\cdot)$ such that $w(\cdot)$ is continuous and takes any value between 0 and 1. It is interesting to note that this argument does not determine $w(\cdot)$ precisely but this is sufficient for determining the analysis procedure.

Returning to models of type 2B i.e. the model (2), the functions f and g are now dependent on parameters i.e. $f = f(x, \theta)$ and $g = g(x, \phi)$ where θ and ϕ represent sets of parameters whose values are to be determined i.e. f and g are not fixed functions of x and they must be estimated. Therefore, because of its meaning as a weight function, the function $w(\cdot)$ could be associated with two GOF tests i.e. (a) fitting $f(x, \hat{\theta})$ to $\frac{t(x)w(x)}{\hat{\pi_0}}$ (with the best $\hat{\theta}$) and (b) fitting $g(x, \hat{\phi})$ to $\frac{t(x)(1-w(x))}{1-\hat{\pi_0}}$ (with the best $\hat{\phi}$). Thus $w(\cdot)$ needs to be chosen such that both these fits are good. This suggests a combined condition should be optimized in the choice of $w(\cdot)$ and simultaneously an estimate of $\hat{\pi_0}$ needs to be found. There are difficulties with this because (1) the presence of the factor $w(\cdot)$ the likelihood is not strictly a valid concept here because the product $t(x)w(x)$ can have fractional values which should be frequencies for the likelihood to have a meaning and (2) a way of combining the two GOF statistics is needed and an optimization method.

An alternative related approach is the use of (a) to determine $w(x)$ by choosing

$$w(x) = \frac{\hat{\pi_0} f(x, \hat{\theta})}{t(x)}. \tag{28}$$

from which

$$\hat{\pi_0} = \int dxt(x)w(x) \tag{29}$$

follows. Then after substituting for $w(x)$ the GOF test for g can be written as comparing $t(x)$ with $\hat{\pi_0} f(x, \hat{\theta}) + (1 - \hat{\pi_0})g(x, \hat{\phi})$, which is the GOF test for the whole composite model. A satisfactory fit is required for the remainder of this analysis to be valid. Thus in this approach, only one GOF test has to be done for the whole model after simultaneously estimating π_0, θ, and ϕ, which will be typically done by maximum likelihood.

Minimizing q using the argument above with arbitrary fixed estimates of θ, ϕ and π_0 shows that the order the data are declared significant (i.e. are from g) is in increasing order of γ i.e. increasing order of $f(x, \hat{\theta})/g(x, \hat{\phi})$. Now it

is clear from Equation (28) that assuming the model fits,

$$w(x) = \frac{\pi_0 f(x, \hat{\theta})}{\pi_0 f(x, \hat{\theta}) + (1 - \pi_0) g(x, \hat{\phi})} \tag{30}$$

i.e. $w(x) = \gamma(x)$.

7. The Proposed Multiple Testing Procedures

The above arguments suggest that the following procedure (MTP1) should be used to test and fit the model (2) to a data set where f and g are alternative sub models that could apply for each data point x.

(1) Use a fitting procedure to simultaneously search for optimal values of θ, ϕ and π_0 (i.e. their estimates denoted by$\hat{}$) such that the overall model $\pi_0 f(x, \theta) + (1 - \pi_0) g(x, \phi)$ best fits the entire data set. The fitting could be done by maximum likelihood (ML) using the Nelder and Mead (1965) simplex method (also described in Press, Teukolsky, Vetterling, & Flannery, 1997) to search for the optimum point.

(2) Do a GOF test for this model to test that this best fitting model actually fits the data, for example simulate data (for the same number m of x values as in the observed data) based on the model r times, calculate the likelihood for each simulated data set and suppose on b occasions the likelihood is less than or equal to the likelihood of the observed data (in practice the logarithms of the likelihood values would be used). Then the p-value should be estimated as $\frac{b+1}{r+1}$ taking into account the correction for the fact that this Monte Carlo technique generates an exact uniform discrete frequency distribution of b under the null hypothesis (Phipson & Smyth, 2010). This one-sided test is appropriate because simulated data sets based on the model are expected to fit the model better than any other data sets.

(3) If the model fits, choose data x_i as "significant" in the order of increasing $f(x_i, \hat{\theta})/g(x_i, \hat{\phi})$ i.e. smallest values are taken first. Otherwise exit the procedure.

(4) In order to estimate the null data subset, the first fraction $1 - \hat{\pi_0}$ of the data can be assumed to be from g while taking the data in the order of increasing $w(x)$. If required, the following steps may be taken.

(5) For each x calculate the p-value (i.e. the smallest value of $\frac{n_{01}}{\pi_0}$ that can be obtained while calling the data point x_i significant) according to Equation (4) using the discreteness correction (Phipson & Smyth, 2010). This would be typically done using a large simulated population of x_i from the density $f(x, \hat{\theta})$.

(6) Finally, calculate the estimates of the q-values $q(x_i)$ i.e. estimates of the smallest value of FDR = $\frac{n_{01}}{n_1}$ that can be obtained by calling data point x_i significant according to Storey and Tibshirani (2003), which only uses the definition of the FDR, using the estimate $\hat{\pi_0}$ obtained above.

In step 2 of MTP1, the number r of simulations may not need to be very great in practice. For example if the true p-value was 0.5 indicating a good fit, and $r = 10$ was chosen, then there would be only a probability of $1/2^{10} = 1/1024$ that $b = 0$ giving the estimated p-value as 1/11 and necessitating more simulations to estimate it more precisely. Therefore if the true p-value was not close to zero it could be inferred relatively inexpensively, and accurate estimation of its value is unimportant.

MTP1 could be compared with a method (MTP2) which is an obvious modification of the method Nixon (2012) used and which is appropriate for the model (2):

1) Choose λ between 0 and 1.

2) Do an initial test generating p-values for each member x of the dataset D. This splits it into two subsets $D_0 = \{D : p > \lambda\}$ and $D_1 = \{D : p \leq \lambda\}$ which are estimates of the null data and alternative data subsets respectively.

3) Repeat {

   (a) Estimate the parameters θ by fitting the model $f(x, \theta)$ to D_0 and ϕ by fitting $g(x, \phi)$ to D_1 by maximum likelihood.
   (b) Calculate the likelihood ratio test statistic for each data point by $l(x) = g(x, \hat{\phi})/f(x, \hat{\theta})$
   (c) Calculate the p-values as $p(x) = \int_{x':l(x')\geq l(x)} f(x', \hat{\theta}) dx'$
   (d) Redefine D_0 and D_1 as in step 2 using the new p-values.
   (e) Calculate ndiff as the number of hypotheses that have changed status in two consecutive cycles of the procedure

} while ndiff > 0

4) Determine whether the two fits in step (a) are satisfactory as in step 2 of MTP1.

5) If they are both satisfactory, calculate q-values as in step 6 of MTP1 otherwise exit the procedure.

Note that if the null and alternative hypotheses were reversed, then the likelihood ratio statistic would be $l^*(x) = f(x,\hat{\theta})/g(x,\hat{\phi})$ and $p^*(x) = \int_{x':l^*(x')\geq l^*(x)} g(x',\hat{\theta})dx' = \int_{x':l(x')\leq l(x)} g(x',\hat{\theta})dx'$. Similarly to the above argument, p^* increases monotonically with l increasing and p decreases monotonically with l increasing therefore p decreases monotonically with p^* increasing. In other words p and p^* determine complementary subsets of D that are significant as should be the case. So analysing the data this way will give the reverse order of significance of the data. This argument works for MTP1 and MTP2.

For the case when the model for g is saturated $\hat{g}(\cdot)$ is entirely data-dependent and describes the distribution of data that don't fit the model $f(\cdot)$. The fitting is in general non-unique because all the data could be described by an appropriate $g(\cdot)$ i.e. a solution with $\hat{\pi_0} = 0$ always fits perfectly, but with $\hat{\pi_0} > 0$ there could be no solutions, one or more. Therefore a new fitting procedure is needed that modifies the first step and omits the second step of MTP1. Because the model fits perfectly for $\hat{\pi_0} \leq$ this upper limit, the likelihood will be independent of $\hat{\pi_0}$ up to this point, so the parameters are ML estimates and the procedure to be defined below (MTP3) is a limiting case of MTP1.

If the data are first smoothed giving say $t^*(x)$, then because $f \geq 0$, the parameters π_0 and θ must be such that for each data point x

$$t^*(x) \geq \hat{\pi_0} f(x,\hat{\theta}). \tag{31}$$

The estimate $\hat{\pi_0}$ should be as large as possible to be maximally informative and depends on $\hat{\theta}$. Therefore from Equation (31) for fixed $\hat{\theta}$, $\hat{\pi_0} = \inf_x \frac{t^*(x)}{f(x,\hat{\theta})}$, and further, $\hat{\theta}$ should be chosen to maximize $\hat{\pi_0}$ i.e.

$$\hat{\pi_0} = \max_{\hat{\theta}} \left\{ \inf_x \frac{t^*(x)}{f(x,\hat{\theta})} \right\}. \tag{32}$$

(An alternative may be to first fix $\hat{\pi_0}$ then fit θ by maximum likelihood. Do this repeatedly to find the largest value of π_0 that allows an adequate fit.) Because the model fits we can write

$$t^*(x) = \hat{\pi_0} f(x,\hat{\theta}) + (1-\hat{\pi_0})\hat{g}(x) \tag{33}$$

so the test statistic is

$$w(x) = \frac{f(x,\hat{\theta})}{\hat{g}(x)} = \frac{f(x,\hat{\theta})(1-\hat{\pi_0})}{t^*(x) - \hat{\pi_0} f(x\hat{\theta})} \tag{34}$$

from which the p-values can be calculated from Equation (4).

Therefore the modified procedure (MTP3) would be in outline as follows:

1) Use Equation (32) to determine both $\hat{\pi_0}$ and $\hat{\theta}$.

2) Use Equation (34) to determine the test statistic $w(x)$.

3) Use Equation (4) to determine the p-values as in step 5 of MTP1.

4) Estimate the null data subset and q-values according to steps 4 and 6 of MTP1.

8. Applying This Method When the Null Hypothesis Is Not Explicit

Suppose that the null hypothesis is not explicitly given, and is replaced by a statement which implies a known probability distribution of a test statistic, i.e. there is a known procedure for associating p-values (denoted by y) with data such that they are uniformly distributed under the null hypothesis, and the alternative distribution is not known *a priori*. Thus y is one-dimensional and is uniformly distributed in the interval $[0,1]$. Then model (2) applies to the variable y with $f(y) = 1$ and the alternative hypothesis $g(\cdot)$ is not known *a priori* and θ is absent. The above arguments suggest MTP3 applies, so Equation (32) implies

$$\hat{\pi_0} = \inf_{y\in[0,1]} \{\text{estimate of the density of } p\text{-values at } y\}. \tag{35}$$

and the test statistic would be $s = \frac{1-\hat{\pi_0}}{t^*(y)-\hat{\pi_0}}$ which is equivalent to using $t^*(y)$ as the test statistic because they are related inversely by a monotonic function (s decreases as t^* increases). The quantity $t^*(y)$ which in practice is an

estimate of the probability density of the p-values at y could be used as the test statistic when evaluated at each p-value i.e. at each data point. The largest values of this test statistic are now the most significant. After sorting the data into the new order most significant first, the final p-value for the data point x (if needed) corresponding to y is obtained as the fraction of data at least as significant as x according to Equation (4).

9. Methods and Models

In order to test this statistical method, the problem of classifying frequency ratios (Nixon, 2012) where the frequencies are small and the number of instances is large will be considered. The parameters for the null hypothesis are also the same as in Nixon's paper so that the results can be compared. Three simulation models were investigated that only differ in the distribution of N_1 in the alternative data and so the same theoretical result of Section 8 still applies. The model for the simulation is as follows: both the alternative and null data (500 each) have $N = N_1 + N_2$ poisson distributed with a mean of 20, but truncated and renormalised so that N cannot be 0 or 1. The null data are such that given N, N_1 is binomially distributed with expected fraction $N_1/(N_1 + N_2) = 0.3$, but truncated and renormalised such that N_1 cannot be zero or N:

$$f(N_1, N_2) = \text{bin}^T(0.3, N_1, N_1 + N_2)\text{Poisson}^T(N_1 + N_2, 20). \tag{36}$$

The alternative data were chosen according to three models conditional on the total N. In model (a), N_1 is uniformly distributed in $[1, N - 1]$:

$$g_1(N_1, N_2) = \left\{ \begin{array}{ll} 1/(N_1 + N_2 - 1) & N_1 \neq 0 \ \& \ N_2 \neq 0 \\ 0 & \text{otherwise} \end{array} \right\} \text{Poisson}^T(N_1 + N_2, 20). \tag{37}$$

In models (b) and (c), N_1 has the fixed values $\max(1, N - 10)$ and $N - 1$ (implying $N_2 = 1$) respectively i.e.

$$g_2(N_1, N_2) = \left\{ \begin{array}{ll} 1 & N_1 = \max(1, N_1 + N_2 - 10) \\ 0 & \text{otherwise} \end{array} \right\} \text{Poisson}^T(N_1 + N_2, 20) \tag{38}$$

and

$$g_3(N_1, N_2) = \left\{ \begin{array}{ll} 1 & N_2 = 1 \\ 0 & \text{otherwise} \end{array} \right\} \text{Poisson}^T(N_1 + N_2, 20). \tag{39}$$

where $\text{Poisson}^T(N, M) = \left\{ \begin{array}{ll} 1/(1 - e^{-M}(1 + M)) & N \geq 2 \\ 0 & \text{otherwise} \end{array} \right\} \text{Poisson}(N, M)$ and $\text{Poisson}(N, M) = \frac{M^N}{N!} e^{-M}$.

Model (a) was chosen to be the minimally informative model that could apply if nothing is known about the alternative data that are sought except that they do not follow the null hypothesis that represents independent random assignment of the clones to the two libraries in proportion to their sizes. Models (b) and (c) are quite the opposite; model (b) represents a specific pattern of non-random behaviour (despite being unrealistic in this case) and model (c) is the most unlikely possible behaviour under the null hypothesis. They were introduced to test the statistical procedure described above based on the density of p-values of individual tests.

For this purpose, 100 samples of 1000 frequency pairs were independently generated as above for each scenario. In order to make this an example of the general theory of Section 8, the only parameter $r = E\{N_1/N\} = 0.3$, i.e. the expected fraction of the total frequency N in the first library, has to be known in the analysis procedures (i.e. 0.3 except where the deliberately wrong value 0.4 was used) because this theory requires no undetermined parameters. The purpose of the analysis is to estimate π_0 and arrange the data in the order of significance, taking the most significant first, such as to minimise the expected frequency of errors at whatever point in the ordered data set is chosen as the cut-off for the significant data.

The initial p-values for the density calculations were obtained using the binomial test (Nixon, 2012) which is to calculate the p-value of (N_1, N_2) by

$$p = \frac{\text{bin}(r, N_1, \phi) + 2.\min\left(\sum_{i=1}^{N_1-1} \text{bin}(r, i, N), \sum_{i=N_1+1}^{N-1} \text{bin}(r, i, N)\right)}{1 - (1 - r)^N - r^N} \tag{40}$$

where $\text{bin}(r, i, N) = r^i(1 - r)^{N-i}\binom{N}{i}$, because it is appropriate to the above null distribution being related to the quantile function of a continuous random variable which is always uniformly distributed. For the case when there is no parameter to be fitted, the iterated ODP (Nixon, 2012) reduces to ordering the data by the likelihood ratio

$g(x)/f(x)$ given by the NP lemma, where f is the null density and g is the alternative density (or equivalently the total density). This statistic is

$$\mathrm{lr}_1(N_1, N_2) = \left\{ \begin{array}{ll} 1/\{(N-1)\mathrm{bin}^T(0.3, N_1, N)\} & N_1 > 0 \ \& \ N_2 > 0 \\ 0 & \text{otherwise} \end{array} \right\} \tag{41}$$

for model (a) and

$$\mathrm{lr}_2(N_1, N_2) = \left\{ \begin{array}{ll} 1/\mathrm{bin}^T(0.3, N_1, N) & N_1 = \max(1, N-10) \\ 0 & \text{otherwise} \end{array} \right\} \tag{42}$$

for model (b) where

$$\mathrm{bin}^T(r, N_1, N) = \left\{ \begin{array}{ll} \mathrm{bin}(r, N_1, N)/(1-(1-r)^N - r^N) & N_1 > 0 \ \& \ N_2 > 0 \\ 0 & \text{otherwise} \end{array} \right\}. \tag{43}$$

These statistics were compared with the other statistics for analysing the simulated data.

9.1 Estimation of Probability Densities

To implement the method described in Section 8, the first step is to implement a probability density estimate based on a one-dimensional set of data having values in [0, 1]. The method chosen was based on Wand and Jones (1995) and consists basically of five steps.

1) The m values p_i were treated as quantiles and converted to instances of a normal random variable z with zero mean and standard deviation σ given by "slope" c divided by $\sqrt{2\pi}$ using the inverse normal cumulative distribution function (c.d.f.) routine available online (Acklam, 2003) i.e.

$$z_i = \frac{c}{\sqrt{2\pi}}.\text{inverse normal c.d.f.}(p_i) \tag{44}$$

2) The bandwidth parameter $h = (\frac{4}{3m})^{0.2}\hat{\sigma}$ (Wand & Jones, 1995, Equation 3.2) was calculated where $\hat{\sigma}$ is the sample standard deviation obtained from the set of z values. This value of h minimises the AMISE (asymptotic mean integrated squared error) for estimating the normal probability density with the sample standard deviation $\hat{\sigma}$ from m data points, using the kernel density estimate with the normal(0,1) kernel function K (implying $R(K) \equiv \int (K(x))^2 dx = 1/(2\sqrt{\pi})$ and $\mu_2(K) \equiv \int x^2 K(x) dx = 1$).

3) The initial estimate of the density was the kernel density estimate $\hat{f}(z) = \frac{1}{mh}\sum_{j=1}^{m} K(\frac{z-z_j}{h})$.

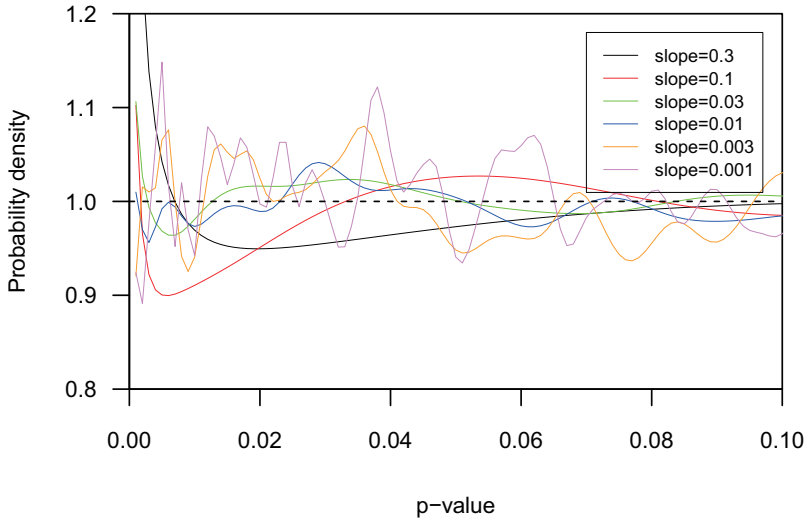
4) This estimate was used as the pilot estimate of the density to obtain the bandwidth $\alpha(z_i) = h(\hat{f}(z_i))^{-0.5}$ associated with the point z_i i.e. according to the square root law (Abramson, 1982).

5) Finally this bandwidth was used to construct the variable kernel density estimate

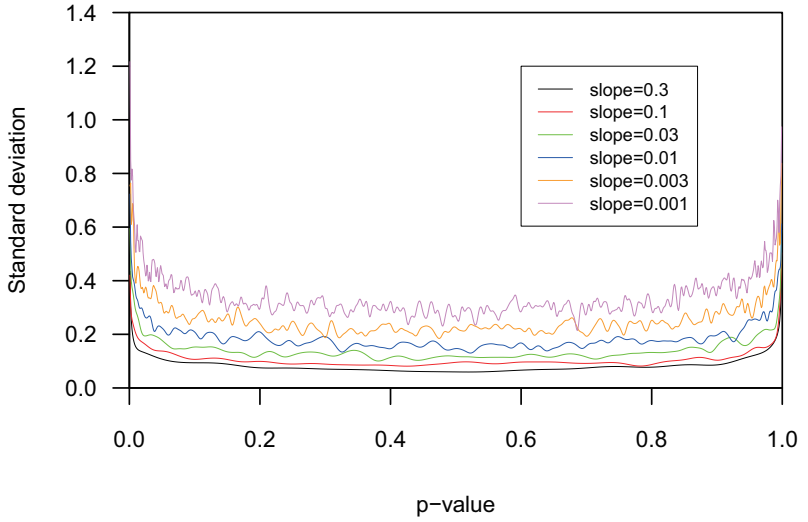
$$\hat{f}(z) = \frac{1}{m}\sum_{i=1}^{m} \{\alpha(z_i)\}^{-1} K\{(z - z_i)/\alpha(z_i)\} \tag{45}$$

(i.e. Wand & Jones, 1995, Equation 2.31).

To assess bias and variance of this estimate and fix the one free parameter c giving a reasonable point in the trade-off between the bias and variance, 100 populations were simulated with $m = 1000$ points from the uniform density and their density estimates were computed as above. The results for the mean and standard deviation of these estimates are shown in Figure 2.



(a) Mean



(b) Standard deviation

Figure 2. Estimates of the uniform probability density from 100 samples of 1000 points

The parameter value $c = 0.01$ was chosen in the following calculations as a reasonable compromise between the high variance and low bias that results when the c is small and the low variance and high bias that results when the slope is large. The main effect is near the ends of the range i.e. 0 and 1 (note that for $c = 0.3$ the mean density at $p = 0.0001$ was 1.51). Of particular concern was to not under or overestimate the density near 0 which corresponds to the most significant data.

There is another aspect to this that arises from the fact that the underlying data are small frequencies. Hence there are many repetitions of the data of individual tests in the entire data set so the data, after the initial analysis, is best represented by a set of pairs, (p-value, frequency). For the applications where the density estimate is required for each test (as opposed to the above case where the density was estimated on an equally spaced grid) the manner in which the frequency information was included in the density calculation had a major effect on the result. There seemed to be two reasonable ways to do this (here the sum is over the distinct values z_j with frequency s_j):

1) Use the same formula for the density as if all the p-values were distinct i.e.

$$\hat{f}_1(z_i) = \frac{1}{m} \sum_j s_j \{\alpha(z_j)\}^{-1} K\left\{(z_i - z_j)/\alpha(z_j)\right\}. \tag{46}$$

2) Use the above formula but use each distinct p-value once only, then correct the resulting density estimate for frequencies by multiplying the density estimates by their corresponding frequency afterwards i.e.

$$\hat{f}_2(z_i) = \frac{s_i}{m^*} \sum_j \{\alpha(z_j)\}^{-1} K\left\{(z_i - z_j)/\alpha(z_j)\right\}. \tag{47}$$

To these a third option was added as a curiosity.

3) Use the uncorrected density estimate from step (2) above i.e.

$$\hat{f}_3(z_i) = \frac{1}{m^*} \sum_j \{\alpha(z_j)\}^{-1} K\left\{(z_i - z_j)/\alpha(z_j)\right\}. \tag{48}$$

where m^* is the number of distinct p-values. Note that correct normalisation will not affect the order that the tests are considered to be significant and so will not affect the expected q-values, but will affect the estimates of π_0. $\hat{f}_1$ and $\hat{f}_3$ are correctly normalised while $\hat{f}_2$ is not and it is not clear how it could be normalised because it appears to be an estimate of a discrete density with the scale factor changing from point to point; this clearly requires more research. To speed up the calculations by avoiding adding large numbers of terms that are very close to zero, the sums were started from $j = i$ and continued to larger values of j while the terms were greater than a fraction 10^{-10} of the largest term. Then similarly the terms for $j < i$ were added in decreasing order of j. Finally the density for p was calculated as

$$\hat{f}(p) = \hat{f}(z)\left(\frac{dp}{dz}\right)^{-1} = \hat{f}(z)c.e^{(\pi z^2/c^2)} \tag{49}$$

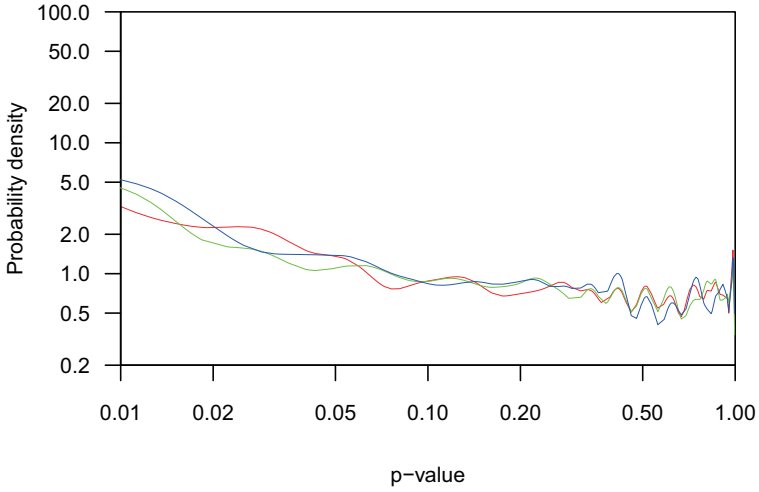
for each of the density estimators.

9.2 Calculation of q-Values

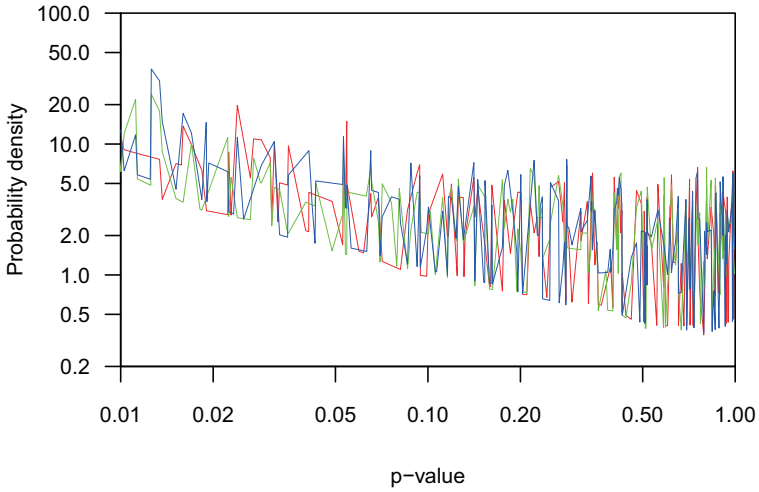
For each simulated data model and each test procedure a set of tuples was constructed that contains a tuple for each distinct data point: N_{1i}, N_{2i}, the number of times s_i in the simulated data set that the data point i (i.e. (N_{1i}, N_{2i})) occurred, and the test statistic. The set of tuples for the whole dataset was constructed as a list in the Standard Template Library of C++ to enable it to be sorted efficiently. Then the number of times each distinct data point was found in the null and alternative data subsets were obtained and added to the appropriate tuple. After sorting the tuples, most significant first, the running totals of the number of data points in the null and alternative data subsets were obtained, and the fraction of the data that are true nulls. This fraction is an estimate of the expected fraction of errors (q-value) that would occur if the given fraction of the data was assumed to be alternative hypotheses.

10. Results

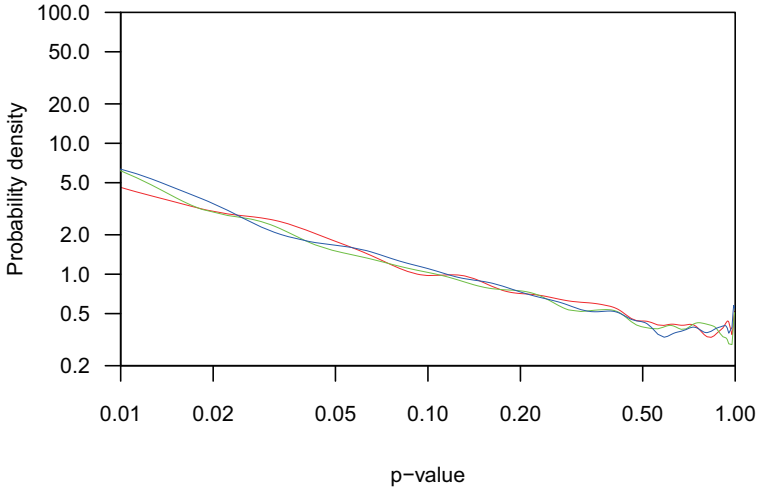
For the analysis of each simulated data set, the q-values were linearly interpolated onto an equally spaced grid of fractions n_1 of the data selected from 0.001 to 0.999 with spacing 0.001, because repetitions of data points (N_1, N_2) in the data sets implies that the n_1 spacing for the q-values directly calculated from the simulated data is irregular, and changes for each simulated data set. The means and standard deviations of the interpolated q-values for 100 data sets at each n_1 point were compared. In many cases independent sets of data sets are compared on the same graphs. To assess the practicality of the p-value density method for doing multiple hypothesis testing, three versions of this method corresponding to the three density estimates (46, 47 and 48) based on the binomial test were compared with using the p-values from the underlying binomial test, with the likelihood ratio test (based on the NP lemma), and with the ODP (Storey, 2007) and its iterative extension (Nixon, 2012) both using the actual and estimated values of r in the binomial test on which these methods are based.



(a) Density estimate $\hat{f}_1$

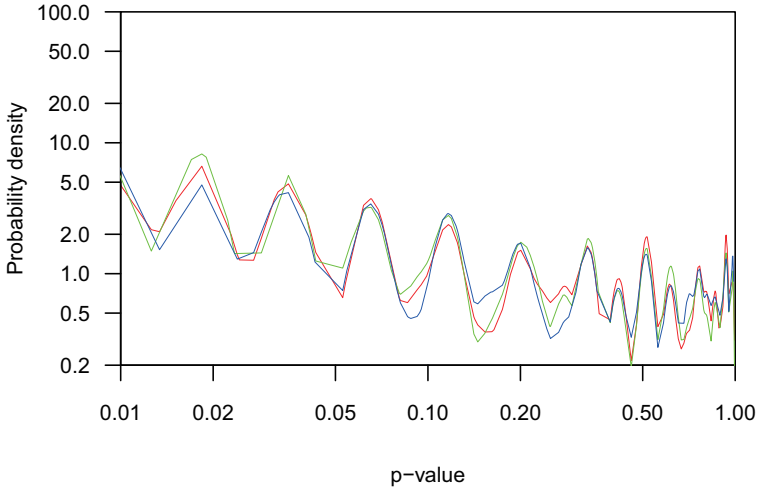


(b) Density estimate $\hat{f}_2$

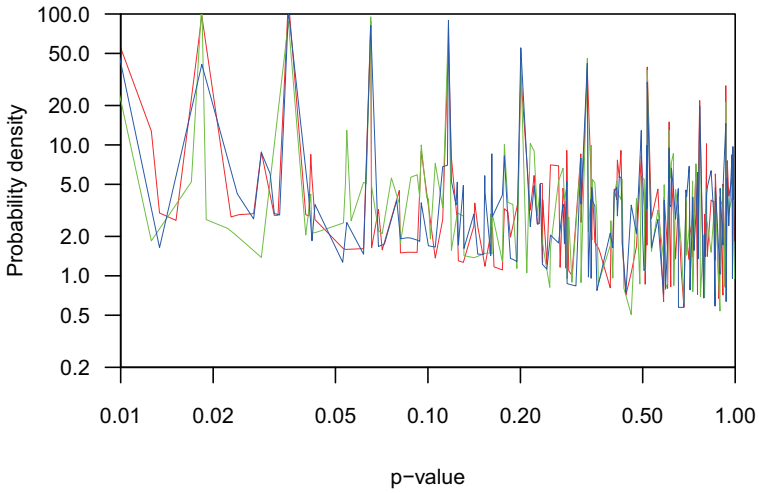


(c) Density estimate $\hat{f}_3$

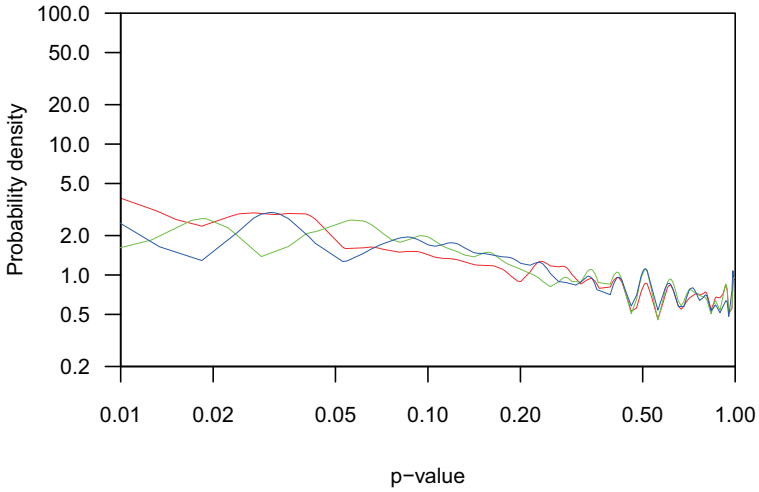
Figure 3. Probability density estimates using the three estimators from samples of 1000 points using model (a)



(a) Density estimate $\hat{f}_1$



(b) Density estimate $\hat{f}_2$



(c) Density estimate $\hat{f}_3$

Figure 4. Probability density estimates using the three estimators from samples of 1000 points using model (b)

The general theory demonstrating the optimality of the density-based method for multiple hypothesis testing in the scenario of Section 8 where only the test statistic and its null distribution are known, and in the large sample (simplest) case, implies that improvement in the practical efficiency of the method must result from improving the density estimate.

The results for model (c) are discussed first because they are very simple: an almost perfect agreement (data not shown) between the ideal case where all the alternatives are selected before any null hypotheses in the order of significance of the data, the p-value density method based on the binomial (using all three density estimators described here), and the binomial procedure itself. In this model, the alternative is the most extreme possible case that could happen according to the null hypothesis, i.e. an outcome with probability extremely close to zero, therefore an almost perfect separation of the alternative and null data sets was expected. The remaining results refer to models (a) and (b) only.

Each plot in each graph (Figures 3a-c for model (a) and Figures 4a-c for model (b)) was obtained from an independent random sample of 1000 frequency pairs from the above data simulation. The binomial test was applied using the actual value $r = 0.3$ to give the p-values, and the density was computed according to the appropriate Equations (46), (47) or (48), together with (49). The log-log plots are essentially linear with a negative slope as expected, indicating that low p-values correlate with high p-value density and vice-versa. What is interesting is the variation from monotonic behaviour which results in the p-values generating a different order of the data from the density values and a potentially better result for the expected error rate (q-value) obtained from the density of p-values rather than from the p-values themselves. Note the variation from sample to sample, and the different fluctuations resulting from the different ways of incorporating the frequency of occurrence of each unique data point into the density estimates. In particular by far the most variable density estimate comes from $\hat{f}_2$ that corrects for multiple identical p-values afterwards, and the smoothest density estimate ignores the multiplicity of identical p-values.

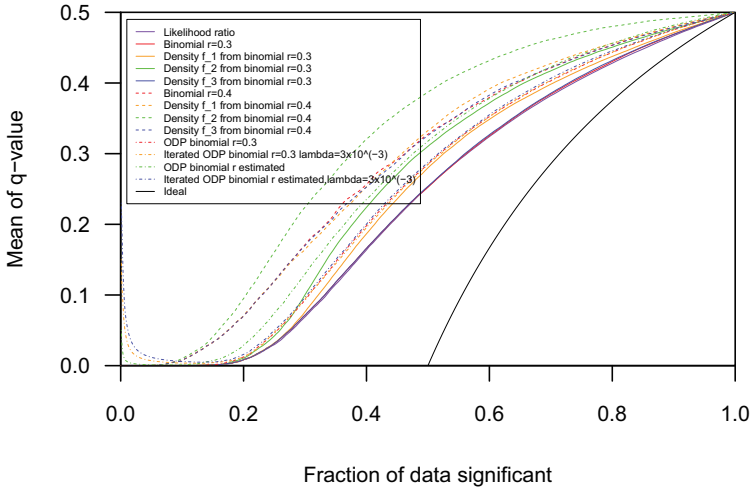
10.1 Results for q-Values Under Model (a)

The means and standard deviations of the q-values (Figure 5) were plotted against the fraction of data chosen as significant (n_1) for each analysis method, using the actual value $r = 0.3$ and independent simulations analysed using the *wrong* binomial parameter $r = 0.4$.

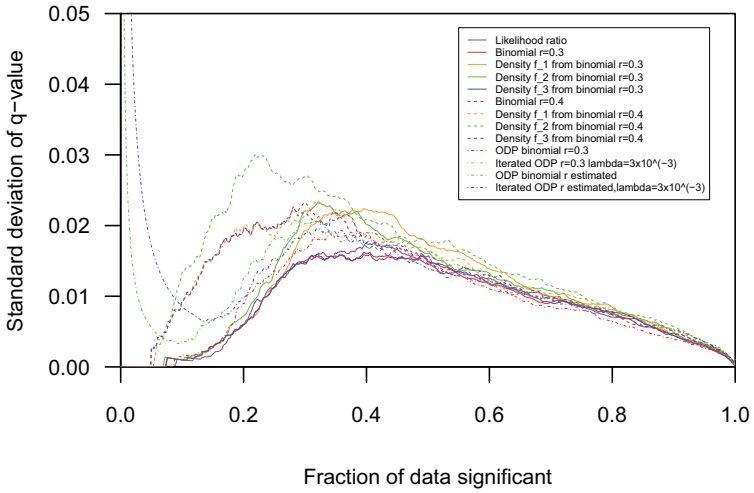
The value $r = 0.3$ for the simulations was chosen to be the same as the value that Nixon (2012) used in the simulations for the null hypothesis and the value $r = 0.4$ is the expected value of r for the entire data set consisting of 50% null hypotheses generated using $r = 0.3$ and 50% of alternative hypotheses generated using the truncated uniform distribution having r which is the expected value of $\frac{N_1}{N_1+N_2} = 0.5$. It is clear from these results that the improvement due to iteration of the ODP (Nixon, 1012) compared with the binomial test was mainly due to r being re-estimated as a value close to 0.3 using the least significant data only i.e. those most likely to be nulls, compared with using r fitted from the entire data set as a value close to $r = 0.4$, which was used in the binomial test then reported.

For analysis with $r = 0.3$ and $r = 0.4$, multiple hypothesis testing based on the binomial test, and multiple hypothesis testing based on f_3 (based on the binomial test) gave practically indistinguishable results. This was confirmed by comparing the mean q-values for the two procedures using two-sided t-tests (assuming a common variance) at 998 equally spaced points and gave roughly uniform p-value distributions (data not shown) with the smallest p-values being 0.02 and 0.007 for $r = 0.3$ and 0.4 respectively. For $r = 0.3$ these were also practically identical to the likelihood ratio test that requires knowledge of both the null and alternative models and is the most efficient way to carry out such tests (NP lemma). This is probably no coincidence because the binomial test was derived assuming no information on the distribution of alternatives, which are in fact distributed with N_1 uniform on $[1, N-1]$. The binomial test for $r = 0.3$ and the f_3 test based on it also had practically the smallest standard deviation of q-value throughout the whole range of n_1. Surprisingly for analysis with $r = 0.3$ and $r = 0.4$ the best of the density-based methods was the one without correcting for the p-value frequency (f_3), in fact for the three density estimation methods, the efficiencies were in the order $\hat{f}_2 < \hat{f}_1 < \hat{f}_3$ where $<$ should be read as "is less efficient than". These analyses were also compared with (1) the ODP (Storey, 2007) initialised by the binomial test but with r fixed at 0.3, and using the cut-off $\lambda = 3\text{x}10^{-3}$ to determine the initial partition of the data into significant and non-significant subsets, and (2) the iterated modification of this method (Nixon 2012) with the same parameters ($\lambda = 3\text{x}10^{-3}$ seemed to be close to the most efficient value for λ for samples of m=1000 data). Finally the original form of both these methods using the estimated value of r, and $\lambda = 3\text{x}10^{-3}$ were also carried out. The annoying spike at small fractions of data chosen as significant, using the iterated Storey (2007) method (Nixon, 2012) is completely removed by using any of the density-based testing methods here. The density-based

method using f_3 and the binomial test, and the practically equivalent binomial test itself (with the correct $r = 0.3$) outperformed all the other procedures tested without the full knowledge of the null hypothesis.



(i) Mean



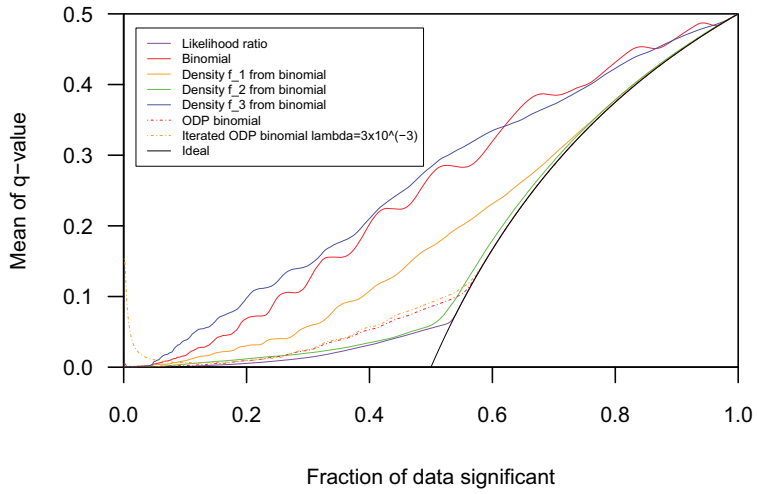
(ii) Standard deviation

Figure 5. Comparison of different analysis methods using the uniform alternative distribution model (a)

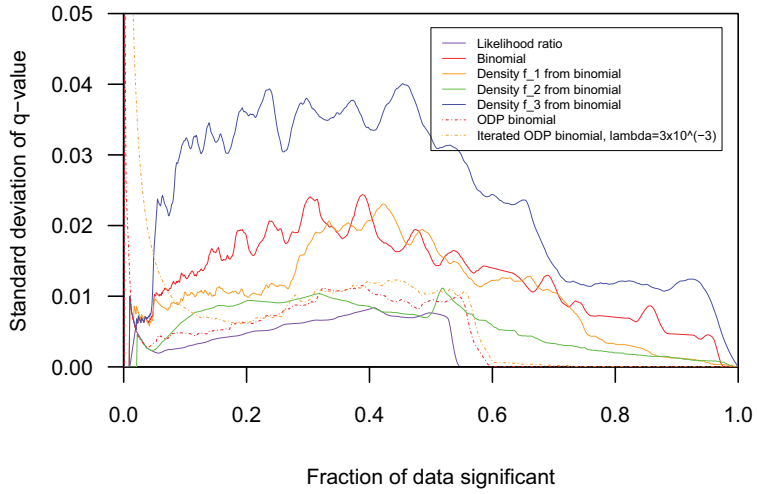
10.2 Results for q-Values Under Model (b) and Comparisons Between the Models

This gave a quite different result (Figure 6) for the q-values as a function of the fraction of the data taken as significant. Surprisingly there is a different (actually reverse) order for the efficiency of the density-based methods ($\hat{f}_3 < \hat{f}_1 < \hat{f}_2$), and two density-based methods are now far superior to the binomial test. The poor performance of the binomial test here was expected because it is not appropriate in this situation. It was included for comparison because the density-based tests are based on it. Again the best of the density-based methods is very close in performance to the optimal likelihood ratio test. The ODP and iterated ODP using the fixed value $r = 0.3$ worked well, with the ODP outperforming the iterated ODP that also had the prominent spike near zero. The results with r estimated as in the original forms of these methods performed more poorly than with the actual value $r = 0.3$ as expected (results not shown).

The ODP and iterated ODP performance cross the density-based method performance with $\hat{f}_2$ and are equivalent to the LR test in performance for large values of fraction significant (n_1). This is probably because the ODP method contains an internal simulation of the null hypothesis, but this information is not known to the density-based methods which only have the information about the test statistic and its distribution under the null hypothesis.



(i) Mean



(ii) Standard Deviation

Figure 6. Comparison of different analysis methods using the new alternative distribution model (b)

For model (a), the best of the f_3 density-based tests using the binomial test p-values, $f_3(p)$ is extremely close to being monotonically related the p-values themselves. This is verified by the closeness of the efficiencies of the binomial and f_3 density based test in Figure 5a. If this statement was exact, it would allow the density to be replaced by these p-values giving an equally good test that is much easier and faster test to carry out. A direct test would have both the advantage of being likely to be less computationally intensive than a density-based test and not subject to the statistical error in the latter mentioned above requiring a large dataset for good results, and should be sought if at all possible. Such a test however would probably require detailed knowledge of both the null and the alternative hypotheses i.e. a knowledge of what you are looking for.

It is also interesting that the standard deviation (SD) of the q-values is fairly insensitive to the statistical method used for model (a), but not for model (b).

10.3 Estimation of π_0

Another use of the p-value density is to obtain an estimate of π_0 which was 0.5 in these simulations. From 100 simulated data sets giving 1000 p-values each obtained from the binomial test as above, their p-value densities and the minimum value of the density was obtained (Table 2) which estimates π_0 (35). Note that $\hat{f}_2$ was not expected to yield a good estimate of π_0 because of the lack of proper normalisation and was included only for completeness. For both model (a) and model (b), the best estimate of π_0 was however given by the density estimate $\hat{f}_2$. There is a

basic difficulty in estimating anything by the smallest value of a quantity rather than some sort of average because the random errors would be expected to be relatively large. Apart from the estimator $\hat{f}_1$ that performs poorly, the estimates for model (b) are better than those for model (a) because it is easier to distinguish model (b) data points from the null hypothesis data points than to distinguish model (a) data points from the null hypothesis data points as the LR test performance shows.

Table 2. Estimates of the fraction of null hypotheses (π_0)

Density estimator (see text)	Mean (SD) Model (a)	Mean (SD) Model (b)
$\hat{f}_1$	0.31 (0.11)	0.202 (0.053)
$\hat{f}_2$	0.352 (0.027)	0.542 (0.055)
$\hat{f}_3$	0.310 (0.025)	0.446 (0.037)

Statistics for minimum probability density as an estimator for $\pi_0 = 0.5$ from 100 simulations based on the binomial test with given $r = 0.3$ and with the two alternative models for N_1 (see text).

11. Discussion

What has been developed in this paper is the multiple testing procedure (MTP1) for the general class of multiple hypothesis testing problems namely the two-groups model where the data are a mixture of those that come from two submodels, each with their own unknown parameters, and its specialisation (MTP3) that is appropriate when one of the submodels is effectively vacuous and would estimated from the data. These procedures are all based on a test statistic and sort the data into order of significance such that whatever cut-off point is chosen determining the fraction of the data selected as being significant, the fraction of true null hypotheses amongst this subset is as small as possible on the average. (Another procedure MTP2 was also mentioned based on a method Nixon (2012) used, but there is no theoretical reason for its optimality.) In the situation where both hypotheses are known without any unknown parameters, all three multiple testing procedures MTP1, MTP2 and MTP3 reduce to testing based on the NP lemma. In the case where the null hypothesis is specified implicitly by a known probability distribution of a test statistic t and the alternative is not specified, MTP3 reduces to the final test statistic which is the probability density of p-values derived from t. This of course leaves open the question as to what density estimation method should be used. This procedure implies a two-stage calculation of test statistics and p-values, once for the individual data sets, and once for the entire data set. This multiple testing procedure therefore adapts to the unspecified distribution of the alternative data to minimise the error rate whatever fraction of the data are selected as significant, but could be computationally more demanding per test than a direct test (because of the density estimation which is not yet optimised) and is subject to additional random error that decreases with the size of the data set arising from the probability density estimation. Of these three procedures (MTP1, MTP3, and the density-based method), the density-based test seems likely to be most important in applications because explicit models for alternative distributions may not be available.

In the numerical test of this method, the probability density was estimated by three methods that differ only in the manner in which the frequencies of occurrence of the data points (themselves being frequency pairs) in the entire data set are handled. Surprisingly it turned out that for testing with two different models for the alternative hypothesis, different versions of the density-based test (with different density estimators) performed best. For the estimation of π_0 (for either model) the same density estimator $\hat{f}_2$ had the smallest magnitude of bias.

The numerical results show that the p-value density method can improve the efficiency of multiple testing procedures resulting from a different order of significance of the data than would be obtained by simply using the p-values directly. That the order of significance was changed is confirmed by the estimated p-value distributions which are not monotonic functions of p. This would be expected to happen for example if there is a peak in the p-value density resulting from a cluster of p-values a little away from zero and there are some smaller more isolated p-values. Tests close to this peak could have the final effective p-values very close to zero and may be more significant than isolated p-values nearer to 0. Moreover the results as a whole confirm that the theoretically optimal test based only on the test statistics for the null distribution can outperform, provided an appropriate density estimator is supplied, the binomial test and under some conditions the ODP (Storey, 2007) and its iterated extension (Nixon, 2012) (that both know about the null hypothesis model), and be very close to the theoretically optimal LR test that requires both the null and alternative hypotheses to be specified.

In general once a null hypothesis is specified, it is possible to find a test statistic having a known distribution

under the null hypothesis, i.e. an algorithm that generates a p-value that is uniform under this null hypothesis. This can in general be done in an infinite number of ways corresponding to different test statistics though only a few may be sensible. The fact that it can be done in an infinite number of ways is obvious e.g. as a result of flipping over a segment $[a, b]$ of the range of p-values to generate a new "p-value", and such transformations with different segments could be repeated indefinitely. Different p-value algorithms are clearly appropriate to different alternative hypotheses for example testing H_0: $\mu_{1i} = \mu_{2i}$ against H_{1a}: $\mu_{1i} < \mu_{2i}$ or against H_{1b}: $\mu_{1i} > \mu_{2i}$. Therefore the specification of an algorithm to generate p-values from data contains implicitly information about both the null and alternative distributions. If only such a p-value algorithm is known, the general theory shows that the p-value density method should be used, which would be clearly unaffected in the large sample case by flipping over segments of the p-value distribution. In this example it would remove the distinction between an upper-tail t-test and a lower tail t-test and give something like the two-tailed test because the alternative model is not known. In this example data most significant by one criterion would be least significant by the other and vice-versa. This is appropriate in the situation where the alternative distribution is unknown as is the case here. In general, from its derivation, it seems that the p-value density test must remove the implicit information about the alternative distribution contained in the test procedure generating the original p-values and instead use the empirical distribution contained in the distribution of p-values.

If the alternative and null distributions are known, the optimal test is MTP1 and reduces to the NP lemma when there are no unknown parameters and orders the data in order of the likelihood ratio between the two models. In this case the statistical test via the density of p-values could be converted to a much more computationally efficient direct test without the disadvantages mentioned above for the density-based test, so the density-based test is likely to be most useful when the distribution of alternatives only known empirically, and its efficiency is expected to improve with the number of true alternative hypotheses amongst the data because the empirical alternative hypothesis will then be more precisely specified.

These observations are all affected by the finiteness of the data set resulting in random errors. There is clearly much theoretical work to be done in the finite sample analysis accounting for the effects of random sampling error and extending the continuously infinite data approach taken in this paper.

Because the models analysed in this paper ignore correlation between the data subsets for the different tests, it is expected that the methods presented here will be most useful in situations where no correlation would reasonably be expected between the different data subsets such as might happen for example when they arise from different genes from *in silico* analysis of sequencing data from RNA Seq gene expression studies.

For the purpose of applying this method to data sets that arise for example from gene expression studies via microarrays, there is much evidence that the data for different genes cannot be considered to be independent of each other, and correction methods such as surrogate variable analysis (SVA) (Leek & Storey, 2007) or its Partial Least Squares variant (SVA-PLS) (Chakraborty, Datta, Somnath, & Datta, Susmita, 2012) need to be applied to correct for this to allow much of this dependence to be accounted for in extra surrogate variables to be fitted. The p-values from the corrected tests for significance of the genes are almost (theoretically exactly) independent (Leek, & Storey, 2008). Then the assumptions behind the p-value density method described here are satisfied and this method theoretically gives the optimal solution of the multiple testing problem.

Acknowledgements

I am grateful to Matthew Links for reading the manuscript and suggesting some improvements and to the reviewer for pointing out some deficiencies in the earlier version of the manuscript.

References

Abramson, I. S. (1982). On bandwidth variation in kernel estimates–a square root law. *Annals of Statistics, 9*, 168-76.

Acklam, P. J. (2003). An algorithm for computing the inverse normal cumulative distribution function. Retrieved from http://home.online.no/ pjacklam/notes/invnorm/

Chakraborty, S., Datta, Somnath, & Datta, Susmita. (2012). Surrogate variable analysis using partial least squares (SVA-PLS) in gene expression studies. *Bioinformatics, 28*, 799-806. http://dx.doi.org/10.1093/bioinformatics/bts022

Dudewicz, E. J., & Mishra, S. N. (1988). *Modern Mathematical Statistics*. John Wiley & Sons, Inc.

Givan, S. A., Bottoms, C. A., & Spollen, W. G. (2012). Computational Analysis of RNA-seq. In H. Jin, & W. Gassmann (Eds.), *RNA Abundance Analysis: Methods and Protocols, Methods in Molecular Biology, 883*, 201-219. Springer Science+Business Media, LLC 2012. http://dx.doi.org/10.1007/978-1-61779-839-9

Hwang, J. T. G., & Liu, P. (2010). Optimal tests shrinking both the means and the variances applicable to microarray data. *Statistical Applications in Genetics and Molecular Biology, 9*(1), 1-31. http://dx.doi.org/10.2202/1544-6115.1587

Leek, J. T., & Storey, J. D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis *PLoS Genetics, 3*(9), e161, 1724-1735. http://dx.doi.org/10.1371/journal.pgen.0030161

Leek, J. T., & Storey, J. D. (2008). A general framework for multiple testing dependence. *Proc. Natn. Acad. Sci. USA, 105*(48), 18718-18723. http://dx.doi.org/10.1073/pnas.0808709105

Lunceford, J. K., Chen, G., Hu, P. H., & Mehrotra, D. V. (2011). Evaluating surrogate variables for improving microarray multiple testing inference. *Pharmaceutical Statistics, 10*, 302-310. http://dx.doi.org/10.1002/pst.466

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal, 7*, 308-313.

Nixon, J. H. (2012). Investigations into refinements of Storey's method of multiple hypothesis testing minimising the FDR, and its application to test binomial data. *Computational Statistics and Data Analysis, 56*, 4381-4398. http://dx.doi.org/10.1016/j.csda.2012.03.026

Phipson, B., & Smyth, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology, 9*(1), 1-12. http://dx.doi.org/10.2202/1544-6115.1585

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1997). *Numerical Recipes in C: The Art of Scientific Computing* (2nd ed., reprinted with corrections). Cambridge University Press.

Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York: Springer-Verlag.

Storey, J. D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Statist. Soc. B, 69*(3), 347-368.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natn. Acad. Sci. USA, 100*, 9440-9445. http://dx.doi.org/10.1073/pnas.1530509100

Wand, M. P., & Jones, M. C. (1995). *Kernel Smoothing*. Monographs on Statistics and Probability No. 60, Chapman and Hall.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews: Genetics, 10*, 57-63. http://dx.doi.org/10.1038/nrg2484