A SEMIPARAMETRIC MIXTURE METHOD FOR LOCAL FALSE DISCOVERY RATE ESTIMATION FROM MULTIPLE STUDIES

BY SEOK-OH JEONG* AND DONGSEOK CHOI† AND WONCHEOL JANG‡

Hankuk University of Foreign Studies*, Oregon Health & Science University/Kyung Hee University†, and Seoul National University‡

Anti-neutrophil cytoplasmic antibody associated vasculitis (AAV) is extremely heterogeneous in clinical presentation and involves multiple organ systems. While the clinical presentation of AAV is diverse, we hypothesized that all AAV share common pathways and tested the hypothesis based on 3 different microarray studies of peripheral leukocytes, sinus, and orbital inflammation disease. For the hypothesis testing, we developed a two-component semiparametric mixture model to estimate the local false discovery rates from the p-values of 3 studies. The two pillars of the proposed approach are Efron’s empirical null principle and log-concave density estimation for the alternative distribution. Our method outperforms other existing methods, in particular when the proportion of null is not that high. It is robust against the misspecification of alternative distribution. A unique feature of our method is that it can be extended to compute the local false discovery rates by combining multiple lists of p-values.

1. Introduction. Anti-neutrophil cytoplasmic antibody associated vasculitis (AAV) is extremely heterogeneous in clinical presentation and involves multiple organ systems including ranges from life threatening pulmonary hemorrhage to limited diseases of skin, nerves, orbit, or eye (Kallenberg, 2014; Macfarlane et al., 1983). While the clinical presentation of AAV is diverse, all may share common pathways. We wanted to test whether we could identify common pathways of AAV from 3 different studies. To test the hypothesis, gene expression data were collected from 3 published studies; a study of peripheral leukocytes (Alcorta et al., 2007), sinus brushings (Grayson et al., 2015), and orbital inflammatory disease (Rosenbaum et al., 2015). While all studies used microarray technology, they used Affymetix HU133 A and B, Affymetrix Human Gene 1.0 ST, and Affymetrix Human Genome U133 Plus 2.0, respectively. Therefore, it was not straightforward to analyze the raw expression data of 3 studies together due to the different

Keywords and phrases: false discovery rate, log-concave, microarray, mixture model, next generation sequencing data
number of probes of each array, different comparisons and control groups etc. From each study, the associated pathways were identified based on the set of differentially upregulated genes. The aim was to identify common pathways with the combined local false discovery rate (md-fdr), which will be defined in Section 2.3, less than 0.1 as shown in Figure 1.

Since the early 21st century, microarrays technology and next generation sequencing technology have revolutionized genomic researches by enabling simultaneous interrogating of tens of thousands genes. A quick search on PubMed.gov with a keyword microarray returns more than 80,000 items since 2000 (NCBI Resource Coordinators et al., 2017). Many authors have deposited their microarrays data to the gene expression omnibus (GEO) database and anyone can easily download gene expression data from multiple studies for a secondary data analysis (Edgar et al., 2002). While one can perform a meta-analysis of raw data from multiple studies, there will be several major challenges in combining data from different platforms/technology such as normalizing or adjusting batch effects. In addition, the full scale analysis can require considerable resources and time. If one only wishes to do a quick screening analysis from the multiple lists of significantly differentially expressed genes or pathways, a method to combine the multiple lists of p-values into one can be extremely handy. Figure 1 shows the probit-transformed p-values of common pathways identified from the 3 studies in our motivated application and the black points represent the combined local false discovery rates less than 0.1. In this paper, we propose a semiparametric mixture model that provides a unified way to compute the local false discovery rate for both single list of p-values and multiple lists of p-values.

Suppose that we observe $N$ cases, each with its own p-value, $p_i$, for $i = 1, 2, \ldots, N$. Let $z_i$, $i = 1, 2, \ldots, N$, be the probit transformed p-values. Each case can be considered as being from either null or alternative with the prior probability $p_0 = \Pr\{\text{null}\}$ or $p_1 = \Pr\{\text{alternative}\} = 1 - p_0$. Hence, $z$-values have the following mixture density

\begin{equation}
    f(z) = p_0f_0 + (1 - p_0)f_1(z),
\end{equation}

where $f_0(z)$ and $f_1(z)$ are null and alternative densities.

The above mixture model appears in three contexts: (i) in multiple testing problems (fMRI, microarray), probit-transformed p-values under $H_0$ follow the standard normal distribution while the marginal probability density function of the probit-transformed p-values associated with $H_1$ is unknown (Efron, 2008a); (ii) in variable/basis selection, a mixture prior is used to achieve sparsity (Johnstone and Silverman, 2005); (iii) in prediction, Fisher’s discriminant function can be regularized using a connection.
Fig 1: Probit-transformed $p$-values of common pathways from 3 AAV microarray studies. Black points represents $md$-$fdr < 0.1$. The red circled points are osteoclast differentiation(K) and cell surface interactions at the vascular wall(R). For more details, see Section 4.
with the local false discovery rate theory (Efron, 2008b).

In multiple testing, adjusting for multiplicity is of great interest. To do so, we may consider controlling for either the local false discovery rate, \( \text{fdr}(z) = \Pr(\text{null} \mid Z = z) \) or the False Discovery Rate, \( \text{FDR}(z) = \Pr(\text{null} \mid Z > z) \). In this paper, we focus on estimating \( \text{fdr}(z) \) that can be viewed as the posterior probability of a case being from the null given \( z \). Under the mixture model (1.1), it is straightforward to show

\[
\text{fdr}(z) = \frac{p_0 f_0(z)}{f(z)}.
\]

It is natural to assume that \( f_0 \) follows the standard normal distribution. However, Efron (2008a) suggested that theoretical null distribution \( N(0, 1) \) may not be suitable for \( f_0 \) and proposed to estimate \( f_0 \) with the empirical null distribution \( N(\mu, \sigma^2) \) where \( \mu \) and \( \sigma^2 \) are to be estimated from data. He used the zero assumption to estimate \( p_0 \) and \( f_0 \), but estimate the marginal density \( f \) separately with Lindsey’s method. As a result, his estimates may not follow the original mixture structure since the estimates are given from separate procedures. In other words, the mixture model only used to define \( \text{fdr} \), but not for making inference of \( \text{fdr} \). This can sometimes lead to problematic \( \text{fdr} \) estimates as shown in our simulations in Section 3.

In this paper, we propose a semiparametric mixture model that estimates \( p_0, f_0 \) and \( f_1 \) simultaneously with alternative being a log-concave density. The main advantages of this approach are three-fold. First, it is reasonable to assume the alternative distribution belongs to a log-concave family as long as probit transformed p-values are considered. We will discuss the robustness of this assumption in detail. Second, log-concave densities can be estimated nonparametrically without a smoothing parameter. Finally, our method can compute the local false discovery rate from multiple lists of p-values while Efron’s method cannot be extended similarly. For example, we observe 3-tuple of p-values \((p_{1i}, p_{2i}, p_{3i})\) for \( i = 1, \ldots, N \) in the 3 AAV microarray studies. Our goal is to estimate

\[
\text{fdr}(z_i) = \Pr(\text{null} \mid Z = z_i),
\]

where \( z_i = (z_{1i}, z_{2i}, z_{3i}) = (\Phi^{-1}(1 - p_{1i}), \Phi^{-1}(1 - p_{2i}), \Phi^{-1}(1 - p_{3i})) \), and \( \Phi \) is the cumulative distribution function of \( N(0, 1) \) so that \( \Phi^{-1} \) defines the probit transform.

Section 2 introduces our semiparametric mixture model. In Section 3, we present numerical studies to show the robustness and performance of our method. Section 4 presents the analysis results of 3 AAV microarray studies. Section 4 shows the robustness of our method. Section 5 concludes this paper.
2. Semiparametric Mixture Model.

2.1. The proposed model. In this section, we propose a semiparametric mixture model for $f$.

\begin{equation}
  f(z) = p_0 \phi_{\mu, \sigma^2}(z) + (1 - p_0)f_1(z),
\end{equation}

where $\phi_{\mu, \sigma^2}$ denotes the density of $N(\mu, \sigma^2)$ and $f_1$ is a log-concave density function for the alternative distribution.

Let $h(t)$ denote the pdf of the alternative distribution of $p$-values. Define $Z = \Phi^{-1}(1 - p)$ to be a probit transformed $p$-value. We usually assume that the $p$-values follow a uniform $[0, 1]$ under $H_0$. For the behavior of $p$-value under $H_1$, see Selke et al. (2001) and Hung et al. (1997). While we do not assume a specific class of family for alternative distribution of $p$-values, we claim that it is reasonable to assume the alternative distribution of $Z$, probit transformed $p$-value belongs to a log-concave distribution family. We will discuss the robustness of our assumption later in Section 3.

Statistical properties of the class of log-concave densities are well studied in Walther (2002, 2009). Walther (2002) also showed the existence of the nonparametric MLE of a univariate log-concave density that can be computed via an efficient algorithm such as an active set algorithm and an iterative convex minorant algorithm (Dümbgen and Rufibach, 2011).

In general, semiparametric mixture models are not identifiable without additional assumptions on the alternative density. Assuming that the alternative distribution belongs to the location-shift family, Bordes et al. (2006) showed the identifiability of the semiparametric mixture model under mild regularity conditions. Genovese and Wasserman (2004) also addressed identifiability of mixture models for $p$-value distribution under the assumption that the alternative is pure, that is, $\text{ess inf}_t h(t) = 0$. The recent work of Hunter et al. (2007) and Balabdaoui and Doss (2018) address the identifiability issues in log-concave mixture models. Note that Balabdaoui and Doss (2018) have an extra assumption about symmetry, but our method assumes the null distribution is a normal.

Our semiparametric mixture model in (2.1) may also suffer from the non-identifiability issue. In order to avoid this issue, we assume that the support of alternative distribution for $Z$ is given by $(a, \infty)$ for some $a$. Note that $(p_0, \mu, \sigma^2)$ can be determined by at least 3 distinct points arbitrarily chosen from the interval $(-\infty, a]$. In other words, the null components are estimable using data points in the interval $(-\infty, a]$, since the data points are certainly from the null distribution. Hence, the identifiability of our model is guaranteed as long as we can choose $a$ such that there exist at least 3 points in
This argument is similar to Efron’s with zero assumption for the support of $f_1$.

Recently Hu et al. (2016) proposed the MLE for log-concave mixture models and showed the existence and consistency of the MLE under fairly general condition. Since a normal distribution is also log-concave, our semiparametric mixture model can be considered as a special case of a finite (two) mixture of log-concave distributions.

We closely follow Hu et al. (2016) to show the consistency of our estimator. The key idea is to prevent the likelihood from being unbounded by constraining the parameter space. Hence, it suffices to check whether our semiparametric mixture model satisfies $f \in \mathcal{F}_\eta$ for some $\eta \in (0, 1]$, where $\mathcal{F}_\eta$ is defined as follows:

$$
\mathcal{F}_\eta = \{ f : f(x) = p \exp(\varphi_1(x)) + (1 - p) \exp(\varphi_2(x)), p \in (0, 1), \varphi \in \Phi_\eta \},
$$

where

$$
\varphi = (\varphi_1(x), \varphi_2(x)) = \left(-\frac{(x - \mu)^2}{2\sigma^2} - \frac{1}{2} \log(2\pi \sigma^2), \log f_1(x)\right)
$$

and $\Phi_\eta$ is a constrained subspace

$$
\Phi_\eta = \{ \varphi = (\varphi_1, \varphi_2) | \varphi_j \text{ is concave}, |S(\varphi)| \geq \eta > 0 \}.
$$

with

$$
S(\varphi) = \frac{\min\{\max_x \varphi_1(x), \max_x \varphi_2(x)\}}{\max\{\max_x \varphi_1(x), \max_x \varphi_2(x)\}} = \frac{\min\{-\frac{1}{2} \log(2\pi \sigma^2), \max_x \log f_1(x)\}}{\max\{-\frac{1}{2} \log(2\pi \sigma^2), \max_x \log f_1(x)\}}.
$$

The condition

$$
|S(\varphi)| \geq \eta > 0
$$

is satisfied as long as the mode of alternative density does not increase to $\infty$, and it is not so much a restrictive one for any alternative density. Therefore, the log-likelihood for our method is bounded and the estimated mixture density is consistent by Theorems 1 and 2 in Hu et al. (2016). Furthermore, the two estimated components and the null probability are also consistent since our model is identifiable.
2.2. *EM Algorithm.* It is natural to consider an EM-type algorithm to fit the proposed semiparametric model. To do so, we first define the likelihood function:

\[
\ell(\Delta, \mu, \sigma^2, f_1 | z_1, \ldots, z_N) = \sum_{i=1}^{N} \{ \Delta_i \log \phi_{\mu,\sigma^2}(z_i) + (1 - \Delta_i) \log f_1(z_i) \} + \sum_{i=1}^{N} \{ \Delta_i \log p_0 + (1 - \Delta_i) \log (1 - p_0) \}
\]

where \( \Delta = (\Delta_1, \ldots, \Delta_N)^T \) is a latent variable vector indicating the group membership.

Following Chang and Walther (2007), we first run the EM algorithm for a Gaussian mixture to get initial estimates of \((p_0, \mu, \sigma^2)\). Then the EM algorithm for fitting the proposed semiparametric model is given below:

1. Initialization: Set \( k = 0 \) and run the EM algorithm for a Gaussian mixture to get initial values \((p_0^{(0)}, \mu^{(0)}, \sigma^2^{(0)}), f_1^{(0)}(z_i) \) for \( i = 1, \ldots, N \). Put \( \gamma_i^{(0)} = 0, i = 1, \ldots, N \).

2. E-step: Compute the posterior probability: for \( i = 1, \ldots, N \),
   \[
   \gamma_i^{(k+1)} = \mathbb{E}(\Delta_i | \mu^{(k)}, \sigma^{2(k)}, z_1, \ldots, z_N) = \frac{p_0^{(k)} \phi_{\mu^{(k)},\sigma^{2(k)}}(z_i)}{p_0^{(k)} \phi_{\mu^{(k)},\sigma^{2(k)}}(z_i) + (1 - p_0^{(k)}) f_1^{(k)}(z_i)}.
   \]

3. M-step: Compute the log-concave estimates \( f_1^{(k+1)}(z_i) \) based on \( z_i \) with weights \( 1 - \gamma_i^{(k+1)} \) for \( i = 1, \ldots, N \). And put
   \[
   \mu^{(k+1)} = \frac{\sum_{i=1}^{N} \gamma_i^{(k+1)} z_i}{\sum_{i=1}^{N} \gamma_i^{(k+1)}},
   \]
   \[
   \sigma^{2(k+1)} = \frac{\sum_{i=1}^{N} \gamma_i^{(k+1)} (z_i - \mu^{(k+1)})^2}{\sum_{i=1}^{N} \gamma_i^{(k+1)}},
   \]
   \[
   p_0^{(k+1)} = \frac{1}{N} \sum_{i=1}^{N} \gamma_i^{(k+1)}.
   \]

4. Replication: If \( \max_{i=1,\ldots,N} |\gamma_i^{(k+1)} - \gamma_i^{(k)}| < \text{TOL} \), then output \( p_0^{(k+1)}, \mu^{(k+1)}, \sigma^{2(k+1)}, f_1^{(k+1)}(z_i) \)'s, \( \gamma_i^{(k+1)} \)'s and STOP. Otherwise, set \( k = k + 1 \) and goto the E-step.

In practice, we used \( \text{TOL} = 5 \times 10^{-6} \).
2.3. Extension to Multiple Studies. Suppose that one forms multivariate statistics $\mathbf{Z} = (Z_1, \ldots, Z_d)$ where each component is collected from a different study and provides unique information for common scientific hypotheses. For example, from each study of 3 AAV studies, the lists of test statistics and corresponding $p$-values are available. To the best of our knowledge, there is no method available to combine such information. It is straightforward to extend the concept of fdr to such multivariate cases using the proposed method:

$$\text{fdr}(\mathbf{Z}) = p_0 \frac{f_0(\mathbf{Z})}{f(\mathbf{Z})}$$

which we call hereafter as md-fdr. However, applying Efron’s method to multivariate is not straightforward. For example, it is not easy to extend Lindsey’s method to estimate the marginal distribution even for 2-dimension.

Up to our knowledge, Ploner et al. (2006) was the first attempt of md-fdr modeling. They proposed a method to combining a common test statistic, say $t$-statistic, for assessing differential expression in microarray studies with its standard error information. They estimated $f_0$ with discrete smoothing of binomial data after binning the data. However, their method still requires a smoothing parameter and has an issue with boundary bias. Furthermore, their method is only applicable to combining $t$-test statistics with standard error estimates for each gene. On the other hand, our method is straightforward to extend to multivariate cases and is not involved in choosing a smoothing parameter.

Cule et al. (2010) extended Dümbgen and Rufibach (2009) to multivariate settings and implemented multivariate log-concave estimation in R package LogConcDEAD (Cule et al., 2009). We use LogConcDEAD to implement our semiparametric mixture model in both single study and multiple studies. The R function and examples are available on the GitHub. (https://github.com/seokohj/SPMix-LocalFDR)

3. Simulation Studies. In this section, we investigated the performance of our semiparametric approach with some simulation studies where we can compare the results with Efron’s method. We conducted $M = 500$ Monte Carlo experiments with $N = 1,000$ and considered performance measures used for classifiers in machine learning applications: false positive rate (FPR) and sensitivity with the threshold set to be $\text{fdr} \leq 0.2$, which is equivalent to the Bayes factor $f_1(z)/f_0(z) \geq 4p_0/(1-p_0)$. This is a very strict level compared with classical testing practice. Note that the multiple testing can be thought as an unsupervised learning where the nullity of each
observation is not known. Since the true nullity of each data point is known during the simulation studies, it is appropriate to use performance measures for classifiers with two classes to investigate the performance of a multiple testing procedure. Also, in order to evaluate the accuracy of the estimated \( fdr \), we computed the Root Mean Squared Error (RMSE) which is given by

\[
\text{RMSE} = \frac{1}{M} \sum_{r=1}^{M} \frac{\left( \frac{1}{N} \sum_{i: \text{fdr}(z_i^{(r)}) \leq 0.5} \left( \hat{\text{fdr}}(z_i^{(r)}) - \text{fdr}(z_i^{(r)}) \right) \right)^2}{\sum_{i=1}^{N} I(\text{fdr}(z_i^{(r)}) \leq 0.5)}.
\]

where \( z_1^{(r)}, z_2^{(r)}, \ldots, z_N^{(r)} \) is the \( r \)th random sample generated during the Monte Carlo simulation. Note that we only took account of \( z_i \)’s with \( \text{fdr}(z_i^{(r)}) \leq 0.5 \) in the definition of the RMSE, because the accuracy of \( \text{fdr} \) estimation is required mainly for the region where the \( \text{fdr}(z) \) is small. For example, the behavior of \( \text{fdr} \) estimates for \( z \)-values with \( \text{fdr}(z) \approx 0.2 \) is of concern as is in our study.

3.1. Simulations of single study. We considered 6 scenarios for simulations, see Table 1. For the first 3 scenarios, we assumed normal distributions for both null and alternative distributions with different \( p_0 \). It does not conform the assumption on the support of the alternative density when considering a normal distribution for the alternative component. However the normal distributions are extremely thin-tailed, we may practically reckon that they are supported on a bounded interval given by the range of \( \pm 3 \) standard deviation around its mean, we consider that the support assumption is valid. The next 3 scenarios are the same as the first 3 except using gamma distributions as an alternative. Gamma distributions are log-concave and are supported on \((0, \infty)\), so that there is no issue on the identifiability.

Figures 2 and 3 show that our proposed approach stably yields reasonable results for a wide range of \( p_0 \) while Efron’s method tends to collapse when \( p_0 \) is relatively small. During our simulation study, R package locfdr was used for implementing Efron’s method (Efron et al., 2015). Note that one of the key assumptions of Efron’s method is the null probability is large, say \( p_0 \geq 0.90 \),

Table 1

<table>
<thead>
<tr>
<th>Scenario</th>
<th>( p_0 )</th>
<th>( f_0 )</th>
<th>( f_1 )</th>
<th>Scenario</th>
<th>( p_0 )</th>
<th>( f_0 )</th>
<th>( f_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.95</td>
<td>( N(0, 1) )</td>
<td>( N(3.5, 0.5^2) )</td>
<td>4</td>
<td>0.95</td>
<td>( N(0, 1) )</td>
<td>( \text{gamma}(12, 0.25) )</td>
</tr>
<tr>
<td>2</td>
<td>0.90</td>
<td>( N(0, 1) )</td>
<td>( N(3.5, 0.5^2) )</td>
<td>5</td>
<td>0.90</td>
<td>( N(0, 1) )</td>
<td>( \text{gamma}(12, 0.25) )</td>
</tr>
<tr>
<td>3</td>
<td>0.80</td>
<td>( N(0, 1) )</td>
<td>( N(3.5, 0.5^2) )</td>
<td>6</td>
<td>0.80</td>
<td>( N(0, 1) )</td>
<td>( \text{gamma}(12, 0.25) )</td>
</tr>
</tbody>
</table>
see (6.11) in Efron (2010). In the case that \( p_0 \) is not large enough, Efron’s method for estimating \( p_0 \) and the null distribution tends to break down and give an unreasonable estimate for \( p_0 \), for example, an estimate greater than 1, by the naive (with no constraint on the range of \( p_0 \) estimate) default fit to \( \log f \) over the central portion of the \( z \)-values. Figure 2 (a)–(d) summarize the results from the simulation setting with \( p_0 = 0.95 \) while (e)–(h) are from the setting with \( p_0 = 0.90 \); and (i)–(l) are from the setting with \( p_0 = 0.80 \). In the settings with \( p_0 = 0.95 \) and \( p_0 = 0.90 \), both Efron’s and our method worked reasonably well. However, when \( p_0 = 0.80 \), Efron’s method tends to overestimate \( p_0 \) and RMSE becomes substantially large. Furthermore, their sensitivity is closer to 0 in most cases while our method still works well. Figure 3 shows similar patterns for gamma alternative distributions.

Our log-concavity assumption is indeed robust in testing problems. To check this, we simulated \( p \)-values from Beta(0.3, 1) following Selke et al. (2001). Note that this distribution does not satisfy log-concavity. But, we verified that the distribution of probit transformed \( p \)-values gets quite close to a log-concave one. Figure 4 shows comparisons between the empirical distributions of the simulated \( p \)-values \( p_i \)’s from Beta(0.3, 1) and that of random numbers generated from the log-concave fit of \( p_i \)’s. Clearly, Figure 4(a) presents a distributional discrepancy between the simulated \( p \)-values and a random sample generated from the fitted log-concave density, and the Kolmogorov-Smirnov (KS) test confirmed the discrepancy with \( p \)-value \(< 10^{-4} \). However, after probit-transformation, one can barely see the difference between the two empirical distributions in Figure 4(b) with KS statistics \( p \)-value = 0.8693. We repeated the above procedure 500 times and Figure 5 summarizes the results. While the most of the KS test \( p \)-values before the probit transformation are small, after the probit transformation only 0.2\% of the KS test have \( p \)-values less than or equal to 0.05.

We conducted the similar numerical studies using noncentral \( t \), noncentral \( \chi^2 \) and noncentral \( F \) as alternative distribution of test statistics. We found the results were similar to the beta distribution case, see the Supplemental materials for the details.

It is possible that the distribution of probit-transformed \( p \)-values from the alternative component may not be unimodal. Suppose the alternative is a two-component mixture distribution normal distributions given by

\[
f_1 = 0.6 \cdot N(2.5, 0.25^2) + 0.4 \cdot N(3.5, 0.5^2),
\]

which clearly does not belong to a log-concave family. Figure 6 demonstrates that our method performed reasonably well with a two-component alternative density in a simulation. Our method overestimates the alternative
Fig 2: Comparison of the performances of the proposed method and Efron’s method for Scenarios 1–3 with the normal alternative distribution in Table 1. Boxplots are drawn with the results from $M = 500$ Monte Carlo experiments. Panels (a)–(d) summarize the results from the simulation setting with $p_0 = 0.95$; (e)–(h) are from the setting with $p_0 = 0.90$; and (i)–(l) are from the setting with $p_0 = 0.80$. 
Fig 3: Comparison of the performances of the proposed method and Efron’s method for Scenarios 4–6 with gamma alternative distribution in Table 1. The boxplots are drawn with the results from $M = 500$ Monte Carlo experiments. Panels (a)–(d) summarize the results from the simulation setting with $p_0 = 0.95$; (e)–(h) are from the setting with $p_0 = 0.90$; and (i)–(l) are from the setting with $p_0 = 0.80$. 
density (underestimating \( fdr \)) and provides deflated \( fdr \) estimates around the valley between the two components in the alternative distribution. However, this area should be declared as alternative, so underestimating \( fdr \) should not be an issue. Indeed, the monotonicity of \( fdr \) estimator can be considered as a desirable property in a two component model from the classification point of view and using the log-concave estimators enforces the monotonicity into \( fdr \) estimator. See Figure 6.

The further simulation results presented in Figure 7 suggest that the proposed method is robust even if the alternative distribution violates the log-concavity assumption. It still outperforms Efron’s again, although his method does not need assumptions on the shape of alternative distribution. As long as \( p_0 \) is moderately large, the potential risk of mis-specification for alternative distribution is likely to be forgiven, since it affects little to the accuracy of \( f \) estimates and, consequently, to that of \( fdr \) estimates.

In general, we are more sensitive to \( fdr \) estimation results around the decision boundary. In other words, we are more interested in whether our estimate for alternative fits reasonably well around the decision boundary area. We want to put an emphasis on that the key is to estimate the \( fdr \) around the boundary between null and alternative distributions and our \( fdr \) estimator around this area is robust even when log-concavity assumption is violated. The risk of mis-specification for the alternative distribution can be forgiven, since it affects little to the \( fdr \) estimates as long as \( p_0 \) is close to 1 in testing problems. Note that the proposed method captures and utilizes the “local” feature of data distribution only because it is affected little by the global shape of data distribution. Indeed the \( fdr \) estimate is not required to fit well all over the support of the marginal distribution and has only to be accurate on a neighborhood near the threshold.

3.2. Simulations of multiple studies. The simulation scenarios are similar to those in the previous subsection. In each scenario, the distributions of null and alternative are the same as the null and alternative distributions of the corresponding scenario in the single study setup. For the dependency of \( z \)'s, we postulated two copula structures: the Gaussian copula for simulations with normal marginal distributions and Frank copula for simulations with gamma marginals, see Yan (2007) for a detailed explanation on implementing copula models with R. By doing so, we want to find whether there is an advantage of \( md-fdr \) compared to \( fdr \) from a single study (1d-fdr).

Figure 8 illustrates the empirical distributions of sensitivities from 500 Monte Carlo experiments for each scenario. All the figures commonly show that \( md-fdr \) outperforms 1d-fdr. Comparing the boxplots for \( md-fdr \) with
Fig 4: A simulation result for Beta(0.3, 1) which violates the log-concavity. Panel (a) is for comparing the empirical distribution of 50 simulated $p_i$ from Beta(0.3, 1) and that of $\tilde{p}_i$, $i = 1, 2, ..., 50$ randomly drawn from the log-concave fit of $p_i$'s. The Kolmogorov-Smirnov test statistic for comparing these two empirical distributions was 0.46 with p-value < $10^{-4}$. Panel (b) is for comparing the empirical distribution of the probit transformed $p$-values $z_i = \Phi^{-1}(1 - p_i)$ and that of $\tilde{z}_i$'s drawn from the log-concave fit of $z_i$'s. The KS statistic was 0.12 with p-value = 0.8693.
Fig 5: From 500 Monte Carlo experiments with the same setup as in Figure 1, 500 p-values of Kolmogorov-Smirnov test were computed for $p_i$’s and $z_i$’s. Panel (a) is the histogram for the p-values of Kolmogorov-Smirnov test which compares the empirical distribution of $p_i$’s and that of $\tilde{p}_i$’s randomly drawn from the log-concave fit of $p_i$’s. Among 500 p-values for KS test, 57.2% were less than or equal to 0.05. Panel (b) is the histogram for the p-values of KS test for $z_i = \Phi^{-1}(1 - p_i)$’s and $\tilde{z}_i$’s drawn from the log-concave fit of $z_i$’s. Only 0.2% (1 case out of 500 experiments) of p-values were less than or equal to 0.05.
Fig 6: A simulation result with a bimodal alternative distribution: $f_1 = 0.6 \cdot N(2.5, 0.25^2) + 0.4 \cdot N(3.5, 0.5^2)$. The dotted lines in the upper panel is the true density functions used for simulating data. The black solid curve in the lower panel is the true fdr and the gray curve is the fdr estimate.
Fig 7: Comparison of the performances of the proposed method and Efron’s method for the scenarios with a bimodal alternative distribution which violates the log-concavity assumption for alternative distributions. Boxplots are drawn with the results from $M = 500$ Monte Carlo experiments. Panels (a)–(d) summarize the results from the simulation setting with $p_0 = 0.95$; (e)–(h) are from the setting with $p_0 = 0.90$; and (i)–(l) are from the setting with $p_0 = 0.80$. 
those from separate 1d-fdr, we confirm that the empirical distributions of sensitivities with md-fdr are much preferable to those in the corresponding 1d-fdr. Combining information with md-fdr could be better since it endows the test procedure with an extra flexibility in deciding the boundary of the rejection region. This is also explained by a similar phenomenon to the classification problem where the error rate may decrease when more variables are used.

Fig 8: Simulation results from bivariate (two studies) setups. Boxplots are drawn with the results from $M = 500$ Monte Carlo experiments. Panels (a)–(c) summarize the simulation results from the setting with the normal + normal mixture marginals coupled with elliptical copula; (d)–(f) are from the setting with the normal + gamma mixture coupled with Frank copula. In each panel, first two boxplots are drawn with the results from fitting 1-dimensional models separately for two components, and the last boxplot is from fitting bivariate model.

4. Common pathways across multiple studies. We wanted to test the hypothesis that common pathways exist in various AAV. Three published studies provided the list of significantly upregulated genes in 1) peripheral leukocytes (Alcorta et al., 2007), 2) sinus brushings (Grayson et al., 2015) and 3) orbital inflammatory disease (Rosenbaum et al., 2015) compared to
healthy controls. Reactome pathway database was used to get the relevant pathways for upregulated genes from each study (Fabregat et al., 2016) with their associated p-values. By applying our model to each list of p-values separately, there were 26, 32, and 26 pathways with 1d-fdr < 0.1 for study 1, 2) and 3), respectively. When we fitted our model to estimate md-fdr of 3 lists of p-values, there were 23 pathways with md-fdr < 0.1. Figure 1 shows the probit-transformed p-values of common pathways from the 3 studies, where black points represent md-fdr < 0.1. Table 2 presents the pathways with md-fdr < 0.1.

Many of the pathways in Table 2 support current knowledge and theories about AAV. For example, neutrophil degranulation pathway was the most significant pathway with 1d-fdr of each study and md-fdr. The involvement of neutrophil in these diseases was previously reported (Soderberg and Segelmark, 2016). Toll-Like receptors cascades pathway had also small 1d-fdr in all three studies as well as md-fdr. This pathway supports the hypothesis that the pathogenesis of these diseases involves infections and innate immunity. In addition, our methods could identify more pathways with md-fdr < 0.1, but not necessarily all the 1d-fdr < 0.1. For examples, osteoclast differential pathway had the 1d-fdr < 0.1 in the sinus brushings study only, while cell surface interactions at the vascular wall pathway had the 1d-fdr < 0.1 in the sinus brushings and the orbital tissue studies. The former is known to be important in the pathogenesis of AAV and the latter is highly relevant in vasculitis. Two large points in Figure 1 represent these two pathways. Three dimensional movie files of Figure 1 are available on the GitHub (https://github.com/seokohj/SPMix-LocalFDR).

Finally, potential novel common pathways were identified such as cell surface interactions at the vascular wall, amb2 integrin signaling, platelet pathways, etc., which should be investigated in future studies. The more comprehensive clinical interpretation can be found elsewhere (Friedman et al., 2019).

5. Discussion. In this paper we were able to combine 3 lists of p-values of pathways from 3 different studies about AAVs by using the proposed semiparametric mixture model of normal and log concave densities. Our method identified more pathways with md-fdr < 0.1, but not necessarily all the 1d-fdr < 0.1. For example, the osteoclast differential pathway had the 1d-fdr < 0.1 in only one study, but the md-fdr is significantly small. The pathogenesis of AAV agrees with our findings. To the best of our knowledge, there is no existing method that can compute combined FDR or fdr for multiple lists of p-values from several microarray studies.
Table 2
Raw p values and estimated fdrs for pathways with the md-fdr < 0.10

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Peripheral Leukocytes</th>
<th>Sinus brushings</th>
<th>Orbital tissue</th>
<th>md-fdr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutrophil degranulation</td>
<td>3.14 · 10^{-14}</td>
<td>1.11 · 10^{-10}</td>
<td>3.28 · 10^{-9}</td>
<td>1.05 · 10^{-12}</td>
</tr>
<tr>
<td>Osteoclast differentiation</td>
<td>1.26 · 10^{-2}</td>
<td>6.19 · 10^{-12}</td>
<td>3.91 · 10^{-7}</td>
<td>3.75 · 10^{-9}</td>
</tr>
<tr>
<td>Cell surface interactions at the vascular wall</td>
<td>8.87 · 10^{-1}</td>
<td>4.49 · 10^{-4}</td>
<td>5.86 · 10^{-6}</td>
<td>4.19 · 10^{-3}</td>
</tr>
<tr>
<td>Signaling by Interleukins</td>
<td>8.91 · 10^{-5}</td>
<td>8.24 · 10^{-8}</td>
<td>1.88 · 10^{-6}</td>
<td>6.13 · 10^{-4}</td>
</tr>
<tr>
<td>Phagosome</td>
<td>4.77 · 10^{-4}</td>
<td>2.88 · 10^{-8}</td>
<td>2.13 · 10^{-7}</td>
<td>3.17 · 10^{-5}</td>
</tr>
<tr>
<td>Leishmaniasis</td>
<td>4.12 · 10^{-2}</td>
<td>1.22 · 10^{-9}</td>
<td>1.72 · 10^{-7}</td>
<td>3.54 · 10^{-5}</td>
</tr>
<tr>
<td>Urokinase-type plasminogen activator and uPAR-mediated signaling</td>
<td>3.54 · 10^{-1}</td>
<td>9.81 · 10^{-8}</td>
<td>9.39 · 10^{-7}</td>
<td>7.77 · 10^{-4}</td>
</tr>
<tr>
<td>Antimicrobial peptides</td>
<td>7.69 · 10^{-6}</td>
<td>1.31 · 10^{-4}</td>
<td>1.14 · 10^{-4}</td>
<td>8.23 · 10^{-4}</td>
</tr>
<tr>
<td>Malaria</td>
<td>4.00 · 10^{-4}</td>
<td>3.77 · 10^{-6}</td>
<td>1.69 · 10^{-4}</td>
<td>1.16 · 10^{-5}</td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>3.96 · 10^{-2}</td>
<td>3.79 · 10^{-8}</td>
<td>2.52 · 10^{-7}</td>
<td>1.91 · 10^{-5}</td>
</tr>
<tr>
<td>IL4-mediated signaling events</td>
<td>5.47 · 10^{-4}</td>
<td>1.07 · 10^{-1}</td>
<td>5.04 · 10^{-4}</td>
<td>2.07 · 10^{-4}</td>
</tr>
<tr>
<td>Signaling by the B Cell Receptor</td>
<td>7.74 · 10^{-4}</td>
<td>8.67 · 10^{-4}</td>
<td>6.68 · 10^{-4}</td>
<td>2.18 · 10^{-4}</td>
</tr>
<tr>
<td>Extracellular matrix organization</td>
<td>4.96 · 10^{-1}</td>
<td>2.51 · 10^{-4}</td>
<td>1.27 · 10^{-4}</td>
<td>2.40 · 10^{-2}</td>
</tr>
<tr>
<td>Toll-Like Receptors Cascades</td>
<td>1.85 · 10^{-5}</td>
<td>2.78 · 10^{-4}</td>
<td>3.39 · 10^{-4}</td>
<td>2.55 · 10^{-2}</td>
</tr>
<tr>
<td>Measles</td>
<td>7.59 · 10^{-1}</td>
<td>1.23 · 10^{-2}</td>
<td>2.84 · 10^{-4}</td>
<td>3.48 · 10^{-2}</td>
</tr>
<tr>
<td>Cytokine-cytokine receptor interaction</td>
<td>6.66 · 10^{-3}</td>
<td>7.78 · 10^{-6}</td>
<td>4.08 · 10^{-4}</td>
<td>3.55 · 10^{-2}</td>
</tr>
<tr>
<td>Response to elevated platelet cytosolic Ca2+</td>
<td>6.70 · 10^{-1}</td>
<td>4.50 · 10^{-6}</td>
<td>9.52 · 10^{-4}</td>
<td>3.82 · 10^{-2}</td>
</tr>
<tr>
<td>Chemokine signaling pathway</td>
<td>3.08 · 10^{-1}</td>
<td>1.97 · 10^{-5}</td>
<td>1.27 · 10^{-2}</td>
<td>4.12 · 10^{-2}</td>
</tr>
<tr>
<td>Complement and coagulation cascades</td>
<td>1.99 · 10^{-4}</td>
<td>2.70 · 10^{-6}</td>
<td>4.78 · 10^{-4}</td>
<td>4.44 · 10^{-2}</td>
</tr>
<tr>
<td>GPVI-mediated activation cascade</td>
<td>4.00 · 10^{-1}</td>
<td>2.10 · 10^{-3}</td>
<td>1.69 · 10^{-3}</td>
<td>7.85 · 10^{-2}</td>
</tr>
<tr>
<td>ambl2 Integrin signaling</td>
<td>2.76 · 10^{-4}</td>
<td>1.90 · 10^{-6}</td>
<td>4.05 · 10^{-4}</td>
<td>8.57 · 10^{-2}</td>
</tr>
<tr>
<td>Platelet homeostasis</td>
<td>2.41 · 10^{-2}</td>
<td>6.56 · 10^{-4}</td>
<td>3.28 · 10^{-3}</td>
<td>9.14 · 10^{-2}</td>
</tr>
<tr>
<td>Inflammatory bowel disease</td>
<td>4.91 · 10^{-5}</td>
<td>3.10 · 10^{-2}</td>
<td>5.80 · 10^{-4}</td>
<td>9.30 · 10^{-2}</td>
</tr>
</tbody>
</table>
We presented an EM-type algorithm to implement the proposed estimators for both single high dimensional and multiple high dimensional testing results. Our method can estimate fdr and the proportion of the null simultaneously and fit the alternative when necessary. Our method is easy to use because it does not require smoothing parameter selection. Our simulation studies showed that the proposed method outperforms other existing method in both single study and multiple studies. We presented an application which demonstrates the unique feature of the proposed semiparametric mixture model and especially the advantage of using md-fdr over 1d-fdr.

Recently Wilson (2019) proposed the harmonic mean p-value (HMP) to combine dependent tests. HMP also has the Bayesian properties (Held, 2019). HMP appears to outperform other procedures that control the false discovery. It would be interesting if one extends the HMP to multiple studies with our method.

Acknowledgement. The authors thanks Professor Bodhisattva Sen and Dr. Rohit Kumar Patra for sharing the Carina data and their codes. We are also grateful to Dr. Marcia Friedman at Oregon Health & Science University for providing us the motivating data used in Section 4. We also thank the editor/associate editor and two referees for their constructive suggestions, which enabled us to improve this manuscript significantly.

Financial support. NIH Grants EY028897, EY020249 and EY010572. None of the funding organizations had a role in the design or conduct of this research. Woncheol Jang’s research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2014R1A4A1007895, No. 2017R1A2B2012816). Seok-Oh Jeong’s work was supported by the research fund of Hankuk University of Foreign Studies and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2006112).

References.


WONCHEOL JANG
DEPARTMENT OF STATISTICS
SEOUL NATIONAL UNIVERSITY
SEOUL, 08826, REPUBLIC OF KOREA
E-MAIL: wejang@snu.ac.kr