Method Selection and Assumption Evaluation: Applications to Gene Expression Data

DEMBA FOFANA
University of Texas Rio Grande Valley,
School of Mathematical and Statistical Sciences,
1201 W University Dr, Edinburg, TX 78539

Abstract

This paper proposes hybrid-testing procedures as a general class of methods that simultaneously addresses the problems of procedure selection and multiple testing. Hybrid-testing procedures apply a set of primary testing procedures to perform the tests of primary interest (e.g. using the t-test or rank-sum test to evaluate equality of univariate means across groups for a large number of variables) and a set of assumption testing procedures to statistically evaluate the assumptions (e.g. the normality of data as an assumption of the t-test for those same variables) of the primary test procedures. The results from each testing procedure are summarized as a set of p-values and empirical Bayesian probabilities (EBPs) of the corresponding null hypotheses. Prior knowledge of the statistical properties of the primary testing procedures according to the validity of the statistically evaluated assumptions is used to define an algorithm. The algorithm selects the best primary testing procedure according to the validity of each assumption testing procedure to determine weights for combining the EBPs from the primary testing procedures into a final set of EBPs for the hypotheses of primary interest (univariate equality of means for the set of variables). This final EBP is a measure of statistical significance that adjusts for multiple-testing and incorporates a formal evaluation of assumptions to combine the results of several hypothesis-testing procedures in a manner guided by prior statistical knowledge. The proposed procedures are applied to gene expression data.

1 Introduction

We propose procedures for testing hybrid hypotheses as a general methodology for addressing the frequently overlooked problem of selecting the most appropriate hypothesis testing procedure to use when performing a very large number of tests in the analysis of high-dimensional data. The procedure selection is complex, usually because distributional test statistics under a specific hypothesis (e.g. the normal distribution assumption for the t-test) may be reasonable for some tests but not valid for other tests. The statistical evaluation of assumptions for each test also introduces an additional layer of multiplicity to the analysis. We propose the use of an estimate of the empirical Bayesian probability (EBP) of the null hypothesis for each test in a manner that simultaneously addresses the multiple testing problem and the procedure selection problem. This overall EBP estimate is the weighted mean of EBPs obtained from the several hypothesis testing procedures performed. The weights are defined in terms of the EBPs that the assumptions hold. Three specific hybrid-testing procedures developed for frequently encountered applications show good performance characteristics in simulation studies and in applications from cancer genomics. An

R package that implements these hybrid-testing procedures and provides "plug-in" capabilities to facilitate the rapid implementation of novel hybrid-testing procedures, HybridMTest, has been developed by Pounds and Fofana (Bioconductor, 2012).

Hypothesis testing is formalized through a null hypothesis and an alternative hypothesis. P-values are computed under the null hypotheses and several other technical assumptions. The computed p-values may not be uniformly distributed if the technical assumptions do not hold. If the technical assumptions are invalid then estimations or control of type I errors and type II errors will be misleading.

Three specific hybrid-testing procedures developed for frequently encountered applications show good performance characteristics in simulation studies and in applications from cancer genomics.

2 Literature Review

Multiple hypothesis testing is an essential step in the analysis of high-dimensional data like genomic or proteomic data. It presents, however, a lot of challenges. There are well known difficulties to face when dealing with multiple testing procedures, such as controlling for multiple testing, see Benjamini & Hochberg (1995), Efron et al. (2001), Storey (2002), taking assumptions into consideration (Pounds & Rai, 2009). These challenges occur very often when analyzing gene expression data.

The need to overcome these challenges provides the motivation for the procedures we develop in this paper. We implement one procedure, hybrid-testing, for analyzing data in a multivariate setting. We develop hybrid-testing procedures to combine tests that are made under a variety of distributional assumptions.

When enormous amounts of data are produced, one statistical task consists of data reduction; Efron et al. (2001) introduce a nonparametric empirical Bayes probability (EBP) model that can be used as a summary statistic. Their model can also be used to make simultaneous inferences.

Some procedures look at averaging models that average across models, for example Bayesian Model Averaging for Linear Regression Models (see Raftery, Madigan and Hoeting, 1997). They emphasize that selecting subsets of predictor variables is a basic part of building a linear regression model. A typical approach would be carrying out a model selection exercise that leads to a single "best" model and then making inferences as if the selected model were the true model. They show that averaging procedures provide better predictive performance than any single model that might reasonably have been selected. As a single model ignores a major component of uncertainty about the model itself.

Hybrid-testing is an averaging procedure too. This procedure estimates the proportion, π_0, of true null hypotheses in multiple-hypothesis set-ups and considers, at the same time, the underlying assumptions of the hypothesis test. The tests are based on observed p-values. Estimating the proportion of true null hypotheses is of interest in many situations where a large number of hypothesis tests are conducted, see Nettleton et al., 2006. To estimate the density function of the p-values, some researchers fit a beta uniform distribution to estimate π_0, see Pounds & Rai (2009), while others use a method called the Grenander density estimation, a nonparametric, maximum likelihood estimator based on the order statistics of the p-values, see Langaas et al. (2005).

Hybrid-testing uses the EBP approach and the Grenander estimator of the p-value density. It simultaneously controls for multiple testing of the error rate and selects the best testing procedures.

3 Notations

Let G represent the number of hypothesis tests to be performed. In an application of interest G will be the number of genes, and let

$$H_{og} : \theta_g = \theta_{og} \tag{1}$$

represent the null hypothesis for gene g, and the alternative hypothesis

$$H_{ag} : \theta_g \neq \theta_{og},\ g = 1, \cdots, G. \tag{2}$$

Suppose also that there are two different methods, M_1 and M_2 that can be used to perform these statistical tests. When M_1 is used, let $\mathbf{T}_1 = \{T_{11}, \cdots, T_{1G}\}$ represent the test statistics obtained and $\mathbf{P}_1 = \{P_{11}, \cdots, P_{1G}\}$, the corresponding P-values. Similarly, let $\mathbf{T}_2 = \{T_{21}, \cdots, T_{2G}\}$ and $\mathbf{P}_2 = \{P_{21}, \cdots, P_{2G}\}$ represent the corresponding quantities for method M_2.

Let

$$H_{oAg} : A_g = 1, \tag{3}$$

represent that the hypothesis from an assumption categorized as 1 is true and

$$H_{aAg} : A_g = 2, \tag{4}$$

the assumption that an alternative assumption categorized as 2 is true. Let $\mathbf{T}_a = \{T_{a1}, \cdots, T_{aG}\}$ be the test statistics and $\mathbf{P}_a = \{P_{a1}, \cdots, P_{aG}\}$ the corresponding P-values.

4 FDRs, Empirical Bayes Probability and Grenander Procedure

False Discovery Rates (FDRs)

In high-dimensional genomic or proteomic data, analyses requiring multiple testings as described in (1), such multiple testings result in inflated type I error. Controlling false discovery rates (FDRs) have proven to be reliable and less conservative statistical method in determining the significance of genomic features. FDR methodologies are currently used in many applications which include gene expression data analysis, spectrometric peak detection, single nucleotide polymorphism (SNP) discovery, and edge selection in genetic networks. FDR was introduced in the seminal papers by Schweder and Spjøtvoll (1982) and Benjamini and Hochberg (1995). Storey (2002) improved it through pFDR.

Consider a test statistic T, with observed values denoted by t. Assume that across hypotheses g, $g = 1, \cdots, G$, the statistics T follows a two-component mixture distribution, with density function

$$f(t) = \pi f_0(t) + (1 - \pi) f_a(t), \tag{5}$$

where f_0 the density function of T under the null hypothesis and f_a the density function of T under the alternative hypothesis. This mixture model may also be written in terms of the distribution functions

$$F(t) = \pi F_0(t) + (1 - \pi) F_a(t), \tag{6}$$

where $F_0(t)$ is the distribution function of T under H_0, and $F_a(t)$ is the distribution function of T under H_a. For an observed value $T = t$, Strimmer (2008) calculates the false discovery

rate as

$$FDR(t) = \pi \frac{1 - F_0(t)}{1 - F(t)}. \tag{7}$$

It should be noted that the expression (7) corresponds to the posterior probability that H_0 is true given $T \geq t$, that is

$$FDR(t) = \mathbb{P}(H_0 \text{ true } \mid T \geq t). \tag{8}$$

Intuitively, FDR is simply a P-value corrected for multiplicity. The idea is to provide an efficient data reduction method, since microarray experiments produce enormous amounts of data that require novel data reduction strategies.

Empirical Bayes Probability

Local FDR, also known as empirical Bayes probability (EBP), was introduced by Efron et al. (2001). Efron and Tibshirani (2002) discuss other aspects of EBP. Using the concept of local FDR by Efron et al. (2001), Strimmer (2008) defines the empirical Bayes probability of the null hypothesis as

$$EBP(t) = \pi \frac{f_0(t)}{f(t)}. \tag{9}$$

Full Bayesian analysis would require prior specification of π, $f_0(t)$, $f_a(t)$, but EBP does not require any prior distribution, Efron and Tibshirani (2002) provide more details.

We illustrate the objective of developing a hybrid EBP by considering a two sample microarray data analysis.

In such analysis, the goal is to identify genes that are differentially expressed relative to the control group and it is not uncommon in such applications to assume that the two samples come from the same scale-location family of distribution such as the Gaussian family. However, in the first step in coming to that assumption is usually to try to transform the pre-processed expression to normality. Hence the first set of hypothesis in the hybrid testing is a test of normality for each gene which is followed by the second layer of testing for equality of mean in expression for each gene.

For this first layer of testing, we use the Grenander density estimation to determine the distribution of the transformed expression data. The empirical Bayes probability for the hybrid-testing is derived from the following theorem.

Theorem 1. *Empirical Bayes probability for hybrid tests*
Let T be a statistic to be used for testing H_0 versus H_a and let H_{01} represent the hypothesis that the distribution of T belongs to family F_1. Symbolically, let

$$H_{01} : A = 1 \text{ vs } H_{a1} : A \neq 1. \tag{10}$$

Then

$$EBP(H_0) = EBP(H_0 \mid H_{01})EBP(H_{01}) + EBP(H_0 \mid H_{a1})EBP(H_{a1}) \tag{11}$$

Proof.

$$\begin{aligned}
EBP(H_0) &= \frac{\pi_0 f_0(z)}{f(z)} \\
&= \frac{P(H_0)P(z|H_0)}{P(z)} \\
&= \frac{P(H_0 \cap z)}{P(z)} \\
&= \frac{P(H_0 \cap z \cap (H_{01} \cup H_{a1}))}{P(z)} \\
&= \frac{P(H_0 \cap z \cap H_{01})}{P(z)} + \frac{P(H_0 \cap z \cap H_{a1})}{P(z)} \\
&= \frac{P(H_0|H_{01})P(z|H_0 \cap H_{01})}{P(z|H_{01})} \times \frac{P(H_{01})P(z|H_{01})}{P(z)} + \\
&\quad \frac{P(H_0|H_{a1})P(z|H_0 \cap H_{a1})}{P(z|H_{a1})} \times \frac{P(H_{a1})P(z|H_{a1})}{P(z)} \\
&= EBP(H_0 \mid H_{01})EBP(H_{01}) + EBP(H_0 \mid H_{a1})EBP(H_{a1})
\end{aligned}$$

□

In order to compute the EBPs and FDRs it is necessary to estimate not only π, but also f, F, F_0, and f_0. However, the raw test statistics are not collected in our study, instead the P-values of the tests are available. Under the null hypothesis, P is uniformly distributed $[0,1]$ when the test statistic is continuous. For one-sided tests the knowledge of P-values can be used to obtain F_0 and f_0 approximately. To estimate π, f, and F a Grenander estimator is described in the following section, (Langaas and Lindqvist, 2005). In the following section we describe the Grenander estimator.

The Grenander Estimator

Denote Θ the set of decreasing density functions on $[0,1]$. Let $p_{(1)} \leq, \cdots, \leq p_{(G)}$ be the ordered observed $p-values$ from the G hypothesis tests. A nonparametric maximum likelihood estimator of f in Θ is given by Langaas and Lindqvist (2005) as

$$\hat{f} = arg \max_{\ell \in \Theta} \left\{ \prod_{g=1}^{G} \ell(p_{(g)}) \right\}, \tag{12}$$

$\hat{f}$ is known as the Grenander estimator, Grenander (1956).
For each g, $g = 1, \cdots, G$, $\hat{f}_g$ is determined by

$$\hat{f}_g = \min_{l \leq g-1} \max_{k \geq g} \left\{ \frac{\hat{F}(p_{(k)}) - \hat{F}(p_{(l)})}{p_{(k)} - p_{(l)}} \right\} \tag{13}$$

letting $\hat{f}$ be constant in each interval $(p_{(g)}, p_{(g+1)}]$.
$\hat{F}$ is defined by

$$\hat{F}(\alpha) = \frac{\#\{p_g \leq \alpha\}}{G}. \tag{14}$$

Several methods are proposed to estimate the parameter π. Langaas and Lindqvist (2005) estimate it as

$$\hat{\pi} = \min_{l \leq G-1} \left\{ \frac{\hat{F}(p_{(G)}) - \hat{F}(p_{(l)})}{p_{(G)} - p_{(l)}} \right\}. \tag{15}$$

Equation (13) shows that the Grenander estimate is the (left-hand) slope of the least concave majorant of the empirical distribution function $\hat{F}$. The estimator can be computed

by using the pool adjacent violators algorithm (Robertson et al., 1988). The Grenander estimator uses $\hat{F}$, the empirical cumulative distribution function (ECDF). To implement our method we use "fdrtool" a statistical package by Strimmer (2008) for the Grenander density estimation.

5 The General Hybrid-Testing Procedure

Pounds and Morris (2003) describe a method that estimates EBPs by fitting a beta-uniform mixture (BUM) model to the p-values and then estimating the EBP for a specific p-value. Efron et al. (2001) use a logistic regression to compute EBPs. The hybrid-testing procedure, however, uses the Grenander estimator density of the p-values.

Specifically, let $\mathbf{B}_1 = \{B_{11}, \cdots, B_{1G}\}$, $\mathbf{B}_2 = \{B_{21}, \cdots, B_{2G}\}$, and $\mathbf{B}_a = \{B_{a1}, \cdots, B_{aG}\}$ be the empirical Bayes probabilities calculated using (9), and let the p-values $\mathbf{P}_1$, $\mathbf{P}_2$, and $\mathbf{P}_a$ be defined as in the notation section. The Grenander estimates of f and π are as (12), and (15), respectively. By using the law of total probabilities, an estimate of the EBP for testing H_{og} versus H_{ag} is given by

$$EBP_g^* = B_{ag} \times B_{1g} + (1 - B_{ag}) \times B_{2g}. \tag{16}$$

In addition, let

$$EBP_g^{**} = \begin{cases} B_{1g}, & \text{if } B_{ag} \geq \varsigma \\ B_{2g}, & \text{if } B_{ag} < \varsigma, \end{cases} \tag{17}$$

with ς a pre-specified threshold, define another hybrid-testing procedure. The decision rule is to reject the null hypothesis $H_{og} : \theta_g = \theta_{og}, g = 1, \cdots, G$ when EBP_g is less than a pre-specified threshold τ.

The hybrid-testing procedure, EBP^*, can be extended to the case where the number of possible testing methods is greater than two. Suppose there are k possible distributional assumptions, $A_1, \cdots, A_k$ for each of the statistic tests. A final weighted EBP for the g^{th} test is computed as

$$EBP_g^* = \sum_{i=1}^{k} w_{ig} B_{ig} \tag{18}$$

where w_{ig} are the weights and satisfy $w_{ig} \geq 0$ and $\sum_{i=1}^{k} w_{ig} = 1$.

6 Some Specific Hybrid-Testing Procedures

We illustrate the above using three hybrid-testing procedures: the hybrid t-Wilcoxon, the hybrid ANOVA-Kruskal-Wallis, and the hybrid Pearson-Spearman. We conduct simulations in order to see the performances of our methods compared to other methods. In addition, we apply the hybrid-testing procedures to real data on human Apendema and AML genes expression data. In the first data set, we compare gene expression levels in three groups and in the second one, we investigate how gene expression is related to DNA synthesis rate data.

Hybrid t-Wilcoxon Testing

When comparing two groups, many different statistical tests can be used. The t-test and the Wilcoxon test are examined here, see Wilcoxon, 1945.

Consider the following multiple hypotheses tests

$$H_{0g}\colon \mu_{1g} = \mu_{2g} \text{ vs } H_{ag}\colon \mu_{1g} \neq \mu_{2g},\ g = 1, \cdots, G. \tag{19}$$

Commonly, under the assumption that the test statistics have normal distribution a t-test is used. However, a Wilcoxon test based on rank sums maybe used when the normality assumption is not valid. It is worthwhile mentioning that when the normally assumption is correct, the t-test may be more powerful than the Wilcoxon test, however, if the assumption of normality is not valid a t-test is not an optimal test and may be anti-conservative. We use the Shapiro-Wilk test to test for normality, Royston (1982) and Shapiro and Wilk (1965). Let $p_1^s, \cdots, p_G^s$ be the G p-values computed from the Shapiro-Wilk's test statistics for the G normality tests, $p_1^t, \cdots, p_G^t$ be the G p-values calculated from the t-test statistics, and $p_1^w, \cdots, p_G^w$ be the G p-values derived from the rank-sum statistics. Let $B_1^s, \cdots, B_G^s$, $B_1^t, \cdots, B_G^t$, and $B_1^w, \cdots, B_G^w$ be the respective empirical Bayes probabilities from the G p-values of the G p-values of Shapiro-Wilk's test statistics, the t-test statistics, and the G p-values of the rank-sum statistics. The weighted empirical Bayes probability, EBP_g^*, is as follow

$$EBP_g^* = B_g^s \times B_g^t + (1 - B_g^s) \times B_g^w. \tag{20}$$

And finally, the selected empirical Bayes probability, EBP_g^{**}, is computed as follows:

$$EBP_g^{**} = \begin{cases} B_g^t, & \text{if } B_g^s \geq \varsigma \\ B_g^w, & \text{if } B_g^s < \varsigma. \end{cases} \tag{21}$$

Hybrid ANOVA–Kruskal-Wallis

Two statistical tests that can be used for comparing multiple groups are ANOVA and the Kruskal-Wallis test. The choice between these two different tests depends on whether the data can be assumed to come from a normal distribution.

Consider the following hypothesis testings

$$H_{0g}\colon \mu_{1g} = \mu_{2g} = \cdots = \mu_{kg} \text{ vs } H_{ag}\colon \mu_{ig} \neq \mu_{jg},\ i \neq j \text{ for at least one } i \text{ and } j \text{ for each } g, \tag{22}$$

$g = 1, \cdots, G$, where μ_{ig} is the mean value of group i for gene g. For each component g, the test determines if there exist at least two groups that have different means. A hybrid-testing procedure can take assumptions for each test into consideration.

We select two well-known methods of testing k-group means, an ANOVA test which assumes the populations are normally distributed, and the non-parametric Kruskal-Wallis test which makes no distributional assumptions. When the data are normally distributed, the ANOVA test is believed to be more powerful than the Kruskal-Wallis test, but when the assumption of normality does not hold the Kruskal-Wallis test may perform better than the ANOVA test. In order to check whether normality may be assumed, a Shapiro-Wilk test of normality can be used. To implement a hybrid-testing procedure for $K-$group comparison, we first compute three different sets of p-values. Let $p_1^s, \cdots, p_G^s$, $p_1^a, \cdots, p_G^a$, and $p_1^k, \cdots, p_G^k$ be the G p-values calculated from the Shapiro-Wilk, the ANOVA, and the Kruskal Wallis, respectively. Let $B_1^s, \cdots, B_G^s$, $B_1^a, \cdots, B_G^a$, and $B_1^k, \cdots, B_G^k$ be the respective empirical Bayes probabilities computed from the G p-values of the Shapiro-Wilk, the ANOVA, and the Kruskal-Wallis test statistics.

The hybrid EBPs are given by

$$EBP_g^* = B_g^s \times B_g^a + (1 - B_g^s) \times B_g^k, \tag{23}$$

and

$$EBP_g^{**} = \begin{cases} B_g^a, & \text{if } B_g^s \geq \varsigma \\ B_g^k, & \text{if } B_g^s < \varsigma. \end{cases} \tag{24}$$

Hybrid Pearson-Spearman

In the literature of gene expression data, it is common to ask whether certain covariates are significantly correlated with a gene expression measurement. Two different statistics that are widely used to measure linear relationship are the Pearson test, and the nonparametric Spearman rho. The choice between one of the procedures is predicated on the normality assumption. The Pearson test assumes that the data come from a normal population, while the Spearman rho test is distribution free.

Suppose that the i^{th} observation (expression value) from gene g satisfies the linear model

$$Y_{gi} = \beta_{g0} + \beta_{g1} X_i + \varepsilon_{gi},\ g = 1, \cdots, G;\ i = 1, \cdots, n, \tag{25}$$

with $\beta_g = (\beta_{g0}, \beta_{g1})$ is a vector of parameters to estimate for gene g, X_i is the measure of a covariate, and ε_{gi} is an error term.

Hypothesis tests for the linear dependency of Y_g on X is

$$H_{0g} \colon \beta_{g1} = 0 \text{ vs } H_{ag} \colon \beta_{g1} \neq 0,\ g = 1, \cdots, G. \tag{26}$$

Let $p_1^s, \cdots, p_G^s$, $p_1^p, \cdots, p_G^p$, and $p_1^{sp}, \cdots, p_G^{sp}$, and be the respective G p-values calculated from the Shapiro-Wilk, Pearson, and Spearman test statistics. Let $B_1^s, \cdots, B_G^s$, $B_1^p, \cdots, B_G^p$, and $B_1^{sp}, \cdots, B_G^{sp}$ be the empirical Bayes probabilities derived from the G p-values of the Shapiro, Pearson, and Spearman test statistics, respectively.

The empirical Bayes probabilities are given by

$$EBP_g^* = B_g^s \times B_g^p + (1 - B_g^s) \times B_g^{sp}, \tag{27}$$

and

$$EBP_g^{**} = \begin{cases} B_g^p, & \text{if } B_g^s \geq \varsigma \\ B_g^{sp}, & \text{if } B_g^s < \varsigma. \end{cases} \tag{28}$$

7 Simulations

In order to compare the hybrid-testing with other methods, we conduct three different series of analysis through simulations, two-group comparison, three-group comparison, and regression analyses and use AUCs, sensitivity, and specificity.

Hybrid t-Wilcoxon Simulation

The hybrid-testing procedure is studied for two-group comparison analysis using simulations. With a total of G genes, the hybrid-testing procedure is compared with other statistical tests. In a two-group comparison study, the problem is, for each gene, to test whether

its expression level is different between two groups. Consider the following hypothesis testings

$$H_{0g}: \mu_{1g} = \mu_{2g} \text{ vs } H_{ag}: \mu_{1g} \neq \mu_{2g},\ g = 1, \cdots, G, \tag{29}$$

where μ_{ig} is group i mean expressing for gene g.

Using three different methods: t-test with equal variances, t-test with unequal variances, and Wilcoxon test can be used to conduct the analysis, and the Bartlett test to check for equality of variances, we compute the hybrid-testing as

$$EBP_g^* = B_g^s \times [B_g^b \times B_g^{evt} + (1 - B_g^b) \times B_g^{uvt}] + (1 - B_g^s) \times B_g^w \tag{30}$$

where B_g^b, B_g^{evt}, and B_g^{uvt} are EBPs from the Bartlett test, the t-test with equal variances, and the t-test with unequal variances, respectively.

We simulate expressions of G genes where n_0 of the G genes are from $H_{0.}$, the null hypotheses, and n_a of the rest of the G genes are from $H_{a.}$, the alternative hypotheses. Among the n_0 genes, some are $N(0, \sigma^2)$, some are $Log - normal(0, \sigma^2)$ and others are $Cauchy(0, \sigma)$, 0 is the location parameter and σ the scale parameter. Also, among the n_a alternative gene expression values some are normally distributed $N(\mu, \sigma^2)$, some are $Log - normal(\mu, \sigma^2)$ and the others are $Cauchy(\theta, \sigma)$. Appendix B contains more details on the simulation setups. In this setup the exact number of expressed genes and the exact number of unexpressed genes are known. In each setup, we conduct several settings. Each setting corresponds to a sample size and each setting is replicated a number of times.

We conduct several simulation studies. In the first setup, there are two groups of sample size varying from 5, 10, 25, and 50. The number of null genes with the normal distribution are 720, number of null genes with the Cauchy distribution are 80. The number of alternative genes is 20 for the Cauchy distribution and 180 for the normal distribution, 200 alternative genes in total, and the number of replications is 1000.

We compute the powers and AUCs in each setting. Powers and AUCs vary according to settings, therefore, for each setting a power and an AUC are computed and the corresponding graphs are provided. Table 1 presents the corresponding powers for the competing procedures and Figure 1 is the corresponding graph. Table 2 shows the AUC results and Figure 2 presents the corresponding graph. The greater power or the greater AUC corresponds to a better methodology. These tables and graphs show that the hybrid procedures are more powerful than other procedures in most of the settings. For instance, in Table 1, when the sample size is 50, the respective powers for the t-test with equal variances, the t-test with unequal variances, the Wilcoxon test, the hybrid-testing EBP^*, and the hybrid-testing EBP^{**} are 0.665806, 0.662936, 0.80269, 0.87464, and 0.810064.

In a second simulation setup, the Cauchy distribution is replaced by the Log-normal distribution. We give more details in Appendix B. The results are presented in Table 3, and Figure 3; in Table 4, and Figure 4. We draw the same conclusion as in the first setup. In most of the settings, hybrid-testing procedures reveal to be more powerful than the other competing procedures.

Table 1: Two-Group Simulation Power Comparison

SS	EV t-test	UV t-test	Wilcoxon	EBP^{*}	EBP^{**}
5	0.002434	0.000846	0	0.02818	0.00069
10	0.08058	0.05682	0.115098	0.270526	0.13047
25	0.444066	0.435188	0.553762	0.674988	0.565108
50	0.665806	0.662936	0.80269	0.87464	0.810064

Notes: This shows powers for different methods of 2−group comparison in multidimensional testing. $SS \equiv$ *Sample size*. EV t-test and UV t-test mean t-test with equal and unequal variances, respectively. EBP^{*} is hybrid-testing as in (30) and EBP^{**} is hybrid-testing as in (17).

Figure 1: Two-Group Simulation Power Comparison

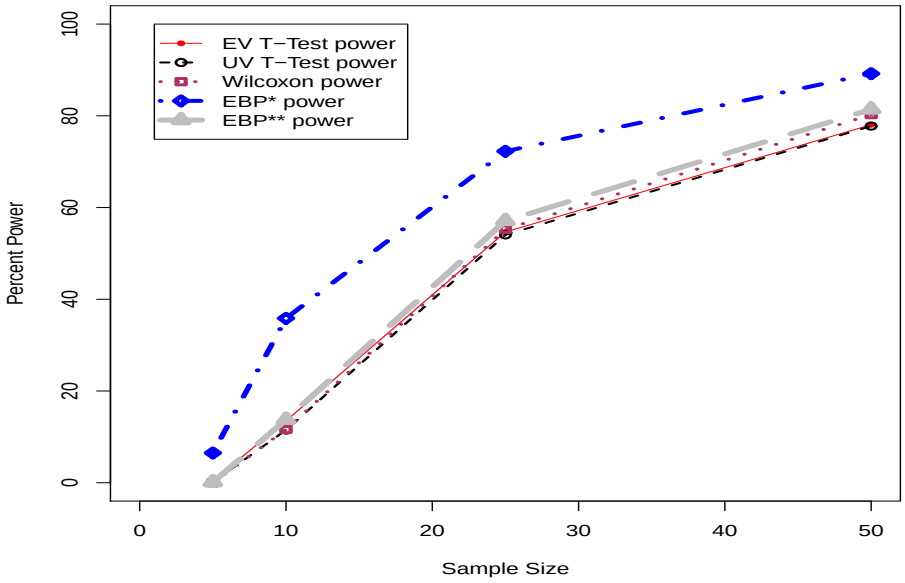


Table 2: Two-Group Simulation AUC Comparison

SS	EV T-test	UV T-test	Wilcoxon	EBP^{*}	EBP^{**}
5	0.64849	0.648356	0.567129	0.64236	0.635831
10	0.730395	0.730853	0.735035	0.755804	0.756211
25	0.829219	0.829531	0.902337	0.910524	0.910104
50	0.868432	0.868546	0.971567	0.976555	0.973893

Notes: This shows AUCs for different methods of 2−group comparison in multidimensional testing. $SS \equiv$ *Sample size*. EV t-test and UV t-test mean t-test with equal and unequal variances, respectively. EBP^{*} is hybrid-testing as in (30) and EBP^{**} is hybrid-testing as in (17).

Figure 2: Two-Group Simulation AUC Comparison

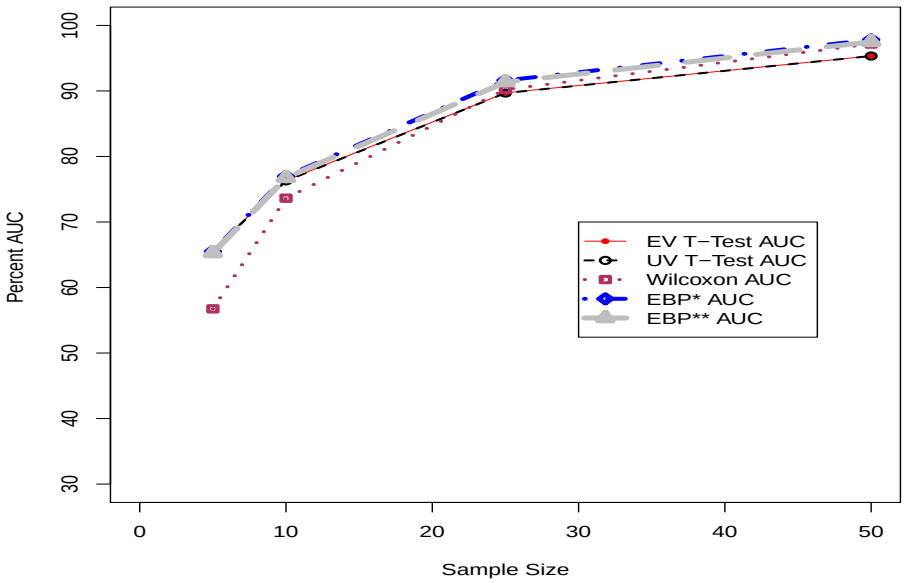


Table 3: Two-Group Simulation Power Comparison

SS	EV t-test	UV t-test	Wilcoxon test	EBP^*	EBP^{**}
5	0.004396	0.00139	0	0.065002	0.001136
10	0.134986	0.11498	0.116222	0.358476	0.137462
25	0.54672	0.540376	0.552902	0.722674	0.570982
50	0.780498	0.777618	0.802668	0.891926	0.813996

Notes: This shows powers for different methods of 2−group comparison in multidimensional testing. $SS \equiv Sample\ size$. EV t-test and UV t-test mean t-test with equal and unequal variances, respectively. EBP^* is hybrid-testing as in (30) and EBP^{**} is hybrid-testing as in (17).

Table 4: Two-Group Simulation AUC Comparison

SS	EV T-test	UV T-test	Wilcoxon	EBP^*	EBP^{**}
5	0.657408	0.656395	0.567454	0.655039	0.65254
10	0.763339	0.762761	0.736313	0.769609	0.767322
25	0.897038	0.89681	0.902189	0.916082	0.91352
50	0.953357	0.953298	0.971567	0.978162	0.974856

Notes: This shows AUCs for different methods of 2−group comparison in multidimensional testing. $SS \equiv Sample\ size$. EV t-test and UV t-test mean t-test with equal and unequal variances, respectively. EBP^* is hybrid-testing as in (18) and EBP^{**} is hybrid-testing as in (17).

Figure 3: Two-Group Simulation Powers Comparison

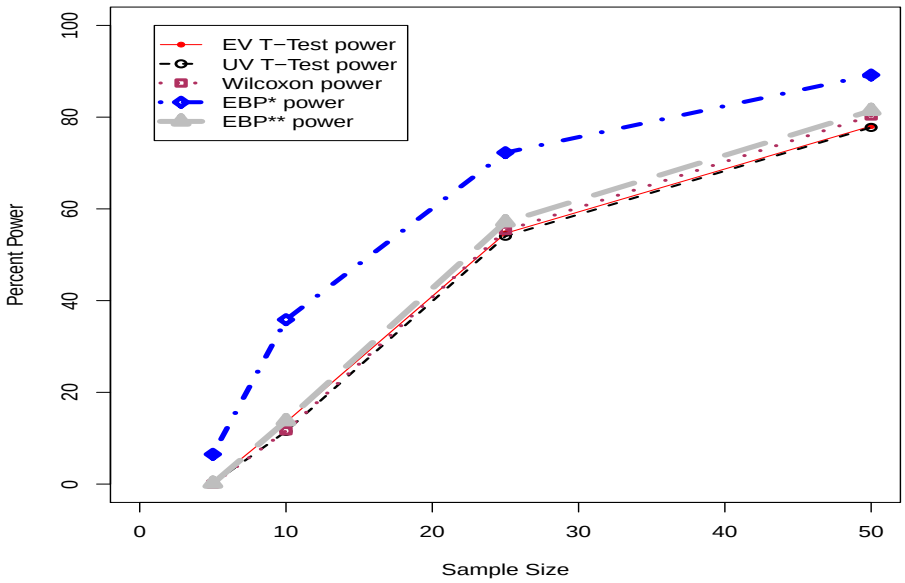
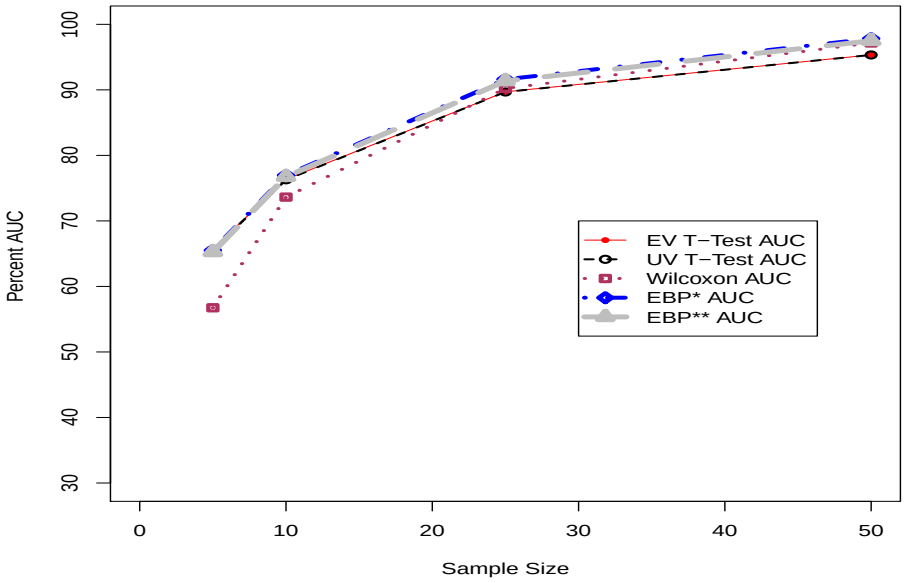


Figure 4: Two-Group Simulation AUC Comparison



Hybrid ANOVA-Kruskal-Wallis Simulation

We apply the hybrid-testing procedure to compare means in a three-group analysis. Consider the following hypothesis testings

$$H_{0g}\colon \mu_{1g} = \mu_{2g} = \cdots = \mu_{kg} \text{ vs } H_{ag}\colon \mu_{ig} \neq \mu_{jg},\ i \neq j \text{ for at least one } i \text{ and } j \text{ for each } g, \tag{31}$$

$g = 1, \cdots, G$. As in two group comparing analysis, we compare the performance of the hybrid-testing procedures with other methods using powers and AUCs. We make the comparison through two different setups. In the first setup, there are three groups with the sample varying from 5, 10, 25, 50. The number of null genes with the normal distribution is 720. The number of null genes with the Cauchy distribution is 80. The number of alternative genes is 20 for the Cauchy distributions and 180 for the normal distributions, with 200 alternative genes in total. The number of replications is 1000. The simulation results are presented in Table 5, and Figure 5; and in Table 6, and Figure 6. The results show that, in most of the settings, hybrid-testing procedures outperform the other methodologies. Figure 5 shows hybrid-testing procedures (EBP^* and EBP^{**}) to be more powerful than the Kruskal-Wallis test and the ANOVA test; the corresponding numerical results are presented in Table 5. Figure 6 shows that the hybrid-testing procedures (EBP^* and EBP^{**}) have greater AUCs than all other methods; the corresponding results are presented in Table 6.

Table 5: Three-Group Simulation Power Comparison

Sample size	ANOVA	Kruskal	EBP^*	EBP^{**}
5	0.00267	0.00011	0.001575	0.00156
10	0.08803	0.047405	0.08517	0.105845
25	0.62652	0.685965	0.713885	0.70887
50	0.90504	0.94124	0.961455	0.945485

Notes: This shows powers for different methods of $k-$group comparison in multidimensional testing. EBP^* is hybrid-testing as in (16) and EBP^{**} is hybrid-testing as in (17).

Figure 5: Three-Group Simulation Power Comparison

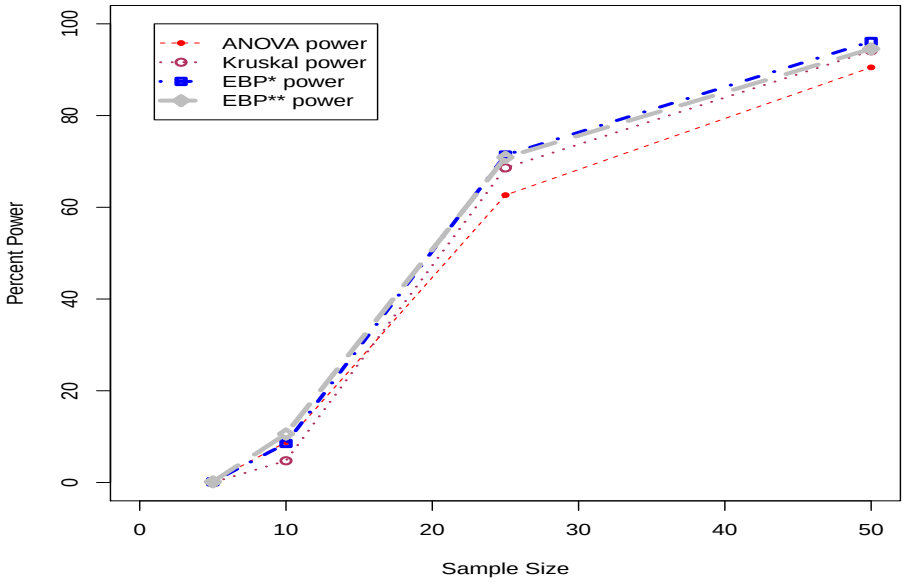
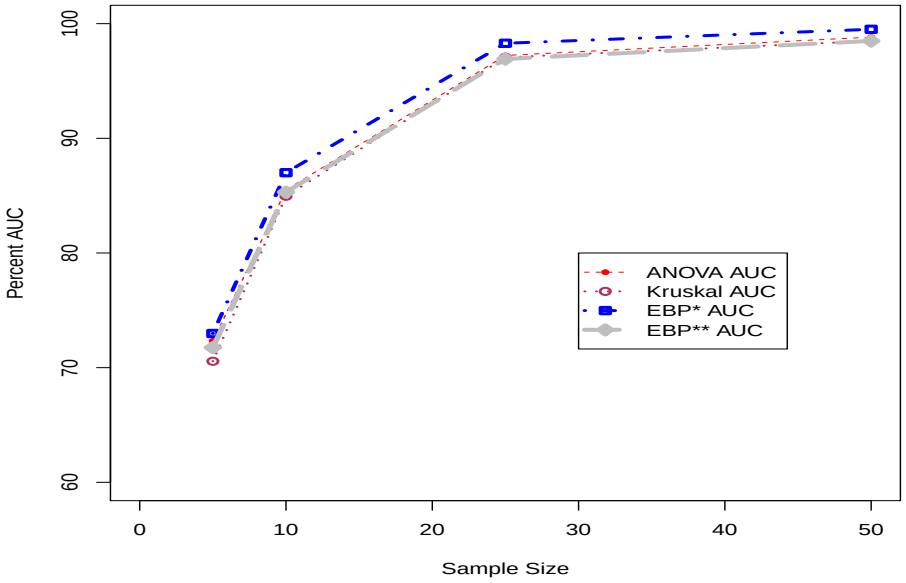


Table 6: Three-Group Simulation AUC Comparison

Sample size	ANOVA	Kruskal	EBP^{*}	EBP^{**}
5	0.723323	0.705571	0.729793	0.717486
10	0.85501	0.849496	0.870027	0.853085
25	0.972265	0.970622	0.982884	0.969216
50	0.988244	0.984893	0.995077	0.984916

Notes: This shows AUCs for different methods of $k-$group comparison in multidimensional testing. EBP^{*} is hybrid-testing as in (16) and EBP^{**} is hybrid-testing as in (17).

Figure 6: Three-Group Simulation AUC Comparison



A second simulation, similar to the first simulation, is performed with the Cauchy distribution replaced by the Log-normal distribution. The results are presented in Table 7, and Figure 7; and in Table 8, and Figure 8. Our proposed methodologies are again more powerful than the other procedures in most of the cases.

Hybrid Pearson-Spearman

Consider a simple regression

$$\mathbf{Y} = \beta\mathbf{X} + \varepsilon, \tag{32}$$

where $\mathbf{Y}$ is a $G \times n$ matrix where each row constitutes gene expression data, G is the number of genes, n is the sample size.

Table 7: Three-Group Simulation Power Comparison

Sample size	ANOVA	Kruskal	EBP^{*}	EBP^{**}
5	0.003705	8.00E-05	0.0023	0.00267
10	0.12584	0.05036	0.12057	0.132475
25	0.700315	0.685965	0.722465	0.72113
50	0.935525	0.94072	0.95545	0.94422

Notes: This shows powers for different methods of $k-$group comparison in multidimensional testing. EBP^{*} is hybrid-testing as in (16) and EBP^{**} is hybrid-testing as in (17).

Figure 7: Three-Group Simulation Power Comparison

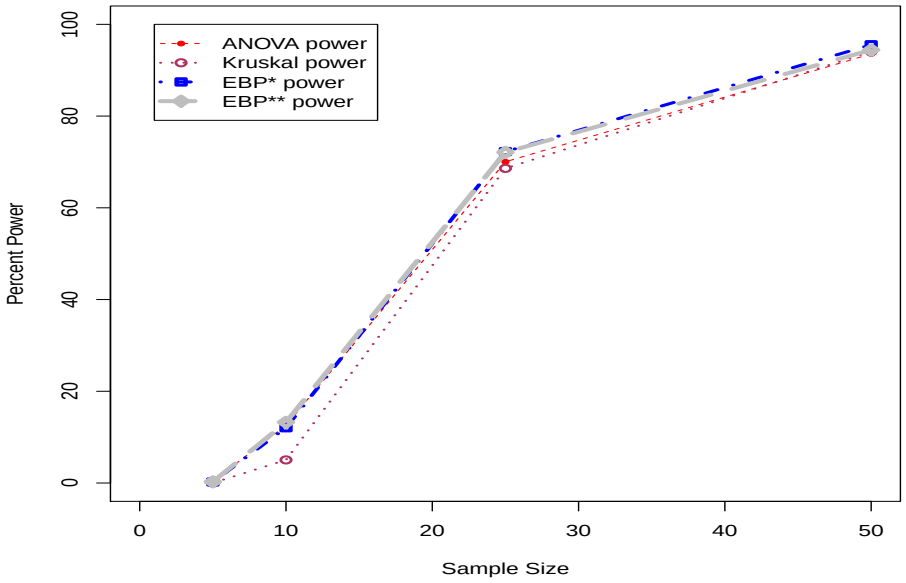


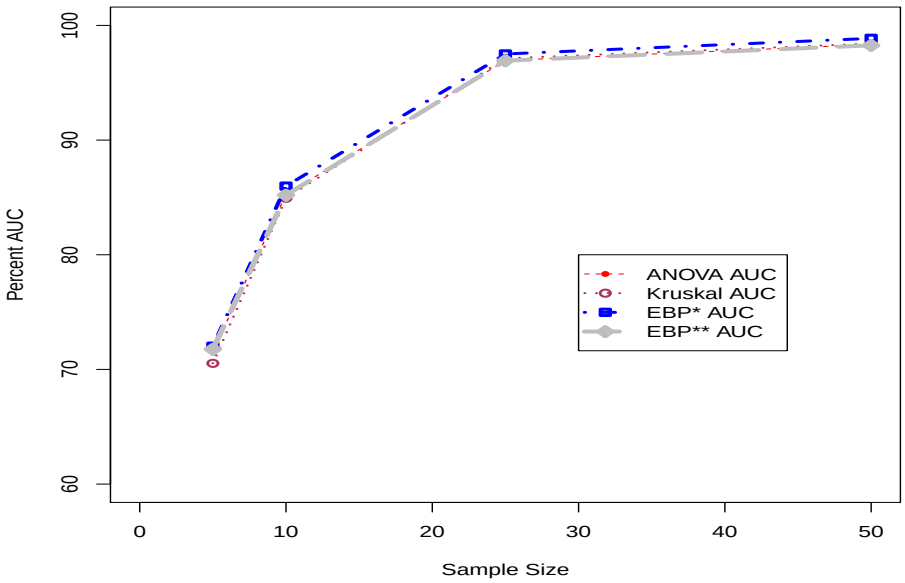
Table 8: Three-Group Simulation AUC Comparison

Sample size	ANOVA	Kruskal	EBP^{*}	EBP^{**}
5	0.715722	0.7053	0.720206	0.717601
10	0.85251	0.849279	0.85984	0.852012
25	0.969388	0.971307	0.975089	0.969274
50	0.983381	0.984229	0.988712	0.98259

Notes: This shows AUCs for different methods of $k-$group comparison in multidimensional testing. EBP^{*} is hybrid-testing as in (16) and EBP^{**} is hybrid-testing as in (17).

And $\mathbf{X}' = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_{10} & \beta_{11} \\ \beta_{20} & \beta_{21} \\ \vdots & \vdots \\ \beta_{G0} & \beta_{G1} \end{bmatrix}$, and $\varepsilon = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1n} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_{g1} & \varepsilon_{g2} & \cdots & \varepsilon_{gn} \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_{G1} & \varepsilon_{G2} & \cdots & \varepsilon_{Gn} \end{bmatrix}$, are matrices of

Figure 8: Three-Group Simulation AUC Comparison



sample covariates, parameters, and random errors with means of zeros and variances of ones, respectively. The hypothesis testings can be summarized as

$$H_{0g}: \beta_{g1} = 0 \text{ vs } H_{ag}: \beta_{g1} \neq 0. \tag{33}$$

We perform two different setups to compare hybrid-testing procedures with the other methodologies. In the first setup, the number of null genes with the normal distribution are 720, number of null genes with the Cauchy distribution are 80. The number of alternative genes is 20 for the Cauchy distribution and 180 for the normal distribution, 200 alternative genes in total, and the number of replications is 1000. The sample size varies from 5, 10, 25, 50. We compare hybrid-testing methodologies (EBP^* and EBP^{**}) with Pearson and Spearman tests using powers, and AUCs. The results are presented in Table 9, and Figure 9; and in Table 10, and Figure 10. The results show that our procedures perform better than other methodologies in most of the cases.

Table 9: Regression Simulation Power Comparison

Sample size	Pearson	Spearman	EBP^*	EBP^{**}
5	0.00094	0	0.000855	0.00115
10	0.084065	0.064045	0.081235	0.10014
25	0.42082	0.46567	0.47968	0.49815
50	0.497415	0.70289	0.693565	0.702665

Notes: This shows powers for different methods of correlation in multidimensional testing. EBP^* is hybrid-testing as in (16) and EBP^{**} is hybrid-testing as in (17).

Figure 9: Regression Simulation Power Comparison

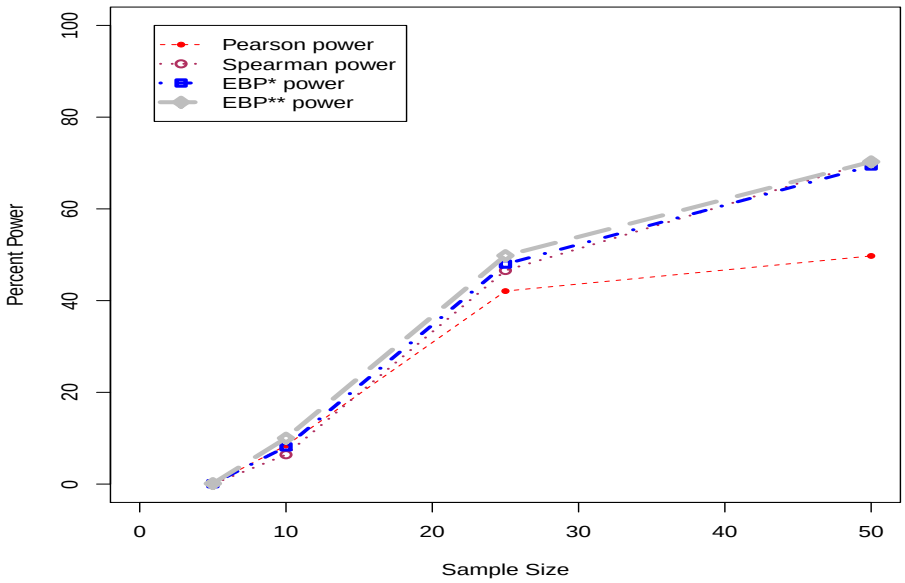


Table 10: Regression Simulation AUC Comparison

Sample size	Pearson	Spearman	EBP^{*}	EBP^{**}
5	0.69203	0.546736	0.691299	0.701406
10	0.824858	0.803001	0.844474	0.859245
25	0.926474	0.958733	0.970049	0.970106
50	0.96073	0.97999	0.989255	0.981823

Notes: This shows AUCs for different methods of correlation in multidimensional testing. EBP^{*} is hybrid-testing as in (16) and EBP^{**} is hybrid-testing as in (17).

A second simulation, similar to the first simulation is conducted where the Cauchy distribution is replaced by the $Log-normal$ distribution. The results are presented in Table 11, and Figure 11; and in Table 12, and Figure 12. Again, the results show that our proposed methods are more powerful than other methodologies in most of the settings.

Figure 10: Regression Simulation AUC Comparison

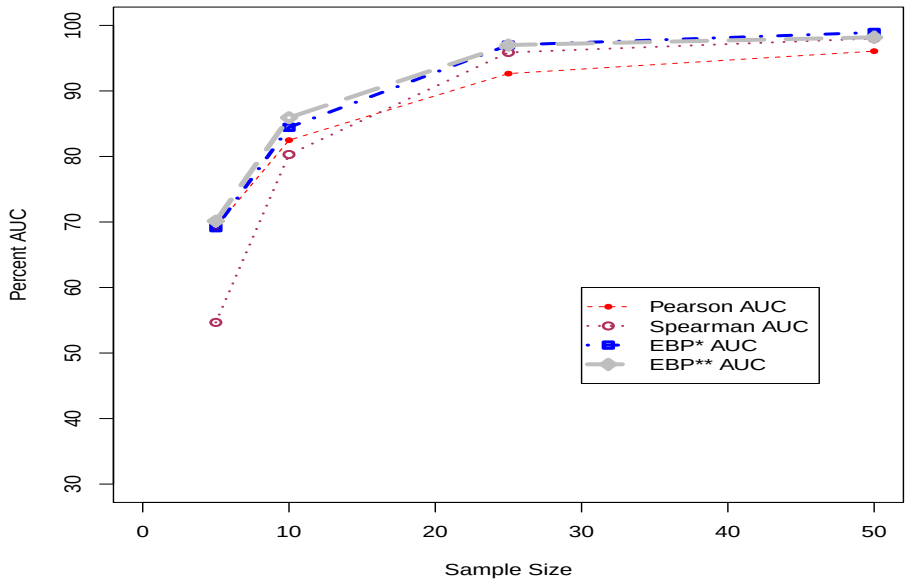


Table 11: Regression Simulation Power Comparison

Sample size	Pearson	Spearman	EBP^*	EBP^{**}
5	0.00423	0	0.003606	0.00479
10	0.16786	0.13519	0.166945	0.183385
25	0.61661	0.71118	0.74184	0.746385
50	0.793995	0.93356	0.941415	0.935395

Notes: This shows powers for different methods of correlation in multidimensional testing. EBP^* is hybrid-testing as in (16) and EBP^{**} is hybrid-testing as in (17).

Table 12: Regression Simulation AUC Comparison

Sample size	Pearson	Spearman	EBP^*	EBP^{**}
5	0.694852	0.547953	0.694589	0.696258
10	0.830779	0.806726	0.851895	0.851087
25	0.9272	0.959929	0.970204	0.965763
50	0.960601	0.978931	0.987718	0.981027

Notes: This shows AUCs for different methods of correlation in multidimensional testing. EBP^* is hybrid-testing as in (16) and EBP^{**} is hybrid-testing as in (17).

Figure 11: Regression Simulation Power Comparison

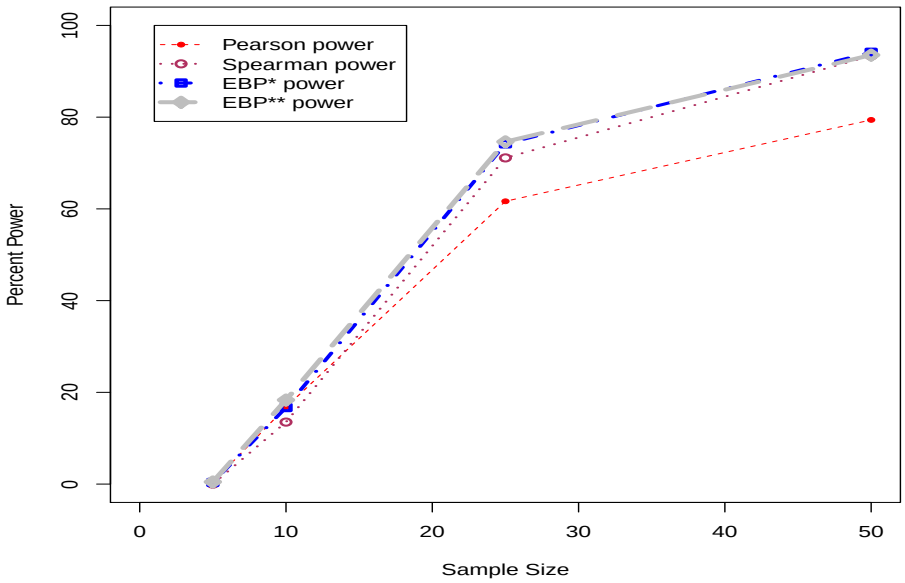
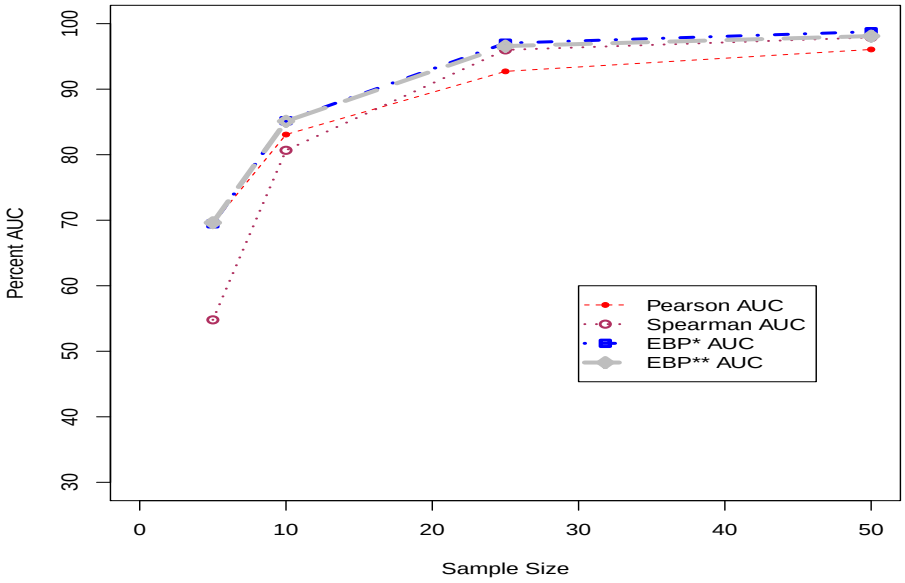


Figure 12: Regression Simulation AUC Comparison



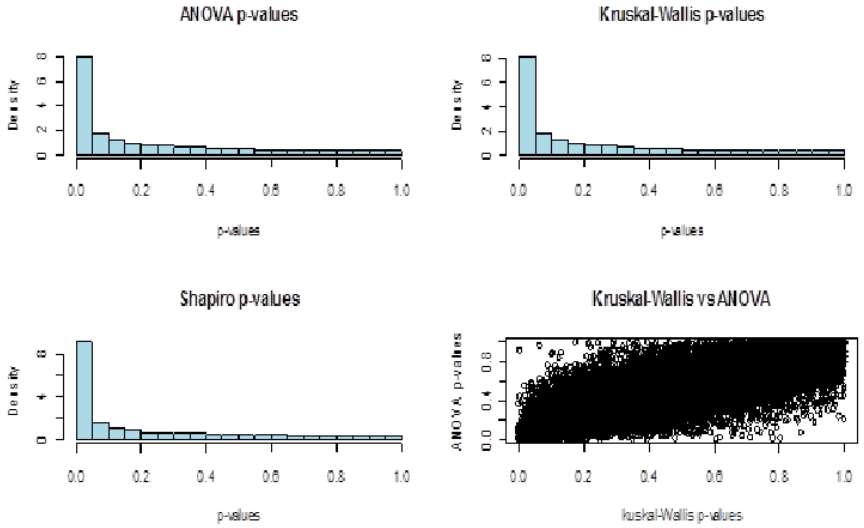
8 Applications

To further compare hybrid-testing methods with other procedures, we analyze the human ependymoma data, and the AML expression data along with the INHIBO data.

Comparison of 3-anatomically human ependymoma

The data sets are human ependymoma data from three distinct anatomic regions: Posterior Fossa (PF), Spine (SP) and Supratentorial (ST). The analysis consists of testing whether mean gene expression levels are equal in these three groups. From Figure 13, the histogram of the p-values of the Shapiro-Wilk test indicates that some of the expression data are normally distributed and others are not, since some p-values are less than 5% level of significance. Thus, the hybrid-testing procedure may be useful for analyzing the data. The box plot given in Figure 14 shows that the three groups of the gene 224132_{at} expression are different. Both the Hybrid ANOVA-Kruskal-Wallis test and the ANOVA test validate that finding, while the Kruskal Wallis test finds no difference among the three groups. Thus the Hybrid ANOVA-Kruskal-Wallis testing procedure performs better than the Kruskal-Wallis test when the normality assumption is valid. The Q-Q plot test for normality in Figure 15 accepts the validity of the normality of the data. Under Figure 14, we report the ANOVA p-value as 0.01 and the ANOVA EBP as 0.06, the Kruskal-Wallis p-value as 0.12 and the Kruskal-Wallis EBP as 0.26, the Shapiro-Wilk p-value as 0.93, and the hybrid-testing EBP as 0.06. With a threshold of τ equals to 0.1, both the hybrid-testing and the ANOVA reject the null hypothesis.

Figure 13: Three-Group Comparison Data



Association of Expression with DNA Synthesis Rate in AML Data

The study consists of investigating if gene expressions from a sample of patients are significantly correlated with the DNA Synthesis Rate (INHIBO). We conduct the analysis using Affymetrix arrays. Specifically, we wish to investigate the relationships between AML gene expression levels with INHIBO. To start with, we evaluate the distribution of p-values of tests for differential expression of the AML data and found in Figure 16 that some of the p-values are less than 0.05. Figure 17 and the Shapiro-Wilk test (p-value=0.172) show that the normality assumption does hold. A linear regression plot in Figure 18 indicates that the expression values of the gene $200081s_at$ are not correlated with INHIBO. An application of

Figure 14: Ependymoma Data: Analysis results

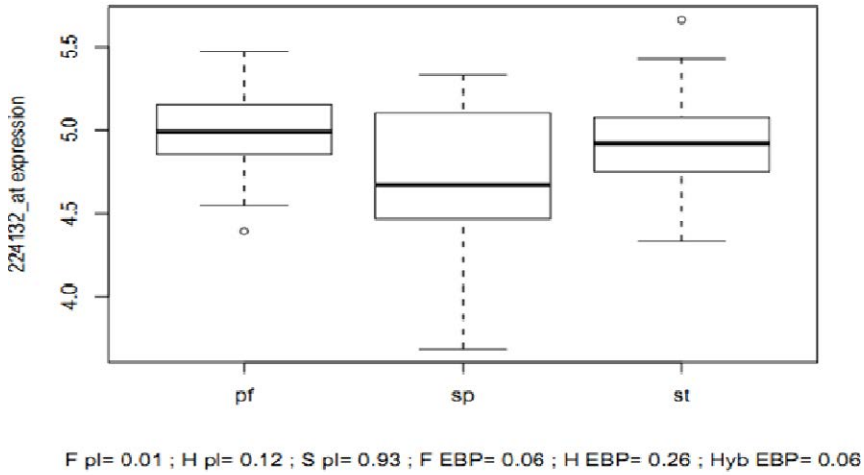
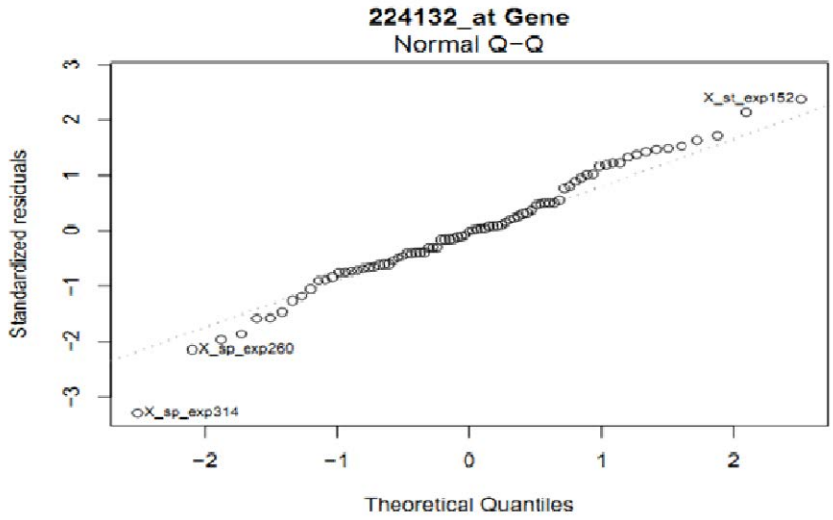


Figure 15: Ependymoma Data: Test for normality



Pearson test in this case accurately supports the hypothesis that the gene expression values are not linearly correlated with DNA synthesis rate. On the other hand, the nonparametric Spearman correlation marginally rejects this hypothesis and reduces the EBP of Pearson from 0.861 to an EBP of 0.152. Using hybrid-testing EBP of 0.506, thus the hybrid-testing performs better than the use of Spearman test alone.

9 Conclusion

We introduce hybrid-testing procedures as a general class of methods that can incorporate procedure-selection and account for multiple-testing in a seamless manner. Theorem 3.1

Figure 16: Regression data

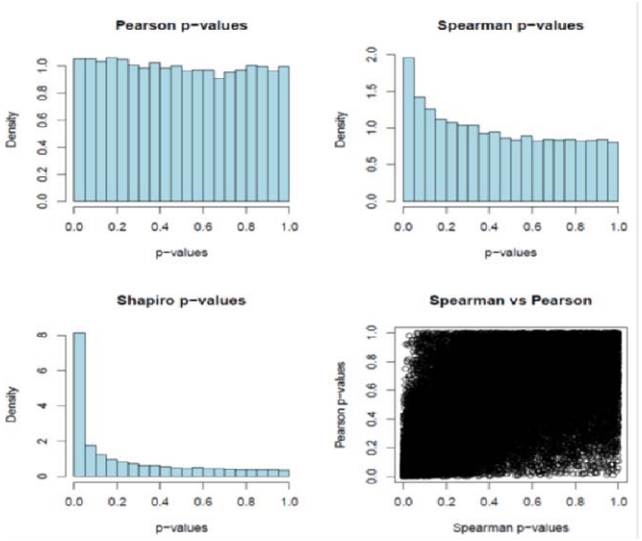
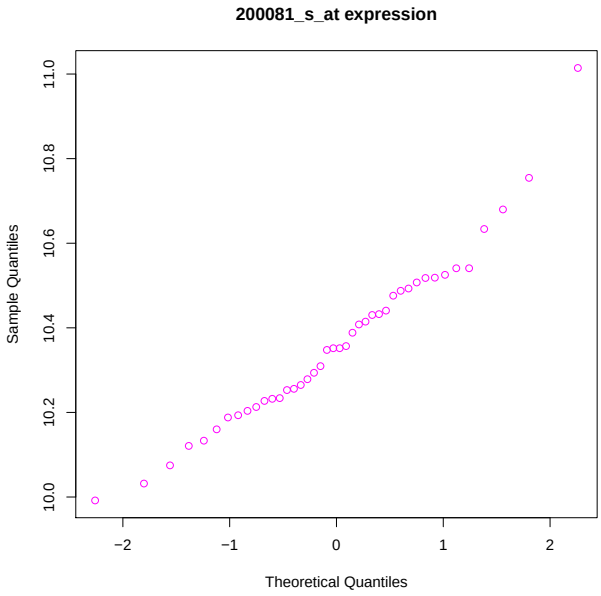


Figure 17: Explore Association of Expression with DNA Synthesis Rate: Normality diagnosis



provides a theoretical foundation for the use of the hybrid-testing procedure. In simulations and in real data analysis, we show that the hybrid-testing procedures perform well. In particular, we apply the hybrid-testing methodology to tumor data to compare expression of genes in three different groups, and to clinical data to study the relationship between AML expression data with INHIBO (DNA Synthesis Rate). The hybrid procedures have good performances in both applications compared to the ANOVA test and the Kruskal-Wallis test for the 3-group comparison analysis and to the Spearman and Pearson tests for the regression analysis.

Figure 18: Explore Association of Expression with DNA Synthesis Rate: Analysis results

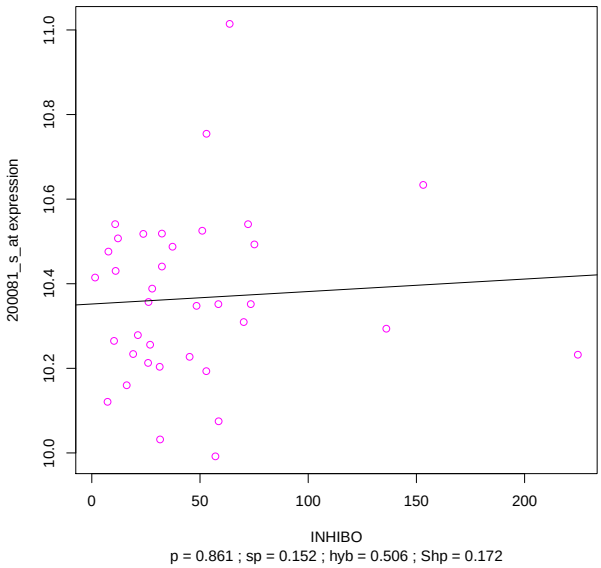
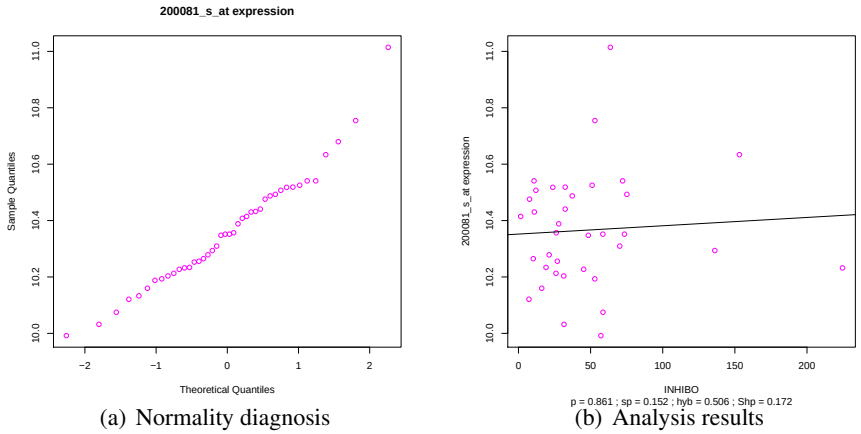


Figure 19: Association of Expression with DNA Synthesis Rate



(a) Normality diagnosis (b) Analysis results

For anyone interested in using our hybrid procedure there is an R package, "HybridMtest", available on the Bio-conductor. It provides, p-values and EBPs from all the methods including the hybrid-testing procedures.

References

[1] Benjamini, Yoav., Yosef Hochberg (1995): Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B (Methodological) 57:289-300.

[2] Besag, J. and D. Higdon (1999): Bayesian Analysis of Agricultural Field Experients. Journal of the Royal Statistical Society B, 61(4), 691-746.

[3] Besag, J., J. York and A. Mollié (1991): Bayesian Image Restoration with Two Applications in Spatial Statistics. The Annals of the Institute of Statistics and Mathematics, 43(1), 1-59.

[4] Dudoit, S, Y. Yang, M. J. Callow and T. P. Speed (2002): Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments. Stat. Sinica, 12, p. 111-139.

[5] Efron, Bradley (2004): Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Am Stat Assoc. 99:96-104.

[6] Efron, Bradley. and Robert Tibshirani (2002): Empirical Bayes Methods and False Discovery Rates for Microarrays. Genetic Epidemiology,23:70-86.

[7] Efron, Bradley., Robert Tibshirani, John D. Storey and Virginia Tusher (2001): Empirical Bayes analysis of a Microarray experiment. J Am Stat Assoc. 96:1151-1160.

[8] Fawcett, Tom. (2006): An introduction to ROC analysis. Pattern Recognition Letters, 27. 861-874.

[9] Gadbury, G.L. (2004): Power and sample size estimation in high dimensional biology. Stat.Meth.Med.Res., 14,325-338.

[10] Grenander, U (1956): On the theory of mortality measurement, Part II. Skan Aktuarietidskr, 39:125-153.

[11] Hochberg, Yosef (1988): A Sharper Bonferroni Procedure for Multiple Tests of Significance. Biometrika, 75 (4): 800802.

[12] Holm, S. (1979): A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 6570

[13] Langaas, Mette., Bo henry Lindqvist and Egil Ferkingstad. (2005): Estimating the proportion of true null hypotheses, with application to DNA microarray data. J. R. Statist. Soc. B 67, 555-5572.

[14] Lee, Duncan (2011): A Comparison of Conditional Autoregressive Models Used in Bayesian Disease Mapping. Spatial and Spatio-Temporal Epidemiology, 2(2), 79-89.

[15] Leroux, B., X. Lei and N. Breslow (1999): Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence. In ME Halloran, D Berry (eds.). Statistical Models in Epidemiology, the Environment, and Clinical Trials, pp. 135-178. Springer-Verlag. New York.

[16] Mantel, N. (1980): Assessing Evidence for Neoplasia. Biometrics, 45 .

[17] Mosig M.O, Lipkin E, Khutoreskaya G, Tchourzyna E, Soller M. (2001): Whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle,by means of selective milk DNA pooling in a daughter design,using an adjusted false discovery rate criterion. Genetics 157:1683-1698.

[18] Nettleton D, Hwang J.T.G, Caldo R.A and Wise R.P. (2006): Estimating the number of true null hypotheses from a histogram of p-values. J. Agri., Bio., Environ. Stat. 11:337-356.

[19] Pounds, Stan and Demba Fofana (2012): Hybrid Multiple Testing. *http://www.bioconductor.org/packages/2.12/bioc/html/HybridMTest.html*

[20] Pounds, Stan and Stephan W. Morris (2003): Estimating the occurrence of false positives and false negatives in Microarray studies by approximating and partitioning the empirical distribution of p-values. Bioinformatics. 19:1236-1242.

[21] Pounds, Stan and Shesh N. Rai (2009): Assumption adequacy averaging as a concept to develop more robust methods for differential gene expression analysis. Comput Stat Data Anal. 53(5): 1604-1612.

[22] Raftery, A. E., David Madigan and Jennifer A. Hoeting. (1997): Bayesian Model Averaging for Linear Regression Models. Journal of American Statistical Association, JASA, Vol. 92, No. 437.

[23] Robertson, T., F. T. Wright and R. L. Dykstra (1988): Order Restricted Statistical Inference. New York: Wiley.

[24] Royston, Patrick (1982): An extension of Shapiro and Wilk's W test for normality to large samples. Applied Statistics, 31, 115124.

[25] Schweder, T. and E. Spjøtvoll. (1982): Plots of p-values to evaluate many tests simultaneously. Biometrika, 69, 493-502.

[26] Shafer, G and I. Olkin. (1983): Adjusting P-Values to Account for Selection Over Dichotomies. Journal of American Statistical Association, JASA, Vol. 81, No. 826-831.

[27] Shapiro SS, Wilk MB. (1965): An analysis of variance test for normality (complete samples). Biometrika 52:591-611.

[28] Shojaie, Ali and George Michailidis (2009): Analysis of Gene Sets Based on the Underlying Regulatory Network. Journal of Computational Biology.

[29] Sidak, Z. K. (1967): Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association.

[30] Storey, John. (2002): A direct approach to false discovery rates. J. R. Statist. Soc. B 64, part3, 479-498.

[31] Strimmer, Korbinian (2008): fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics 1461-1462.

[32] Strimmer, Korbinian (2008): A unified approach to false discovery rate estimation. BMC Bioinformatics 9:303.

[33] Wilcoxon, F. (1945): Individual comparisons by ranking methods. Biometrika 1:80-83.