

Finite Skew-Mixture Models for Estimation of Positive False Discovery Rates

Gordon J. Bean¹, Elizabeth A. DeRose¹, Laina D. Mercer¹,
Laura K. Thayer¹, Anindya Roy¹ *and Subhashis Ghosal²

¹ Department of Mathematics and Statistics, UMBC, Baltimore, MD

² Department of Statistics, NCSU, Raleigh, NC

Abstract

We propose a mixture model framework for estimating positive false discovery rates in multiple testing problems. The distribution of a test statistic or a transformed p-value is modeled by a finite mixture of skewed distributions. We argue that a mixture of skewed distributions like the skew-normal is better capable of addressing some features in modeling than more commonly used mixture of normal distributions. Using the fitted distributions, we estimate the proportion of true null hypotheses, the positive False Discovery Rate and other important functionals in multiple testing problems. We investigate the performance of our methodology via extensive simulation and illustrate the effectiveness of the proposed procedure using real-data examples. We also discuss the roles of an empirical null in place of the theoretical null distributions in the context of common biomedical applications.

Key Words: Multiple testing ; p-value density; shape restriction

1. Introduction

Recent advances in biomedical research such as microarrays have made it necessary to consider testing many hypotheses simultaneously. To control the number of falsely rejected hypotheses, Benjamini and Hochberg (1995) introduced the false discovery rate (FDR) and described a procedure to control FDR. Various modifications of their procedure and extension to dependent test statistics were considered in the literature; see Sarkar (2002, 2004, 2006, 2007) for refinements. Storey (2002, 2003) considered a setting where hypotheses are true or false according to a random mechanism and looked at a slight modification of the FDR, called the positive False Discovery Rate (pFDR). Suppose H_{01}, \dots, H_{0m} are m null hypothesis that are being tested. Let H_r denote the indicator that H_{0r} is true for $r = 1, \dots, m$. Also, let I_r denote the indicator that H_{0r} is rejected based on the observed test statistic and p_r denote the p-value obtained from the r th test. Assuming the H_1, \dots, H_m are independent and identically distributed (i.i.d.) as Bernoulli(π_0), Storey showed that the pFDR at nominal level α is given by

$$\text{pFDR}(\alpha) = \text{P}(H_r = 1 | I_r = 1) = \frac{\pi_0 \alpha}{F(\alpha)}, \quad (1)$$

*Corresponding author: anindya@umbc.edu; phone: 410-455-2435; fax: 410-455-1066

where F is the marginal distribution of the p-values p_1, \dots, p_m . For null hypotheses which are true, the p-value may be thought to be uniformly distributed on the unit interval. Thus, F contains a uniform component with weight π_0 . Using (1), Storey (2002, 2003) introduced an estimation based approach for maintaining pFDR under a desired level in a given multiple testing problem. The approach consists of estimating the null proportion π_0 and $F(\alpha)$ using the empirical distribution of p-values and then fixing a rejection region so that the estimated pFDR falls within a given tolerable limit.

An appropriate model for p-values may increase the efficiency of estimation compared to the empirical procedure. Given that the distribution under the alternative is generally unknown and false null hypotheses need not come from a fixed alternative, specific parametric assumptions would be too restrictive and prone to misspecification. On the other hand, the estimation error in nonparametric estimators of the marginal p-value density may be larger to mitigate the advantages gained from the robustness of the nonparametric procedures. Mixture models provide a nice compromise between the competing priorities of efficiency of the parametric procedures and the flexibility of the nonparametric procedures.

Tang, Ghosal and Roy (2007) proposed a beta mixture model for the marginal p-value density and used a Bayesian approach based on Dirichlet process mixture model for estimating the mixture. While the beta mixture model is natural for p-value distributions in common situations and gives improved estimates, sometimes it may be necessary to consider domain beyond the unit interval, such as when a test statistic is directly modeled, rather than p-values. Also being restricted to the interval $[0, 1]$ makes it hard to generalize to a multivariate setting, where dependence among test statistics is allowed. The availability of various multivariate families with flexible correlation structure such as the multivariate normal family allows flexible joint modeling of statistics taking values in the entire real line. If p-values are considered as statistics, appropriate transformations can map them to the entire real line and then use a multivariate mixture model to fit the transformed p-values and estimate functionals such as the pFDR based on the fitted mixture model may be used on the transformed p-values. Efron (2004, 2007) suggested using the probit transform of the p-values, to be called probit p-values, and then fitting a Gaussian mixture model on them. A somewhat unpleasant consequence of this approach is that the resulting p-value density on the original scale will have a bowl-shaped density. This is in sharp contrast with Propositions 1 and 2 of Ghosal, Roy and Tang (2007), where it was shown that the p-value density is decreasing if the test statistic has the monotone likelihood ratio property. It is desirable to consider a flexible mixture model with a kernel function that will allow the original p-values to have decreasing density, by appropriately restricting parameter values. The skew-normal family, which contains the normal family as a special case, is able to produce decreasing densities on the original scale of p-values if the skewness parameter is considerably bigger than zero depending on the value of the scale parameter; see equation (4). For problems where the test statistic has a continuous distribution and the null hypothesis is simple, the null distribution of the p-value is Uniform $[0,1]$. Because the distribution of a Uniform $[0,1]$ random variable under the probit transformation is standard normal, the null distribution of probit p-values for problems where the test statistics has a continuous density is the standard normal distribution. In this article we concentrate only on problems where the test statistics has a continuous density with respect to the Lebesgue measure. To obtain a model for the probit p-value distribution which contains the null distribution as a special case, a skew-normal mixture model can be considered, or more generally, such a model can be used for test statistics with normal null distributions. Further, the skewness adds another dimension, which boosts the

flexibility of mixtures to handle occasional extra skewness in the distribution of the transformed p-value under the alternative. The approach may be extended to other skew distributions, such as the skew-logistic distribution, which, for instance, will be appropriate for modeling logit transform of p-values.

In this paper, we develop a method of estimating pFDR based on finite skew-mixture models. We employ the Expectation-Maximization (EM) Algorithm [Dempster, Laird, and Rubin (1977)] for estimating the parameters of the mixture model. We also incorporate the empirical null distributions of the p-values within the mixture model framework, which generally give a more reliable estimate of the pFDR.

The paper is organized as follows: In Section 2, we describe the skew-mixture model and the corresponding estimation procedure. In Section 3, we report results of a simulation study. In Section 4, we apply the proposed methodology on some important biomedical data sets and discuss issues regarding the theoretical versus the empirical null distribution of the p-values. We conclude the paper with a discussion in Section 5.

2. Skew-mixture models

Due to the range restriction of the p-values, it is more convenient to model a transformed version of the p-values and capture the salient features of the distribution using a flexible mixture model. There are common transformations such as logit and probit that one might consider to make the range of the variable unrestricted. In general, consider an absolute continuous density g on the entire real line. Let G denote the corresponding cumulative distribution function (c.d.f.). Then the transformation $X = G^{-1}(p)$, where p is the p-value, will be used for modeling. Thus, for our modeling exercise the observed quantities will be the m transformed p-values, X_1, \dots, X_m . Assuming that the distribution of the p-value is uniformly distributed over $[0, 1]$ under the null hypothesis, the null distribution of the transformed p-value will be g . Our objective of modeling the marginal distribution of X as a mixture of densities imposes the condition that the mixture family must contain the null density g as a candidate. For the subsequent development it will be easier to let g be a symmetric density about zero. The reason for such a choice is that we could use that as our basic kernel which is then modified to give the density of the p-value under the alternative hypothesis. The modification that we consider here is skewing the kernel g to reflect extra features in the p-value density under the alternative. For each such kernel g , one can formulate a skewed version as

$$q_g(x; \mu, \sigma, \lambda) = 2\sigma^{-1}g\left(\frac{x - \mu}{\sigma}\right)G\left(-\lambda\left(\frac{x - \mu}{\sigma}\right)\right),$$

which include the density g as a special case when the parameters are $(\mu, \sigma, \lambda) = (0, 1, 0)$. Thus, only under the null hypothesis the density of the transformed p-value X is symmetric about zero. We will use a finite mixture (in terms of the location μ , scale σ and the skewness parameter λ) of the skewed density q_g to model the marginal density of the transformed p-values X . Specifically, if $h(x; \theta)$ denotes the marginal density of X , then we will represent h as

$$h(x; \theta) = \pi_0 g(x) + \sum_{j=1}^k \pi_j q_g(x; \mu_j, \sigma_j, \lambda_j), \quad (2)$$

where $\theta = (\pi_0, \pi_1, \dots, \pi_k, \mu_1, \dots, \mu_k, \sigma_1, \dots, \sigma_k, \lambda_1, \dots, \lambda_k)$ is the parameter. The parameter space is $\Theta = \mathcal{S}^k \times \mathbb{R}^k \times (\mathbb{R}^+)^k \times \mathbb{R}^k$, where $\mathcal{S}^k = \{(\pi_0, \pi_1, \dots, \pi_k) : 0 \leq \pi_0, \pi_1, \dots, \pi_k \leq 1, \pi_0 +$

$\sum_{j=1}^k \pi_j = 1$ denotes the k -dimensional simplex. As discussed in the introduction, there are some natural shape restriction in the p-value density that one may want to incorporate in the model. In particular, it is desirable that the model for the marginal density of X induces a density for the original p-values p which is decreasing. If we assume the density of X to be $q_g(x; \mu, \sigma, \lambda)$ then the p-value density in the original scale is

$$h_p(p) = q_g(G^{-1}(p); \mu, \sigma, \lambda) / g(G^{-1}(p)).$$

Since $(G^{-1}(p) - \mu) / \sigma$ is a monotone increasing function of p , deriving restriction on the parameter for decreasing p-value density is equivalent to deriving condition for the function $h_z(z; \mu, \sigma, \lambda) = \sigma^{-1} q_g(z; 0, 1, \lambda) / g(\mu + \sigma z)$ to be decreasing for all z . Thus we would like the set

$$\Theta_c = \{(\mu, \sigma, \lambda) : \frac{\partial}{\partial z} h_z(z; \mu, \sigma, \lambda) \leq 0 \text{ for all } z\}$$

to be non-empty and possibly large enough to provide enough flexibility in the mixture model under the restriction of decreasing p-value density. We adopt the maximum likelihood approach to estimate the parameter θ . If the shape restriction is imposed, then each $(\mu_j, \sigma_j, \lambda_j)$, $j = 1, \dots, k$, is required to belong to Θ_c . The decreasing property of the density of p under mixture distribution is ensured by the convexity of the class of decreasing densities. The parameter k , controlling the complexity of the model is also unknown and will have to be chosen based on the data.

Let $g_{\mu, \sigma}(x) = \sigma g(\mu + \sigma x)$ and let $\ell(x; \mu, \sigma) = \log \left[\frac{g(x)}{g_{\mu, \sigma}(x)} \right]$. Also let $H_g(x) = g(x) / [1 - G(x)]$ denote the hazard function. Then we may rewrite the restricted space as

$$\Theta_c = \{(\mu, \sigma, \lambda) : \frac{\partial}{\partial z} \ell(z; \mu, \sigma) \leq \lambda H(\lambda z) \text{ for all } z\}. \quad (3)$$

Even though the restricted space is implicitly defined, in specific examples one may reduce the restrictions to explicit restrictions on (μ, σ, λ) . Ghosal and Roy (2001) showed that when $g(x) = \phi(x)$, the standard normal density, then the decreasing p-value density is obtained as long as the mixture is supported on the set

$$\Theta_c = \{(\mu, \sigma, \lambda) : \sigma \geq 1; \lambda > \sqrt{\sigma^2 - 1}; \mu \leq \lambda \sigma^{-1} \varphi((\sigma^2 - 1) / \lambda^2)\} \quad (4)$$

where $\varphi(t) = \inf\{H_\phi(x) - tx : x \in \mathbb{R}\}$.

Another natural transformation that can be applied on the p-values to map $[0, 1]$ into \mathbb{R} is the logit transformation $p \mapsto \text{logit}(p) := \log(p / (1 - p))$, to be called the logit p-value. Under the logit transformation, the theoretical null distribution of the p-value is a standard logistic distribution with c.d.f. $L(x) = e^x / (1 + e^x)$ and p.d.f. $l(x) = L(x)(1 - L(x))$. The skewed version of the kernel is $v(x; \mu, \sigma, \lambda) = 2\sigma^{-1} l((x - \mu)\sigma) L(-\lambda(x - \mu) / \sigma)$. the corresponding mixture density for the logit p-value is

$$h(x; \theta) = \pi_0 l(x) + \sum_{j=1}^k \pi_j v(x; \mu_j, \sigma_j, \lambda_j), \quad (5)$$

Conditions for decreasing shapes of the density of p-value when the logit p-value has density given by a skew-logistic density $v(x; \mu, \sigma, \lambda)$ can be characterized in terms of (μ, σ, λ) .

THEOREM 1. Let X have density $v(x; \mu, \sigma, \lambda)$ and $p = L(X)$. Define

$$\tau(z; \sigma, \lambda) = \text{logit} \left\{ \frac{\sigma - 1 + \lambda L(\lambda z) + 2L(z)}{2\sigma} \right\} - \sigma z. \quad (6)$$

Then p has decreasing p.d.f. if and only if

$$\sigma \geq 1, \lambda \geq \sigma - 1, \mu \leq \mu^*(\sigma, \lambda), \quad (7)$$

where $\mu^*(\sigma, \lambda) = \inf\{\tau(z; \sigma, \lambda) : z \in \mathbb{R}\}$.

Proof. From (3) and the form of the logistic distribution, it follows that the p-value density is decreasing if and only if

$$1 - 2L(z) - \lambda(1 - L(-\lambda z)) - \sigma(1 - 2L(\mu + \sigma z)) \leq 0 \text{ for all } z, \quad (8)$$

and a necessary condition for the density to be decreasing is that $\sigma \geq 1$ and $\lambda \geq 0$. Under these restrictions, Condition (8) also gives that the density is decreasing if and only if $\mu \leq \tau(z; \sigma, \lambda)$ for all z . Algebraic calculations show that a necessary condition for $\inf_z \tau(z; \sigma, \lambda)$ to be finite is that $\lambda \geq \sigma - 1$. This is derived from the fact that unless $\lambda \geq \sigma - 1$, $\lim_{z \rightarrow \infty} \tau(z; \sigma, \lambda) = -\infty$, in which case there are no possible solutions for μ . If we define $z^*(\sigma, \lambda) := z^*$ as the value such that $\lambda L(\lambda z) + 2L(z) = \sigma + 1$ (which exists since $\lambda \geq \sigma - 1$), then from the form of τ we see that for each $\sigma \geq 1$ and $\lambda \geq \sigma - 1$, $0 < \tau(z; \sigma, \lambda) < \infty$ for $-\infty < z < z^*$. For $z \geq z^*$, the condition (8) is satisfied for all μ . The result now follows noting that the function τ is convex in the region $-\infty < z < z^*$ and hence admits a unique infimum over that region. \square

2.1. Parameter estimation and false discovery control

The maximum likelihood estimator (MLE) of $\theta = (\pi_0, \dots, \pi_k, \mu_1, \dots, \mu_k, \sigma_1, \dots, \sigma_k, \lambda_1, \dots, \lambda_k)$ is obtained by maximizing $\prod_{i=1}^m h(X_i; \theta)$ with respect to θ , where the density h is given by (2). If the decreasing shape restriction of the original p-value density is desired, then the maximization will have to be restricted to those θ -values for which $(\mu_j, \sigma_j, \lambda_j)$, $j = 1, \dots, k$, satisfy (4) and (7) respectively for skew-normal and skew-logistic mixtures. However, maximization under the constraint is cumbersome. If one wants to maintain the shape restriction hypothesis, then unrestricted solutions may be projected back to the restricted set to obtain solutions on the boundary. A convenient version of EM type algorithm for fitting skew-normal mixture is given in Lin, Lee and Yen (2007). They also propose a method of moment approach for selecting the initial values of the parameters. We follow their approach for choosing the initial values. No significant difference in accuracy of the fit between the standard EM version for fitting mixture model and the modified version given in Lin, Lee and Yen (2007) was found in our simulations. However, the convergence was faster while using the latter version.

Plugging in the MLE $\hat{\theta}$ of θ into model (2), the MLE of the density of the transformed p-values, $X = G^{-1}(p)$ is $h(x, \hat{\theta}) = \hat{\pi}_0 g(x) + \sum_{j=1}^k \hat{\pi}_j q_g(x; \hat{\mu}_j, \hat{\sigma}_j, \hat{\lambda}_j)$. Let the c.d.f of the distribution of $p = G(X)$ be $\hat{F}(\cdot)$ where X is distributed as $h(x, \hat{\theta})$. Then following (1) we can estimate the pFDR as

$$\widehat{\text{pFDR}}(\alpha) = \frac{\hat{\pi}_0 \alpha}{\hat{F}(\alpha)}.$$

In order to control pFDR at a level γ , one can choose the nominal level α for each test to satisfy $\widehat{\text{pFDR}}(\alpha) \leq \gamma$.

The number of components, $k + 1$, is a critical parameter in mixture model estimation. One can use a penalized approach like the Akaike Information Criterion (AIC) to make a data-dependent selection of k . But EM algorithm augmented with such a selection procedure has severe convergence problem and is very slow. Lin, Lee and Yen (2007) avoided incorporating estimation of the number of skew-normal components in the mixture by assuming the number of components to be known. Vlassis and Likas (2002) proposed a greedy version of EM algorithm that starting from a minimum number of components sequentially adds components to the mixture with a stopping rule based on the value of the likelihood. Although the methodology in Vlassis and Likas (2002) is designed for Gaussian mixtures, the extension to skew-mixture is straightforward. The methodology starts with a single component and conditional on the existing components finds a new 'best' component to add by doing a partial search EM. We followed this approach for our simulation with a stopping rule and a merging criterion. We used eight as the maximum number of components in the mixture. We also merged two components if the estimated parameters associated with two components were closer in Euclidean distance to some pre-specified threshold. We also merged any components with estimated probability less than some pre-specified small number to the components nearest to it in terms of the Euclidean distance between the estimated parameters. The exact values of the thresholds used in the simulation and in the examples are given in the next section. Based on our simulation, we found that the method for selection of number of components can substantially add to the computation time. A smaller scale comparative simulation (not reported here) of the method used and the method used in Lin (2007) found that choosing the number of components as fixed makes the computation much faster without sacrificing statistical accuracy provided the true model can be described adequately with the chosen number of components. Thus, in applications we recommend to fix the number components whenever such prior information is available. Other component selection methods for mixture distributions such as Ueda *et al.* (2000) may also be investigated in the context of skew-mixture models.

3. Numerical Results

For data simulation we use two different models taken from Tang, Ghosal and Roy (2007). We will describe them here for the convenience of the reader. The first is a two-way mixed model where the test of interest is the equality of the fixed effect across two groups and the alternative value is unknown but fixed for all the tests. Specifically, the model for the raw data (and hence the corresponding p -values) for the r^{th} test is

$$\text{Model I: } y_{ijk} = \mu + \tau_i + g_{ij} + \varepsilon_{ijk}, \text{ for } i = 1, 2, j = 1, \dots, 5, k = 1, \dots, 5,$$

where g_{ij} are random subject effects and assumed to be i.i.d. $N(0, 0.5)$ and the model errors ε_{ijk} are i.i.d. $N(0, 1)$. The τ_i are fixed effects (e.g. treatment effects) and the r th hypothesis is

$$H_{0,r} : \tau_1 = \tau_2 \text{ vs. } H_{1,r} : \tau_1 > \tau_2, \quad r = 1, 2, \dots, m.$$

Then a test for $H_{0,r}$ will be the α -level t-test based on the averages $z_{ij} = \frac{1}{5} \sum_k y_{ijk}$. The alternative value for the false null hypotheses is fixed at $\tau_1 - \tau_2 = 1.5$.

The second simulation model is a one-way model with mixture alternatives. Specifically, the data (test statistics) for the r^{th} hypothesis are generated from

$$\text{Model II : } T_i \sim N(\mu, 1),$$

and the tests of interest are $H_{0,r} : \mu = 0$ vs. $H_{1,r} : \mu > 0$. To generate the data for the false null hypotheses, the alternative value of μ is taken from a mixture distribution. The test statistic under the alternative model is simulated according to

$$T_i \sim 0.25 N(1, 1) + 0.4 N(1.5, 1) + 0.2 N(2, 1) + 0.1 N(2.5, 1) + 0.05 N(3, 1).$$

We performed limited simulation based on Model I and applied both skew-normal mixture and skew-logistic mixture methods for estimating parameters. The results from both mixtures were comparable. We augmented the procedure with selection of number of components. We used an upper bound of eight components for the simulation. Thus, the algorithm terminated if either adding an additional component to the mixture did not improve the likelihood or number of components reached eight. Further to reduce complexity of the estimated models, whenever two or more estimated components are similar in the sense that the Euclidean distance between the parameter estimates of location, scale and shape of the two components was less than a pre-specified quantity δ , these components were collapsed into a single component. In the present simulation, we used $\delta = 0.1$. Also, if any of the component probability π_j was estimated to be less than 0.0001, the component was merged with the component closest to it in Euclidean norm. The algorithm started with $k = 1$, i.e., only two component mixture with one $N(0, 1)$ component and another skew-normal component. The starting values of the skew-normal parameters were chosen by the method of moments approach described in Lin, Lee and Yen (2007).

Figure 1 shows the distribution of the number of components. In most cases the number of components giving the maximum value of the likelihood was between 3 and 5. However, in some cases the number of selected components was 7 or 8. A boxplot of the running time for convergence for each number of components is plotted in Figure 1. In one case, the algorithm was terminated with the number of components equal to eight and the algorithm did not converge. IN 6 of the total 3 of the total 100 cases the algorithm failed to converge or had difficulty converging when left running with running time ranging from 10-30 hours. Since there are $4k$ parameters in all for a k -component mixture, and maximum likelihood estimation of the skewness parameter λ can be unstable for small sample sizes, the method may suffer from the curse of dimensionality if k is too large.

In the simulation, we observe that the final mean squared error (MSE) for the proportion of true nulls, π_0 , and that for the pFDR were not strongly affected by the instability of estimation of λ . However, for situations where the maximum likelihood estimation of λ is unstable, we recommend either putting an upper bound on the skewness parameter or use a more stable estimator such as the one proposed by Sartori (2006) where a bias correction is made to the MLE of λ .

To study the empirical properties of the estimators, for both models, we set $m = 1000, 5000$ for the total number of hypotheses, $\pi_0 = 0.80, 0.90, 0.95$ for the proportion of true null hypotheses, and $\alpha = 0.005, 0.05, 0.01$ for the significance levels. To calculate $\hat{\pi}_0$, we carried out 500 iterations of each of the twenty-four simulations. Table 1 shows mean and root-MSE (rMSE) of $\hat{\pi}_0$ for probit and logit models, respectively. The estimates for Model I were slightly negatively biased but the bias decreases with increasing sample size. The overall mean squared error for both models were

reasonably small for both probit and logit p-values. The Monte-Carlo rMSEs compared favorably with those from other methods reported in Tang, Ghosal and Roy (2007). The pFDR estimates for the logit p-values were comparable to those for the probit p-values and we only report those for the probit p-values in Table 2 for the $m = 1000$ case. The values of the pFDR based on the proposed skew mixture method is given under the column labeled “Skew”. For comparison we also give the method based on Storey and Tibshirani (2003) that estimates pFDR by using a bootstrap cross-validation method. The estimates corresponding to that estimator is given under the column labeled “Boot”. The estimates are very similar for the two methods. However, the proposed method has the added advantage that of being able to provide a full description of the p-value distribution which in turn maybe used to compute other relevant measures in the multiple testing setting. The performance of the estimator seems to be a function of the significance level α with performance declining with smaller α .

It should be noted that the performance of the EM algorithm does depend on how accurate the initial values for the location, scale and skew parameters are. Thus, a data-dependent choice of the initial parameter values is advocated. We have followed the approach given in Lin, Lee and Yen (2007) but there is still room for improvement. For example, the method of moment estimator of the skewness parameter could be very unstable, resulting in unusually high initial values for λ . Methods such as bump-hunting algorithms may be also used to guess the initial values for the parameters.

4. Data Analysis

In this section we analyze three different data sets pertaining to gene expression. In the context of these examples we discuss some issues that arise in modeling data from real life multiple testing experiments.

4.1. Empirical Null

Generally, the theoretical null distribution of the p-values is $\text{Unif}[0, 1]$, making the probit p-values distributed as $N(0, 1)$ when the null hypothesis is true. As argued by Efron (2004), the theoretical null model may not be appropriate for the observed p-values in many real-life applications. He shows that a small difference between the theoretical null, $N(0, 1)$, and an empirical null can substantially affect the conclusions of significance findings. In large scale multiple testing situations, empirical estimation of the null distribution is typically possible, such as by using an appropriate parametric model or a nonparametric density estimation technique.

If the true null distribution of the p-values is uniform, then one of the components is the standard normal (logistic) distribution for the probit (logit) p-values, and hence we can define $(\mu_0, \sigma_0, \lambda_0) = (0, 1, 0)$. However, for cases where the null distribution needs to be estimated from the model we can let $(\mu_0, \sigma_0, \lambda_0)$ be additional parameters and estimated them as well using the maximum likelihood technique. Since in most applications where this methodology is relevant has a large proportion of true null hypotheses, we constraint the estimate of π_0 to be greater than 0.50.

We consider two data sets obtained from the National Center for Biotechnology Information (NCBI) database to demonstrate the effect of considering the empirical null distribution in the analysis. Since the issue is separate from the shape restriction of the p-value density, we do not impose the shape restriction for this exercise. We analyze each dataset in two ways — one where the

distribution under the true null hypotheses is constrained to be the theoretical null and another where the skew-mixture is fitted without any constraints. We check the similarity between the theoretical null and the empirical null, and evaluate the benefits of using an empirical null in cases where the p-value distribution appears to deviate from the theoretical null.

The first set of data comes from a study that compared a normal elderly control group of people to an Alzheimer’s disease group in order to identify genes with disease and gender expression patterns [Maes *et al.* (2007)]. Peripheral blood mononuclear cells were obtained from each group of seven female subjects and the expressive profiles were determined using the National Institute on Aging Human Mammalian Gene Collection cDNA microarray.

The constrained fit without the shape restriction for the Alzheimer’s data has one null component and three non-null components and the estimated null proportion is 0.914. The fitted constrained model is given by

$$0.913 N(0, 1) + 0.033 SN(-1.57, 0.44, 2.3) + 0.032 SN(-0.93, 0.13, 0.005) + 0.022 SN(-0.86, 0.12, -3.02),$$

where SN stands for the skew-normal distribution. The skew-mixture fit captures the sharp peak evident in the data, a feature that illustrates the potential and flexibility of the skew-mixture models. The sharper peak not only has a very rapid slope but also exhibits skewness in a smoothed version of the histogram. A Gaussian mixture model would require significantly more number of components to capture these features. However, from Figure 2(a), it is evident that there is room for improvement in the constrained fit near the smoother mode of the histogram. The unconstrained fit with five components gives a more satisfactory approximation to the histogram for this dataset. A closer examination of the fitted components reveals that the component with the largest proportion (0.724) is $SN(0.267, 0.895, 0.034)$. This component is stochastically larger than the theoretical null, which is the standard normal distribution. Another component which is similar to the theoretical null distribution ($SN(-0.225, 0.901, 0.0365)$) has a proportion of 0.177. In the constrained case, the algorithm is forced to fit only one theoretical null component in the region spanned by these two components and therefore does an inferior job of approximating the mixture of two non-normal components with proportions 0.724 and 0.177 by a single normal component with proportion 0.91. Even though the probit p-values corresponding to the two components are similar to those arising from true null distribution, this example clearly demonstrates the need for flexibility in modeling the null distribution of the probit p-values and hence that of the original p-values. The component $SN(0.367, 0.895, 0.034)$, which is stochastically larger than standard normal will result in a p-value density which shifted more to the right than the uniform density. Thus, under such a component, larger p-values are more likely to occur than under uniform distribution, and hence can be definitely considered as a model for the null p-values. Deviations from the uniform distribution can occur either when the p-values are dependent or if the testing situation is complicated and it is not possible to construct test statistics that gives rise to uniform p-values under null.

The second dataset is from an experiment performed to identify transcripts that are differentially expressed in the mammary gland at different stages of development in wild type mice [Ron *et al.* (2007)]. We used the data from virgin mice and pregnant mice. Three biological replicates were obtained from the virgin females and two biological replicates for the pregnant females. In contrast to the Alzheimer’s data, in the microarray data comparing mammary activity between virgin and pregnant mice, the constrained estimation method gives absurd fit from being forced to fit a standard normal component with probability greater than 0.25 where the unconstrained fit is also much

more adequate. With the constraint, the null proportion π_0 is estimated as 0.890, that is, 89% of the probit p-values come from the $N(0, 1)$ theoretical null. With the constraint lifted, 0.808 proportion comes from $SN(0.420, 1.337, 0.037)$ and 0.081 proportion comes from $SN(-0.404, 0.915, -4.065)$. The superiority of the unconstrained fit can be seen in Figure 2(b).

4.2. Application

We apply the proposed method to a dataset obtained from the NCBI database. We estimate the p-value density using the skew-mixture model, using the estimated model we then find the α -levels that would be needed to control pFDR at a chosen level and look at the resulting set of hypotheses that are considered discoveries after applying the pFDR control. The data is from a study of gene expression profiles in white blood cells in response to exercise [Buttner *et al.* (2007)]. Five male subjects performed an exhaustive treadmill test ET at 80% of their VO_{2max} until individual exhaustion. Blood samples (9ml) were drawn before and one hour past the tests. White blood cells were isolated by the erythrocyte lysis method. Gene expression profiles were measured using the Affymetrix GeneChip technology.

Figure 3(a) shows that the fitted model follows the data adequately. The proportion corresponding to the standard normal component is only 0.7. The fit can slightly improved in the unconstrained case. However, the two components that correspond to the null model ($SN(-0.05, 1, 0)$ and $SN(0, 1.02, 0.05)$) are almost identical with the standard normal component and are merged together. The modified fitted model has three components with a standard normal with mixing proportion equal to 0.7 and two other components with parameters $(-2.06, 1.16, -1.98)$ and $(-2.97, 1.03, 1.94)$ and mixing proportions 0.19 and 0.11, respectively. Again, the fitted values belong to the restricted set corresponding to decreasing density. Figure 3(b) gives the estimated pFDR from the fitted model for a range of values of α . For this data the value of the nominal level α needed to control pFDR under 20% is given by 0.096 and that for a pFDR of 10% is 0.035. A histogram plot of the original test statistic is given in Figure 3(c) and the values corresponding to significant findings are highlighted.

5. Conclusions

Skew-mixture models of transformed p-values provide a flexible framework for estimating important quantities in single-step multiple testing situations and the estimates obtained from the skew-mixture models are generally more as efficient as those obtained from mixture models that model the p-value directly. Thus, in terms of parameter efficiency the methodology compares favorably with other methods such as the beta-mixture method of Tang, Ghosal and Roy (2007) but the proposed procedure has the distinct advantage that the data are not restricted to $[0, 1]$. In our investigation, we found that the observed null distribution can be substantially different from the assumed theoretical null distribution. If the observed null is stochastically larger than the theoretical null, the resulting pFDR estimate is a more conservative estimator of the true pFDR. This fact is important in controlling overall error. Another important feature of the proposed method is that shape restriction on the p-value densities can be imposed on the obtained estimators.

Acknowledgments

Research of Anindya Roy was partially supported by NSF grant number DMS-0803531 and that of Subhashis Ghosal by NSF grant number DMS-0803540. All other authors were supported by NSF REU grant number DMS-0354034.

References

- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc., Ser. B*, **57**, 289–300.
- Buttner, P., Mosig, S., Lechtermann, A., Funke, H. and Mooren, F. C. (2007). Exercise affects the gene expression profiles of human white blood cells. *J. Appl. Physiol.*, **102**, 26–36.
- Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc., Ser. B*, **39**, 1–38.
- Efron, B. (2004). Large-scale simultaneous hypothesis testing. *J. Amer. Statist. Assoc.* **99**, 96–104.
- Efron, B. (2007). Size, power and false discovery rates. *Ann. Statist.*, **35**, 1351–1377.
- Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.*, **32** 1035–1061.
- Ghosal, S. and Roy, A. (2011). Predicting false discovery proportion under dependence. *J. Amer. Statist. Assoc.*, **106**, 1208–1218.
- Ghosal, S., Roy, A. and Tang, Y. (2007). Posterior consistency of Dirichlet mixtures of beta densities in estimating positive false discovery rates. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* (Balakrishnan, N. et al., Eds.) IMS Collection **1**, Institute of Mathematical Statistics, Breechwood, OH.
- Lin, T. I., Lee, J. C. and Yen, S. Y. (2007). Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, **17**, 909–927.
- Maes, O. C., Xu, S., Yu, B., Chertkow, H. M., Wang, E. and Schipper, H. M. (2007). Transcriptional profiling of Alzheimer blood mononuclear cells by microarray. *Neurobiol Aging*, **28**, 1795–809.
- Ron, M., Israeli, G., Seroussi, E., Weller, J. I., Gregg, J. P., Shani, M. and Medrano, J. F. (2007). Combining mouse mammary gland gene expression and comparative mapping for the identification of candidate genes for QTL of milk production traits in cattle. *BMC Genomics*, **20**, 183.
- Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.*, **30**, 239–257.
- Sarkar, S. K. (2004). FDR-controlling stepwise procedures and their false negatives rates. *J. Statist. Plann. Inf.*, **125**, 119–137.

- Sarkar, S. K. (2006). False discovery and false non-discovery rates in single-step multiple testing procedures. *Ann. Statist.*, **34**, 394–415.
- Sarkar, S. K. (2007). Stepup Procedures Controlling Generalized FWER and Generalized FDR. *Ann. Statist.*, **35**, 2405–2420.
- Sartori, N. (2006) Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *J. Statist. Plann. Inf.*, **136**, 4259–4275.
- Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc., Ser. B*, **64**, 479–498.
- Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.*, **31**, 2013–2035.
- Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences USA*, **100**, 9440–9445.
- Tang, Y., Ghosal, S., and Roy, A. (2007). Nonparametric Bayesian estimation of false discovery rates. *Biometrics*, **63**, 1126–1134.
- Ueda, N., Nakano, R., Ghahramani, Z., and Hinton, G. E. (2000). SMEM algorithm for mixture models. *Neural Computation*, **12**, 2109–2128.
- Vlassis, N. and Likas, A. (2002). A greedy EM algorithm for Gaussian mixture Learning. *Neural Processing Letters*, **15**, 77–87.

Table 1: Estimation of π_0 using probit and logit p -values. Subscripts a and b denote mean and $rMSE$, respectively.

π_0	Probit				Logit			
	Model I		Model II		Model I		Model II	
	$m = 1000$	$m = 5000$	$m = 1000$	$m = 5000$	$m = 1000$	$m = 5000$	$m = 1000$	$m = 5000$
0.80	0.782 ^a	0.790	0.840	0.827	0.775	0.777	0.846	0.839
	0.019 ^b	0.017	0.061	0.046	0.024	0.022	0.052	0.044
0.90	0.872	0.877	0.900	0.907	0.866	0.873	0.899	0.902
	0.031	0.022	0.033	0.019	0.042	0.023	0.029	0.022
0.95	0.915	0.932	0.926	0.937	0.908	0.919	0.930	0.934
	0.052	0.026	0.048	0.024	0.059	0.030	0.036	0.026

Table 2: Probit $pFDR$ estimates for $m = 1000$. The values for the proposed method are given under “Skew” and that for Storey’s method based on bootstrap cross-validation are given under ‘Boot’. Subscripts a and b denote mean and $rMSE$, respectively.

π_0/α	Model I			Model II		
	true pFDR	<i>Skew</i>	<i>Boot</i>	true pFDR	<i>Skew</i>	<i>Boot</i>
0.8/0.005	0.035	0.035 ^a	0.035	0.087	0.093	0.090
		0.002 ^b	0.002			
0.8/0.01	0.055	0.054	0.053	0.127	0.135	0.129
		0.002	0.002			
0.8/0.05	0.177	0.174	0.176	0.287	0.311	0.295
		0.006	0.006			
0.9/0.005	0.076	0.075	0.074	0.178	0.182	0.179
		0.006	0.006			
0.9/0.01	0.116	0.117	0.116	0.247	0.252	0.241
		0.008	0.007			
0.9/0.05	0.327	0.311	0.312	0.475	0.481	0.460
		0.018	0.019			
0.95/0.005	0.148	0.144	0.145	0.314	0.326	0.306
		0.016	0.015			
0.95/0.01	0.217	0.210	0.209	0.409	0.413	0.416
		0.020	0.021			
0.95/0.05	0.476	0.476	0.474	0.657	0.633	0.642
		0.040	0.040			

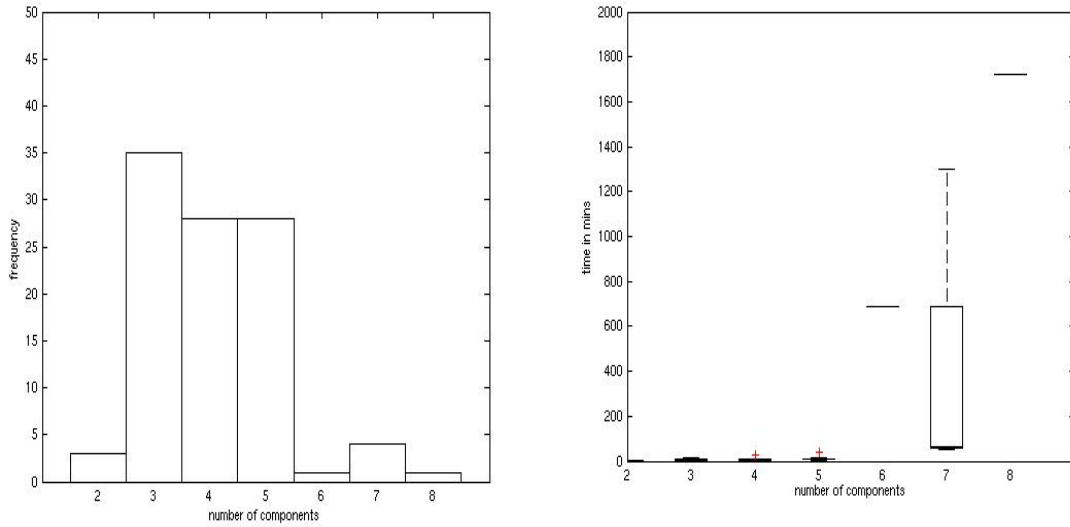


Figure 1: Distribution of number of components and average time for 100 simulations based on Model I

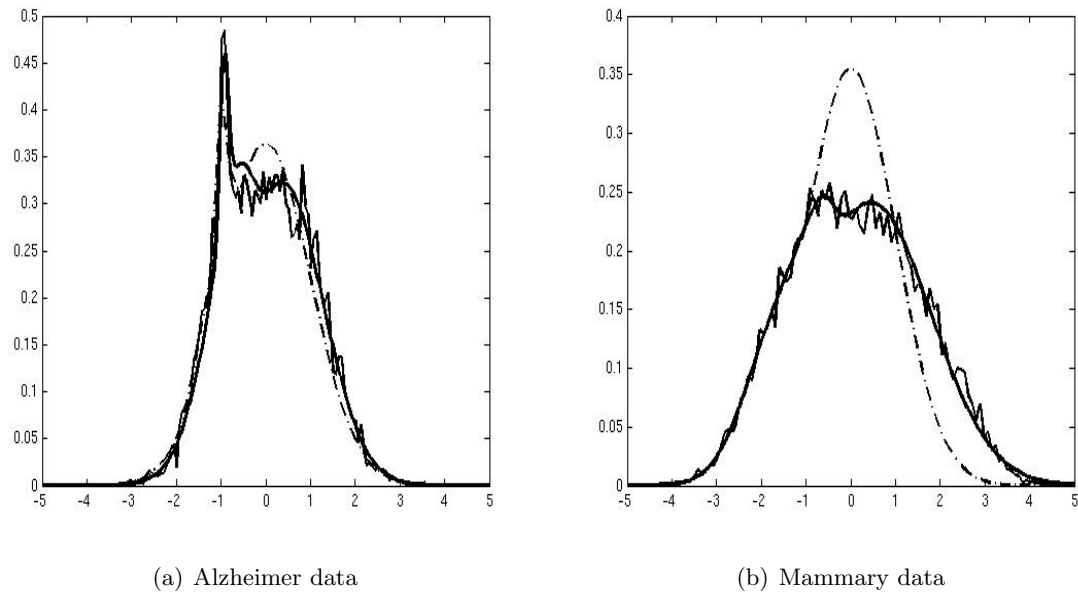


Figure 2: Constrained fit: dashed line; unconstrained fit: bold line; probability histogram polygon: jagged line.

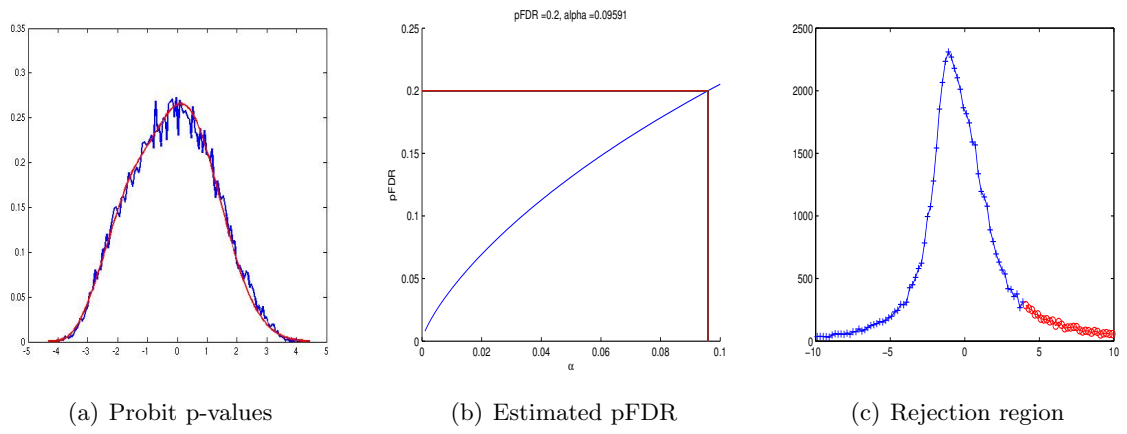


Figure 3: Exercise data; fitted distribution: bold line; probability histogram polygon: jagged line.