Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Two-Component Mixture Model in the Presence of Covariates

Nabarun Deb, Sujayam Saha, Adityanand Guntuboyina & Bodhisattva Sen

To cite this article: Nabarun Deb, Sujayam Saha, Adityanand Guntuboyina & Bodhisattva Sen (2022) Two-Component Mixture Model in the Presence of Covariates, Journal of the American Statistical Association, 117:540, 1820-1834, DOI: 10.1080/01621459.2021.1888739

To link to this article: https://doi.org/10.1080/01621459.2021.1888739

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 117, NO. 540, 1820–1834: Theory and Methods
https://doi.org/10.1080/01621459.2021.1888739

Two-Component Mixture Model in the Presence of Covariates

Nabarun Deb[a], Sujayam Saha[b], Adityanand Guntuboyina[c], and Bodhisattva Sen[a]

[a]Department of Statistics, Columbia University, New York, NY; [b]Google Inc., Mountain View, CA; [c]Department of Statistics, University of California at Berkeley, Berkeley, CA

ABSTRACT

In this article, we study a generalization of the two-groups model in the presence of covariates—a problem that has recently received much attention in the statistical literature due to its applicability in multiple hypotheses testing problems. The model we consider allows for infinite dimensional parameters and offers flexibility in modeling the dependence of the response on the covariates. We discuss the identifiability issues arising in this model and systematically study several estimation strategies. We propose a tuning parameter-free nonparametric maximum likelihood method, implementable via the expectation-maximization algorithm, to estimate the unknown parameters. Further, we derive the rate of convergence of the proposed estimators—in particular we show that the finite sample Hellinger risk for every 'approximate' nonparametric maximum likelihood estimator achieves a near-parametric rate (up to logarithmic multiplicative factors). In addition, we propose and theoretically study two 'marginal' methods that are more scalable and easily implementable. We demonstrate the efficacy of our procedures through extensive simulation studies and relevant data analyses—one arising from neuroscience and the other from astronomy. We also outline the application of our methods to multiple testing. The companion R package `NPMLEmix` implements all the procedures proposed in this article.

ARTICLE HISTORY
Received December 2018
Accepted February 2021

KEYWORDS
Expectation-maximization algorithm; Gaussian location mixture; Identifiability; Local false discovery rate; Nonparametric maximum likelihood; Two-groups model

1. Introduction

Consider independent and identically distributed (iid) observations $Y_1, \ldots, Y_n$ from the following two-component mixture model:

$$Y_i \sim \bar{\pi} F_1^* + (1 - \bar{\pi}) F_0, \quad \text{for } i = 1, \ldots, n, \tag{1}$$

where F_0 is assumed to be a completely *known* distribution function (DF) while F_1^*, along with $\bar{\pi}$, are the unknown quantities of interest. We will call F_0 the *noise* distribution, F_1^* the *signal* distribution and $\bar{\pi}$, the *signal* proportion. Such a model has received a lot of attention in the statistical literature, particularly in the context of multiple testing problems (microarray analysis, neuroimaging, etc.) where it is usually referred to as the *two-groups* model (see, e.g., Storey 2002, 2003; Efron 2008; Cai and Jin 2010; Efron 2010, chap. 2). In the multiple testing problem, the obtained p-values or z-values (Y_i's as per (1)), from the numerous (independent) hypotheses tests, have a Uniform$(0, 1)$ or $\mathcal{N}(0, 1)$ distribution (under the null hypothesis), which we call F_0, while their distribution (i.e., F_1^*) under the alternative is *unknown*; here $\bar{\pi}$ is the proportion of nonnull hypotheses. The two-groups model has also been used in contamination problems, where the (unknown) distribution F_1^* may be contaminated by the known distribution F_0, yielding a sample drawn from F as in (1) (see, e.g., Lemdani and Pons 1999; McLachlan and Peel 2000; Dai and Charnigo 2007; Walker et al. 2009).

However, quite often in real applications, additional information is available on each observation in the form of *covariates* which is ignored by the two-groups model. The following two examples describe two such applications and illustrate the need to model the observed covariates.

Example 1.1 (Neuroscience example). Scott et al. (2015) analyzed data arising from a multi-unit recording experiment consisting of measurements from 128 units (either neurons or multi-unit groups) from the primary visual cortex of a rhesus macaque monkey in response to visual stimuli (see Kelly et al. 2007 for details). The goal of the experiment was to detect fine-time-scale neural interactions ("synchrony"). The data consisted of thousands of test statistics Y_i's, each one arising from testing the null hypothesis of no interaction between a pair of units. Let F_0 be the null distribution of Y_i (assumed to be known) and F_1^* the unknown signal distribution. A natural approach for modeling the distribution of Y_i's is via the two-groups model (see, e.g., Scott et al. 2015). However, the dataset also included two interesting covariates: (a) physical distance between units, and (b) tuning curve correlation between units. Figure 1 illustrates the relationship between the observed test statistics and the two covariates. It clearly shows that the covariates are related to the Y_i's. However, as was also observed by Scott et al. (2015), the two-groups model (1) inappropriately ignores the known spatial and functional relationships among the neurons. This motivates the need to develop and study models that generalize (1) to include covariates. We discuss this data and its analysis in more detail in Section 7.1.

Example 1.2 (Astronomy example). Walker et al. (2009) analyzed data on individual stars obtained from nearby dwarf spheroidal

CONTACT Bodhisattva Sen bodhi@stat.columbia.edu Department of Statistics, Columbia University, New York, NY 10027.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

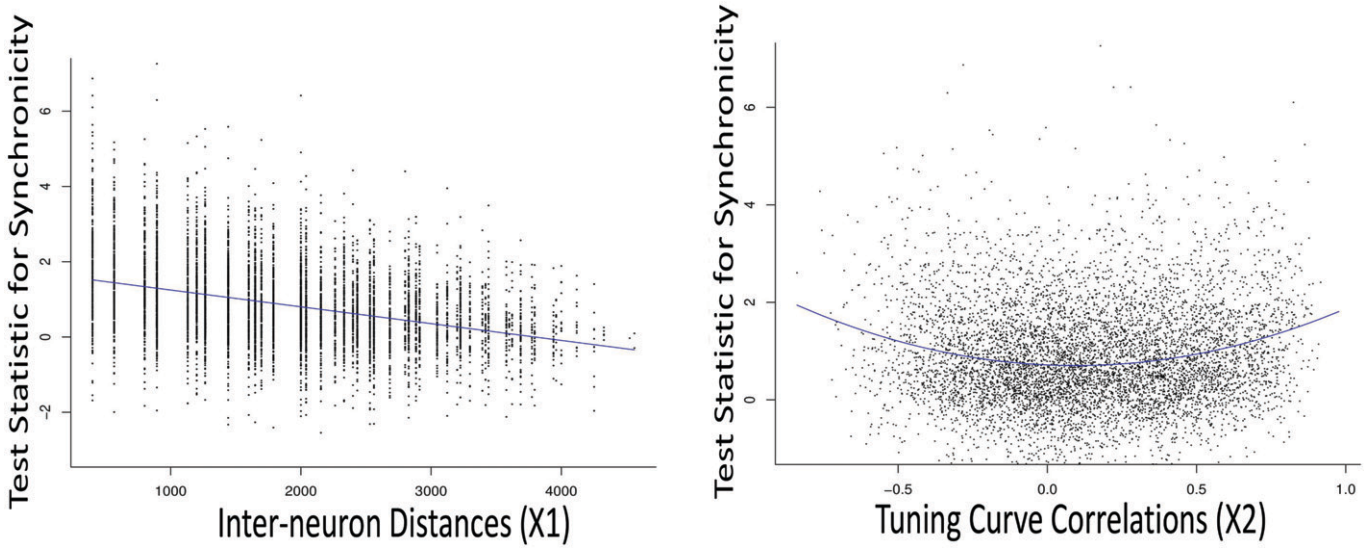

Figure 1. Scatterplots of test statistics computed on each pair of units (higher is more significant), plotted against covariates: distance between units (left), tuning curve correlation between units (right). The loess fit is overlaid upon each scatterplot, visually indicating that the test statistics are dependent upon covariate information.

(dSph) galaxies. The data contain measurements on line-of-sight velocity (denoted by Y), projected distance from the center of the dSph galaxy (denoted by X), and other variables (e.g., metallicity) for around 1000–2500 stars per dSph, including some fraction of contamination from foreground Milky Way stars (in the field of view of the dSph galaxy) (see, e.g., Walker, Mateo, and Olszewski 2009). The primary goal is to identify the dSph galaxy stars in the sample and recover their line-of-sight velocity distribution. Due to foreground contamination, Y is distributed marginally as in the two-component mixture model (1); see the right panel of Figure 2. Here we plot the estimated density (obtained using kernel density estimation) of the observed Y_i's (for the Carina dSph) along with (scaled) f_0—the density of F_0—which is known from the Besancon Milky Way model (see Robin et al. 2003). However, the left panel of Figure 2, which shows the scatterplot of X and Y, reveals that Y indeed depends on X which the *two-groups* model fails to capture. In this article, we develop a methodology that incorporates this covariate information to yield: (a) better estimation of F_1^*, the distribution of the line-of-sight velocity for stars in the dSph; and (b) more reliable "posterior" probability estimates of each star (in the sample) being a dSph member; see Appendix F in the supplementary materials for details.

Applications such as Examples 1.1 and 1.2 motivate the need to generalize (1) to incorporate covariate information; also see Schildknecht, Tabelow, and Dickhaus (2016) and Li and Barber (2017) for two more relevant applications in neural imaging and genetics data, respectively. Toward this direction, suppose that $(Y_1, X_1), \ldots, (Y_n, X_n)$ are iid having a distribution on $\mathbb{R} \times \mathbb{R}^p$ ($p \geq 1$). As studied in Scott et al. (2015) and Walker et al. (2009), a natural way to model the joint distribution of (Y, X) that generalizes (1) would be to consider

$$Y|X = x \sim \pi^*(x)F_1^* + (1 - \pi^*(x))F_0 \quad \text{and} \quad X \sim m, \qquad (2)$$

where:

1. m is a fixed probability measure supported on some space $\mathcal{X} \subseteq \mathbb{R}^p$.
2. The random variable Y takes values in a subset $\mathcal{Y}$ of $\mathbb{R}$ (e.g., $\mathcal{Y} = [0, 1]$ or $\mathcal{Y} = \mathbb{R}$) and $F_0 \neq F_1^*$ are two DFs on $\mathbb{R}$. We assume that F_0 is *known* (see Remark 1.2 for the case when F_0 is not completely specified) and F_1^* is *unknown* and belongs to a parametric or nonparametric class $\mathcal{F}$. Note that model (2) assumes that F_0 and F_1^* do not depend on the covariates.
3. $\pi^* : \mathcal{X} \to [0, 1]$ is an *unknown* function belonging to a parametric or nonparametric class of functions Π.

The crucial difference between models (2) and (1) is that (2) allows the prior probability of an observation coming from the signal distribution to depend on the covariates. In fact, model (2) is indeed a generalization of the two-groups model (which is obtained by taking $\pi^*(\cdot)$ to be the constant function). It is worth mentioning that (2) can be treated as a regression model with a special structure: Suppose that Z is the unobservable latent variable corresponding to Y that decides which of the two populations (F_0 or F_1^*) Y is drawn from; that is, $Y|Z = 0 \sim F_0$ and $Y|Z = 1 \sim F_1^*$. Then, under model (2), Y is *conditionally independent* of X given Z; of course, Y is dependent on X unconditionally. This observation can be interpreted in the following way: Model (2) implies that X provides some information about Y, but X does not provide any additional information about Y if we knew the value of Z.

To motivate model (2) further, we mention a few special cases of (2) that are of significant interest in the multiple testing problem. Let us start with two natural examples of $\mathcal{F}$ and F_0.

Decreasing densities: In this case, $\mathcal{F}$ denotes the class of all DFs having a nonincreasing density on $[0, 1]$ and F_0 is the uniform distribution on $[0, 1]$. This situation naturally arises in multiple testing problems where Y denotes the p-value corresponding to a hypothesis test and we assume that under H_0, the p-values have the uniform distribution on $[0, 1]$ (see, e.g., Genovese and Wasserman 2004; Efron 2010). Further, in this problem it is quite natural to assume that, under the alternative, the p-values will tend to be stochastically smaller (or they will have a nonincreasing density on $[0, 1]$) (see,

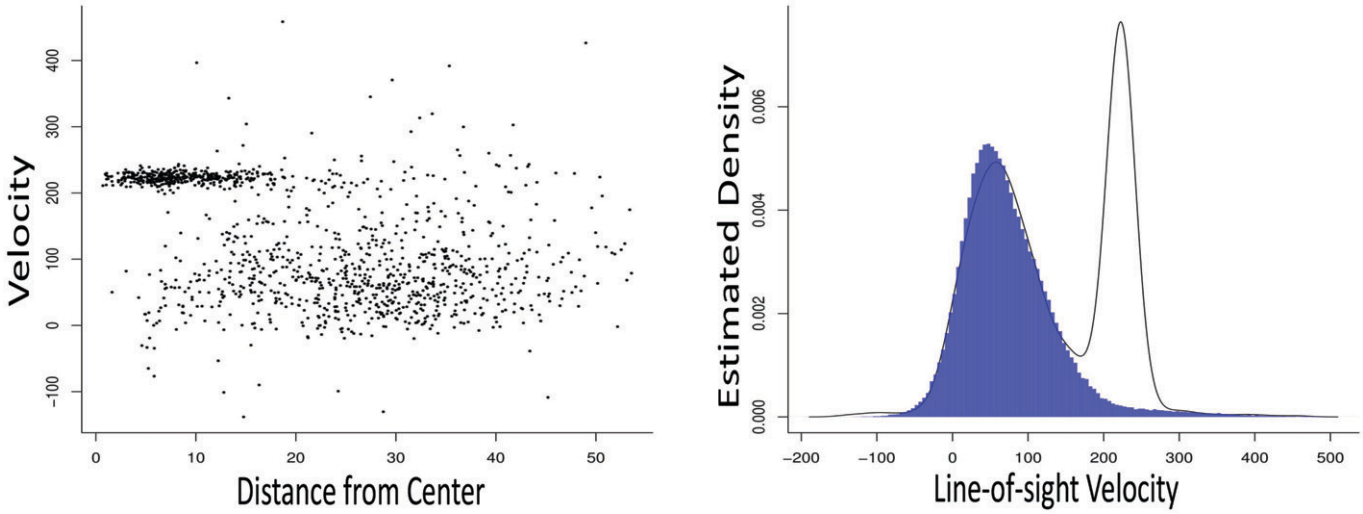

Figure 2. The left-hand figure shows the joint scatterplot of Y versus X. The right-hand plot shows the standard kernel density estimate of Y in this dataset with a scaled f_0 overlaid in blue. This indicates that the tight point cloud to the top left on left figure comprises mostly of stars from Carina galaxy, while the sparser point cloud to bottom center comprises mostly of stars from Milky Way. From the scatterplot, Carina stars are clearly more frequent at lower values of X (distance from center). Thus, a classification procedure which uses X should be more accurate.

e.g., Schweder and Spjøtvoll 1982; Langaas, Lindqvist, and Ferkingstad 2005). Let us denote the class of all distributions with nonincreasing densities on $[0, 1]$ by $\mathcal{F}_{\downarrow}$.

Gaussian location mixtures: In this case, $\mathcal{F} \equiv \mathcal{F}_{\text{Gauss}}$ denotes the class of all Gaussian location mixtures, that is, any $F_1^* \in \mathcal{F}_{\text{Gauss}}$ has the form $F_1^*(x) := \int \Phi(x-\theta)dG(\theta)$ for $x \in \mathbb{R}$ where G is some unknown probability measure on $\mathbb{R}$ and Φ is the standard normal DF. Moreover, we take $F_0 := \Phi$ (see, e.g., Cai and Jin 2010; Scott et al. 2015). In the above, G models the *effect size* distribution (see Appendix H.2 in the supplementary materials for the details) and naturally arises when dealing with z-scores (as opposed to p-values). Note that $\mathcal{F}_{\text{Gauss}}$ contains all finite Gaussian location mixtures (with unit variance).

Next we consider some natural models for the class Π.

Constant functions: Let us first consider the case when Π consists of all constant functions. This reduces model (2) to the well-known two-groups model (see (1)). We shall denote this class by $\Pi_{\equiv}$.

Nondecreasing functions: Assume $p = 1$ and $\mathcal{X}$ is a subinterval of $\mathbb{R}$. Quite often when testing a set of (ordered) hypotheses, the practitioner may have reason to believe that the test statistics earlier in the set are less likely to be signals (see, e.g., Li and Barber 2016, 2017). In such a situation, it is natural to consider Π to be the class of all nondecreasing functions on $\mathcal{X}$. We shall denote this class by $\Pi_{\uparrow}$.

Generalized linear model: In the absence of strong prior information on the class Π, a general modeling strategy would be to consider the following class of functions: $\pi(x) := g(\beta_0+\beta^\top x)$ as (β_0, β) varies over $\mathbb{R}\times\mathbb{R}^p$. Here $g : \mathbb{R} \to [0, 1]$ is a fixed and known link function. We shall denote this class of functions by Π_g. When $g(z) := (1+\exp(-z))^{-1}$ (logistic link), we shall denote Π_g by Π_{logit}. This is a special case of the model considered in Scott et al. (2015). When $g(z) := \Phi(z)$, we denote Π_g by Π_{probit}. We will study these classes in detail in this article.

1.1. Our Contributions

In this article, we propose and study likelihood based methods for estimating the functions $\pi^*(\cdot)$ and F_1^* (and its density f_1^*) as described in model (2). We conduct a systematic study of the statistical and computational properties of our proposed methods, which very naturally yield a multiple testing procedure; see Appendix E in the supplementary materials. We summarize our contributions below:

Identifiability: Model (2), as posited, need not be identifiable. In Section 2, we study identifiability of model (2) and give easily verifiable necessary and sufficient conditions in a rather general setting; see Lemma 2.1. In addition, we demonstrate how to use Lemma 2.1 to prove identifiability for a wide range of choices of Π and $\mathcal{F}$, including the ones used in Scott et al. (2015) (see Lemma 2.2). To the best of our knowledge, the issue of identifiability in (2) has not been properly addressed before. Note that the two-groups model, as posited in (1), is not identifiable (see, e.g., Genovese and Wasserman 2004; Patra and Sen 2016). However, it is interesting to note that the presence of covariates can make model (2) identifiable.

Joint maximum likelihood: In Sections 3 and 4, we develop a general (nonparametric) maximum likelihood based procedure to estimate π^* and f_1^* from iid observations drawn according to model (2). We propose iterative procedures based on the expectation-maximization (EM) algorithm (see Dempster, Laird, and Rubin 1977; Lange 2016) to compute the maximum likelihood estimates (MLEs). Our procedure can handle both parametric and nonparametric specifications for $\mathcal{F}$ and Π and, in particular, covers the important scenarios discussed above. The resulting estimates of π^* and f_1^* yield accurate estimators of the conditional density of Y given X. We show in Theorem 3.1 that when we maximize the likelihood over the nonparametric class of all Gaussian location mixtures ($\mathcal{F}_{\text{Gauss}}$), the resulting estimator of this conditional density has a *parametric* rate

of convergence, up to logarithmic factors (see Section 3.2). In fact, Theorem 3.1 holds for a much larger class of estimators (we call these *approximate MLEs*) which includes the MLE as a special case. This generalization is important for analyzing the statistical properties of our estimators as we are dealing with a nonconvex optimization problem where exact maximizers are computationally difficult to obtain. We also propose specialized algorithms for solving the M-step in the EM algorithm for estimating π^* and f_1^*, depending on the choices of Π and $\mathcal{F}$.

Marginal methods: We propose two other methods for estimating $\pi^*(\cdot)$ and f_1^* that are based on appropriately marginalizing the joint distribution of (X, Y); see Section 5 for the details. These marginal methods bypass the joint maximization of the likelihood (which is a nonconvex problem in general) and are easily implementable. These marginal methods can also be successfully used to properly initialize the EM algorithm to compute the joint MLE. We establish a finite sample risk bound of our estimator of f_1^* (see Theorem 5.1) and derive the asymptotic distribution of the coefficient vector for certain parametrically specified link functions $\pi^*(\cdot)$ (see Theorem 5.2).

Even though we can handle nonparametric classes $\mathcal{F}$ (and Π), both our proposed methods—namely, the joint maximization and marginal procedures—are tuning parameter-free, and are thus completely automated.

Simulations and real data examples: We conduct extensive simulation studies (see Section 6) that point to the superior performance of the proposed estimators, when compared to its competitors. A direct consequence of our proposed methodology is a comprehensive procedure that addresses the multiple testing problem (see Appendix E in the supplementary materials). We demonstrate the accuracy of the estimated local false discovery rate (lFDR) through extensive simulations. Further, we analyze the two real data examples introduced above (see Sections 7.1 and Appendix F). These illustrate the applicability of our methods. Both marginal methods and the joint maximum likelihood method have been implemented in the companion R package `NPMLEmix` (see Deb et al. 2020) which is available on CRAN; see *https://CRAN.R-project.org/package=NPMLEmix*. It also includes relevant codes for all our simulations and data analyses.

Before considering estimation in the framework of (2), as we have done above, it is natural to ask: "Do the covariates indeed have any effect in the multiple testing problem?" and "Does the signal distribution, that is, the distribution of the nonnull p-values/z-values, depend on the covariates?" In Appendices B and C (see the supplementary materials), we show that the above questions can be reformulated as hypothesis testing problems and we propose natural testing procedures to address them.

The accompanying supplementary materials contain proofs of our main results, detailed discussions on some of the algorithms we propose in the article and additional computational studies. Before ending this subsection, we would like to point out two important aspects of our proposals in this article, through the following two remarks.

Remark 1.1. In our problem setting (see (2)), the (oracle) optimal testing procedure should reject hypotheses with low lFDRs; see Basu et al. (2018) where the authors view the multiple testing problem from a decision theoretical perspective and prove such an optimality result. Thus, we focus on accurate estimation of lFDRs (and the associated model parameters). This is the crucial point of difference between our approach and some of the more recent papers in this area (see Barber and Candès 2015; Lei and Fithian 2016; Ignatiadis and Huber 2017). We believe that methods focusing primarily on finite sample FDR control can sometimes be quite conservative.

Remark 1.2 (Estimating an empirical null). In Efron (2004) and Scott and Berger (2006), the authors have observed that the theoretical null (e.g., specifying F_0 as the CDF of a standard normal or uniform distribution) poorly describes many datasets and consequently, an "empirical null" needs to be estimated. Such an example of an "empirical null" would be when F_0 is known to be the CDF of a $\mathcal{N}(\mu, \sigma^2)$ distribution with μ and σ^2 unknown. One of the strengths of our approach is that our proposed methodology can deal with the estimation of an empirical null by a simple preprocessing step (just like FDRreg in Scott and Berger (2006)). We provide a concrete example in our neural synchrony data analysis (see Section 7) where the maximum likelihood approach proposed in Efron (2004) has been used to estimate an empirical null (similar to Scott and Berger (2006)).

1.2. Literature Review

The two-groups mixture model (without covariates) has been studied and applied extensively (see, e.g., McLachlan, Bean, and Peel 2002; Storey 2002, 2003; Efron 2004, 2008, 2010, chap. 2; Genovese and Wasserman 2004; Johnstone and Silverman 2004; Meinshausen and Rice 2006; Scott and Berger 2006; Müller, Parmigiani, and Rice 2007; Robin et al. 2007; Walker et al. 2009; McLachlan and Wockner 2010; Patra and Sen 2016). However, in a variety of multiple testing applications, as in our motivating applications, there is often additional information available on the individual test statistics (e.g., p-values or z-scores)—for example, the p-values may be naturally ordered, grouped, contain inherent clusters, etc. A natural strategy to incorporate such auxiliary information is through the use of weights corresponding to p-values (see, e.g., Genovese, Roeder, and Wasserman 2006; Hu, Zhao, and Zhou 2010; Benjamini and Bogomolov 2014; Dobriban 2016). We believe that modeling the weights can itself be a difficult problem in the absence of strong prior information and there is no generally accepted strategy. These limitations have prompted some recent advances in this area which we discuss below.

Ignatiadis et al. (2016) proposed grouping the hypotheses and choosing weights for each group so as to maximize the number of rejections after a usual reweighing procedure. In Ignatiadis (2018), using a slightly modified censoring p-value based approach, the authors are able to guarantee finite sample FDR control. Such p-value masking techniques have also been used in Li and Barber (2016) and Lei and Fithian (2016). These articles actually consider a further generalization of model (2) where the distribution of nonnull p-values are allowed to

depend on the covariates. However, their proposed methods are geared toward guaranteeing finite sample FDR control whereas we take a more direct approach by proposing natural models (see, e.g., Π_g, $\Pi_\uparrow$, $\mathcal{F}_\downarrow$, $\mathcal{F}_{\text{Gauss}}$) for p-values or z-scores and focus on accurate estimation of the unknown quantities. This is particularly useful if the analyst is also interested in estimating the distribution of the nonnull p-values, for example, in the contamination problem mentioned in Example 1.2. Moreover, our approach avoids grouping hypotheses based on covariates (which may be difficult if the covariate space is complex) and does not need the choice of any tuning parameters.

The article (Scott et al. 2015) is perhaps the closest to our work. The authors use $\Pi = \Pi_{\text{logit}}$, $F_0 = \Phi$ and $\mathcal{F} = \mathcal{F}_{\text{Gauss}}$ and illustrate the superiority of such a model over the traditional two-groups model (1) in terms of signal detection through extensive simulations and by analyzing the neural synchrony data (Example 1.1). The big difference between our article and Scott et al. (2015) is that our main recommended procedure is based on (nonparametric) MLE while their recommended procedure (which they call FDRreg) is more like one of our marginal ones (see Appendix H.1 in the supplementary materials for the details). Note that Scott et al. (2015) also proposed a full Bayes procedure and an empirical Bayes procedure. We, however, resort to a frequentist approach and obtain estimators by maximizing the likelihood function. Moreover, Scott et al. (2015) did not provide any theoretical guarantees for their estimators (as we do in Theorem 3.1). In Section 6, we argue through extensive simulations that our method yields more accurate estimates of $\pi^*(\cdot), f_1^*$ and lFDRs (particularly when the "signal" varies significantly with the covariates).

2. Identifiability in Model (2)

Identifiability issues arise naturally in the study of mixture models (see, e.g., Teicher 1961; Titterington, Smith, and Makov 1985, sec. 3.1) and model (2) is no exception. We detail these issues in this section before proceeding to estimate $\pi^*(\cdot) \in \Pi$ and $F_1^* \in \mathcal{F}$ from model (2).

Recall that $X \sim m$ having support $\mathcal{X} \subset \mathbb{R}^p$. For a fixed $\pi(\cdot) \in \Pi$ and $F_1 \in \mathcal{F}$, let P_{π,F_1} denote the joint distribution of (X, Y) defined in (2). Also let $\mathcal{P} := \mathcal{P}(\Pi, \mathcal{F})$ denote the class $\{P_{\pi,F_1} : \pi \in \Pi, F_1 \in \mathcal{F}\}$. The main issue with identifiability arises from the fact that, in general, it is possible to represent a given $P \in \mathcal{P}$ as P_{π,F_1} for two (or more) different choices of $\pi \in \Pi$ and $F_1 \in \mathcal{F}$.

Definition 2.1 (Identifiability). We say that $P_{(\pi^*,F_1^*)} \in \mathcal{P}(\Pi, \mathcal{F})$ is *identifiable* if for every function $(\pi, F_1) \in \Pi \times \mathcal{F}$, the condition $P_{(\pi^*,F_1^*)} = P_{(\pi,F_1)}$ implies $\pi(x) = \pi^*(x)$ for m-almost everywhere (a.e.) x, and $F_1(y) = F_1^*(y)$ for all $y \in \mathbb{R}$.

Although model (2) has been considered before by Scott et al. (2015) there has not been a rigorous study of the associated identifiability issues. The following lemma characterizes identifiability in the setting of (2).

Lemma 2.1. Let π, π' be two functions from $\mathcal{X}$ to $[0, 1]$ and let F_1, F_1' be two DFs on $\mathbb{R}$. Consider the following two statements:

(a) The probability distributions P_{π,F_1} and $P_{\pi',F_1'}$ are identical.
(b) There exists a real number $c \neq 1$ such that

$$\pi'(x) = \pi(x)/(1-c) \qquad \text{for } m\text{-a.e. } x, \text{ and} \tag{3}$$

$$F_1'(y) = cF_0(y) + (1-c)F_1(y) \qquad \text{for every } y \in \mathbb{R}. \tag{4}$$

Then,

1. The second statement (b) always implies the first one (a).
2. If we have the conditions $F_0 \neq F_1$ and $\pi(x) > 0$ with positive probability under m (or $F_0 \neq F_1'$ and $\pi'(x) > 0$ with positive probability under m), then the first statement (a) implies the second statement (b).

Remark 2.1 (Nonidentifiability under two-groups model without covariates). When $\Pi := \Pi_{\equiv}$ and $\mathcal{F}$ denotes any of the classes $\mathcal{F}_\downarrow$ or $\mathcal{F}_{\text{Gauss}}$, then the model $P_{(\pi^*,F_1^*)} \in \mathcal{P}(\Pi, \mathcal{F})$, where $\pi^* \in (0, 1)$, is not identifiable. This is an immediate consequence of Lemma 2.1 and has also been observed in Genovese and Wasserman (2004) and Patra and Sen (2016), among others. Thus, for many nonparametric classes $\mathcal{F}$, the absence of covariate information always leads to a nonidentifiable model and it is not possible to recover $\bar{\pi}$. However, there is indeed a way of defining an identifiable mixing proportion in these problems (see, e.g., Genovese and Wasserman 2004; Patra and Sen 2016).

Remark 2.2 (Nonidentifiability under two-groups model with covariates). Quite often there is a natural ordering among the hypotheses to be tested (see, e.g., Li and Barber 2016). In this scenario, a natural choice for the parameters in model (2) are $\Pi := \Pi_\uparrow$, $F_0 \sim \text{Uniform}(0, 1)$ and $\mathcal{F} := \mathcal{F}_\downarrow$. In this setting Lemma 2.1 immediately yields that the model $P_{(\pi^*,F_1^*)} \in \mathcal{P}(\Pi, \mathcal{F})$, where $\pi^*(x) < \delta < 1$ for m-a.e. $x \in \mathcal{X}$ (for some δ), is not identifiable. As a result, for the multiple testing problem when we have p-values for each test, the natural model $F_0 \sim \text{Uniform}(0, 1)$ and $\mathcal{F} := \mathcal{F}_\downarrow$ is nonidentifiable if we model the nonnull proportion as a nondecreasing function of the covariates.

Remark 2.3 (Presence of covariates can restore identifiability). Let $\pi^* \in \Pi$ and $F_1^* \in \mathcal{F}$. Lemma 2.1 shows that if $c\pi^*(\cdot)$ does not belong to Π, for any $c \in (0, 1)$, then $P_{(\pi^*,F_1^*)}$ is identifiable. This shows that for many reasonable model classes Π and $\mathcal{F}$, the presence of covariates (if we can model the observed data correctly) can lead to identifiability. Some examples of such model classes are provided below.

Let us recall the definitions of Π_{logit}, Π_{probit}, $\mathcal{F}_\downarrow$, and $\mathcal{F}_{\text{Gauss}}$ from Section 1. In the following discussion, we will use Lemma 2.1 to investigate the issue of identifiability (in the sense of Definition 2.1) in model (2) when $\Pi = \Pi_{\text{logit}}$ or Π_{probit} and $\mathcal{F} = \mathcal{F}_{\text{Gauss}}$ or $\mathcal{F}_\downarrow$. The following result states that under some assumptions on $\mathcal{X}$ and m, the probability measure $P_{(\pi^*,F_1^*)}$, where $\pi^* \in \Pi_{\text{logit}}$ or Π_{probit}, $F_1^* \in \mathcal{F}_{\text{Gauss}}$ or $\mathcal{F}_\downarrow$, is identifiable as long as $\pi^*(\cdot)$ is not a constant function for m-a.e. x and $F_1^* \neq F_0$.

Lemma 2.2. Consider the class of distributions $\mathcal{P}(\Pi, \mathcal{F})$, with $\Pi := \Pi_g$ where $g(x) = (1 + \exp(-x))^{-1}$ or $g(x) = \Phi(x)$,

and $\mathcal{F} := \mathcal{F}_{\text{Gauss}}$ or $\mathcal{F}_{\downarrow}$. Suppose that the set $\mathcal{X}$ contains a non-empty open subset $\mathcal{X}'$ of $\mathbb{R}^p$ such that the probability measure m assigns strictly positive probability to every open ball contained in $\mathcal{X}'$. Assume that $F_0 \neq F_1^*$ and $\pi^* \in \Pi_g$ is given by $\pi^*(x) := g(\beta_0^* + (\beta^*)^\top x)$ for $x \in \mathcal{X}$ and some $(\beta_0^*, \beta^*) \in \mathbb{R} \times \mathbb{R}^p$. Then $P_{(\pi^*, F_1^*)}$ is identifiable if $\beta^* \neq 0$.

It is worth noting that the assumption on m in Lemma 2.2—namely, there exists an open set $\mathcal{X}'$ such that m assigns positive probability to every non-empty open subset of $\mathcal{X}'$, is not very stringent as any absolutely continuous (with respect to the Lebesgue measure) distribution satisfies this. The other key assumption in Lemma 2.2 is that $\beta^* \neq 0$. This means that if the covariates are relevant (i.e., $\beta^* \neq 0$), then identifiability is restored; compare this with the two-groups model (which corresponds to $\beta^* = 0$) in which case we already know that (1) is not identifiable.

However, the way Lemma 2.2 has been stated, it may not accommodate all discrete covariates alongside the test statistics (p-values or z-scores). Corollary H.1 (also see Appendix H.3 in the supplementary materials) is aimed at addressing this issue. In the supplementary materials, we present a simple example (see Remark H.1) which shows that in the presence of discrete covariates, without certain additional assumptions, model (2) may fail to be identifiable.

3. (Nonparametric) Maximum Likelihood Estimation

In this section, we propose and discuss our main estimation strategy—maximum likelihood—for estimating the unknown parameters in model (2), and state our main theoretical result on the estimation accuracy of our proposed estimators. We will assume in this section that every $F \in \mathcal{F}$ admits a probability density on $\mathbb{R}$ and will denote the class of probability densities corresponding to DFs in $\mathcal{F}$ by $\mathfrak{F}$. Our main examples for $\mathcal{F}$ will be $\mathcal{F}_{\text{Gauss}}$ and $\mathcal{F}_{\downarrow}$; we have already seen that these classes arise naturally in multiple testing problems. Our examples for Π will be $\Pi_{\equiv}$, Π_g and $\Pi_{\uparrow}$. Further, we will denote by $\mathfrak{F}_{\text{Gauss}}$ and $\mathfrak{F}_{\downarrow}$ the classes of densities corresponding to $\mathcal{F}_{\text{Gauss}}$ and $\mathcal{F}_{\downarrow}$, respectively. As we will show, the nonparametric classes $\mathfrak{F}_{\text{Gauss}}$ and $\mathfrak{F}_{\downarrow}$ lend themselves to tuning parameter-free estimation through the method of maximum likelihood. Further, for estimation in the class $\mathfrak{F}_{\text{Gauss}}$ we establish an almost parametric rate of convergence of the MLE (see Theorem 3.1).

3.1. Maximum Likelihood Estimation

Let us denote by f_1^* the unknown density of F_1^*. This reduces model (2) to

$$Y|X = x \sim \pi^*(x) f_1^* + (1 - \pi^*(x)) f_0 \quad \text{and} \quad X \sim m, \tag{5}$$

where f_0 is a known density (corresponding to the DF F_0), and $\pi^*(\cdot) \in \Pi$ and $f_1^* \in \mathfrak{F}$ are the unknown parameters of interest. Here, we discuss estimation of (π^*, f_1^*) based on the principle of maximum likelihood. For any $\pi \in \Pi, f_1 \in \mathfrak{F}$, let us denote the normalized log-likelihood at (π, f_1), up to a constant not depending on the parameters, by

$$\ell(\pi, f_1) := \frac{1}{n} \sum_{i=1}^{n} \log \Big(\pi(X_i) f_1(Y_i) + (1 - \pi(X_i)) f_0(Y_i) \Big) \tag{6}$$

and consider the MLE

$$(\hat{\pi}, \hat{f}_1) := \operatorname*{argmax}_{\pi \in \Pi, f_1 \in \mathfrak{F}} \ell(\pi, f_1). \tag{7}$$

As $\mathfrak{F}$ and Π can be nonparametric classes of functions, the estimator $(\hat{\pi}, \hat{f}_1)$ can be thought of as the nonparametric (NP) MLE in model (5). However, the optimization problem in (7) is often nonconvex which makes it difficult to guarantee the convergence of algorithms to global maximizers. To bypass this issue, we define another class of estimators: call any estimator $(\hat{\pi}^A, \hat{f}_1^A)$ satisfying

$$\prod_{i=1}^{n} \frac{\big(1 - \hat{\pi}^A(X_i)\big) f_0(Y_i) + \hat{\pi}^A(X_i) \hat{f}_1^A(Y_i)}{(1 - \pi^*(X_i)) f_0(Y_i) + \pi^*(X_i) f_1^*(Y_i)} \geq 1 \tag{8}$$

an *approximate NPMLE* (AMLE). In other words, $(\hat{\pi}^A(\cdot), \hat{f}_1^A)$ is an AMLE if it yields a higher likelihood (as in (6)) compared to the true model parameters $(\pi^*(\cdot), f_1^*)$.

3.2. Gaussian Location Mixtures

Let us specialize to the case where f_0 is standard normal, $f_1^* \in \mathfrak{F}_{\text{Gauss}}$ and $\pi^* \in \Pi$ for some class of functions Π. Note that this setting has received a lot of attention in the multiple testing literature (see Scott et al. 2015). In the following discussion, we quantify the Hellinger accuracy of any AMLE in estimating (π^*, f_1^*). As is common in regression problems, we state our results conditional on the covariates $X_1, \ldots, X_n$. For each $i = 1, \ldots, n$, and any $\tilde{\pi} \in \Pi, \tilde{f}_1 \in \mathfrak{F}_{\text{Gauss}}$, define $h_i^2\big((\tilde{\pi}, \tilde{f}_1), (\pi^*, f_1^*)\big)$ as

$$\int \Big(\sqrt{(1 - \tilde{\pi}(X_i)) f_0(y) + \tilde{\pi}(X_i) \tilde{f}_1(y)} - \sqrt{(1 - \pi^*(X_i)) f_0(y) + \pi^*(X_i) f_1^*(y)} \Big)^2 dy.$$

Thus, $h_i^2\big((\hat{\pi}, \hat{f}_1), (\pi^*, f_1^*)\big)$ denotes the squared Hellinger distance between the true and estimated conditional density of Y_i given X_i. Our loss function will be the average of h_i^2, for $i = 1, \ldots, n$:

$$\mathfrak{D}^2\Big((\tilde{\pi}, \tilde{f}_1), (\pi^*, f_1^*)\Big) := \frac{1}{n} \sum_{i=1}^{n} h_i^2\Big((\tilde{\pi}, \tilde{f}_1), (\pi^*, f_1^*)\Big). \tag{9}$$

Our main result below gives a nonasymptotic finite sample upper bound on $\mathfrak{D}((\tilde{\pi}, \tilde{f}_1), (\pi^*, f_1^*))$ conditional on the covariates $X_1, \ldots, X_n$. The bound will involve the complexity of the class Π as measured through covering numbers and metric entropy (see van der Vaart and Wellner 1996, chap. 2, pp. 83–86 for the definitions).

Theorem 3.1. Suppose that the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ are drawn from model (5) for some $\pi^* \in \Pi$ and $f_1^* \in \mathfrak{F}_{\text{Gauss}}$ which can be written as $f_1^*(x) = \int \phi(x-u)dG^*(u)$, $x \in \mathbb{R}$, for some probability measure G^* that is supported on $[-M, M]$ for some $M > 0$, where $\phi(\cdot)$ denotes the standard normal density. Also let $M^* := \max(M, \sqrt{\log n})$. Define the sequence $\{\epsilon_n\}$ as

$$\epsilon_n^2 := n^{-1} \max \left(M^* (\log n)^{3/2}, \inf_{\gamma > 0} \left\{ n\sqrt{\gamma} M^* + H(\gamma, \Pi_n, L^\infty) \right\} \right),$$

where $H(\gamma, \Pi_n, L^\infty)$ is the γ-metric entropy of the class of n-dimensional vectors $\Pi_n := \{(\pi(X_1), \ldots, \pi(X_n)) : \pi \in \Pi\}$ with respect to the uniform metric. Then, given an AMLE $(\hat{\pi}^A, \hat{f}_1^A)$ for estimating (π^*, f_1^*), there exists a universal positive constant K such that for every $t \geq 1$ and $n \geq 2$, we have

$$\mathbb{P}\left\{ \mathfrak{D}\left((\hat{\pi}^A, \hat{f}_1^A), (\pi^*, f_1^*) \right) \geq tK\epsilon_n \middle| X_1, \ldots, X_n \right\} \leq 2n^{-t^2}. \tag{10}$$

Moreover, there exists a universal positive constant C such that for every $n \geq 2$, we have

$$\mathbb{E}\left[\mathfrak{D}^2 \left((\hat{\pi}^A, \hat{f}_1^A), (\pi^*, f_1^*) \right) \middle| X_1, \ldots, X_n \right] \leq C\epsilon_n^2. \tag{11}$$

Remark 3.1. Note that $(\hat{\pi}, \hat{f}_1)$ as defined in (7) clearly satisfies (8) and thus Theorem 3.1 implies that (10) and (11) are true with $(\hat{\pi}^A, \hat{f}_1^A)$ replaced by $(\hat{\pi}, \hat{f}_1)$.

Remark 3.2. The optimization problem in (7) is nonconvex and thus there may be multiple local maxima. Consequently, our proposed algorithms (see Section 4) do not guarantee convergence to a global maximizer. Therefore, Theorem 3.1 is of particular importance (more generally useful in estimation involving nonconvex optimization problems) as it establishes finite sample risk bounds for any AMLE. Moreover, our simulations in Section 6 and Appendix H.5 (in the supplementary materials) illustrate that our proposed algorithms almost always yield estimates that are AMLEs.

Remark 3.3. Under suitable technical assumptions, the same proof technique as that of Theorem 3.1 can be adopted to yield near-parametric rates of convergence for other location mixtures (beyond Gaussian), that is, when $f_1^* \in \mathcal{F}_K$ where

$$\mathcal{F}_K := \{f : f(y) = \int K(y-\theta)\, dG(\theta),\ G(\cdot) \text{ is a probability distribution}\}.$$

Another class of interest is when $f_1^* \in \mathcal{F}_\downarrow$, as often used in p-value modeling (see, e.g., Schweder and Spjøtvoll 1982; Langaas, Lindqvist, and Ferkingstad 2005; Cao, Chen, and Zhang 2020). Note that the metric entropy of $\mathcal{F}_\downarrow$ is larger than that of $\mathcal{F}_{\text{Gauss}}$. Consequently, we do not expect a near-parametric rate of estimation for the conditional density, although our current proof does not cover this case.

The above theorem might look a bit abstract at first glance. Let us consider a typical function class Π to demonstrate the conclusions of Theorem 3.1. Let Π be given by a generalized linear model, that is, each function $\pi \in \Pi$ is of the form $x \mapsto g(x^\top \beta)$ for some $\beta \in \mathbb{R}^p$ and known link function $g(\cdot)$. Then Theorem 3.1 gives a parametric rate of convergence p/n, up to a logarithmic factor of n, in the average Hellinger metric (see (9)), for all standard choices of $g(\cdot)$. This is illustrated in the subsequent corollary and remarks.

Corollary 3.1. Suppose $g : \mathbb{R} \to [0,1]$ is a fixed link function that is Lipschitz with some constant $L > 0$, that is, $|g(z_1) - g(z_2)| \leq L|z_1 - z_2|$, for all $z_1, z_2 \in \mathbb{R}$. Suppose that the covariate space $\mathcal{X}$ is contained in a p-dimensional Euclidean ball of radius T and that the function class Π is given by $\{\pi_\beta : \beta \in \mathbb{R}^p, \|\beta\| \leq R\}$ for some $R > 0$ where $\pi_\beta(x) := g(x^\top \beta)$ for $x \in \mathcal{X}$. Then, under the same assumptions on f_1^* as in Theorem 3.1, inequalities (10) and (11) both hold with $\epsilon_n^2 = \frac{1}{n} \max\left(M^*(\log n)^{3/2}, M^* + p \log\left(1 + 2LTRn^2\right)\right)$. The quantities L, M, R, and T can be taken to be either fixed or changing with n.

Remark 3.4. The most common example of the link function g in Theorem 3.1 is the logistic link given by $g(z) := (1+e^{-z})^{-1}$, for $z \in \mathbb{R}$. This function g is clearly Lipschitz with constant $L = 1$ because $|g'(z)| = e^z(1+e^z)^{-2} \leq 1$ for every $z \in \mathbb{R}$. Another example of the link function $g(\cdot)$ in Theorem 3.1 is the probit link given by $g(z) := \Phi(z)$ for $z \in \mathbb{R}$. This function g is also Lipschitz with constant $L = (2\pi)^{-1/2}$ because $|g'(z)| = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) \leq (2\pi)^{-1/2}$, for every $z \in \mathbb{R}$. Both the logit and probit links arise from symmetric (about 0) densities which may sometimes be undesirable, specially in some survival models. As a result, often the complementary log-log link is recommended in survival models (e.g., Jenkins 1995). In this case $g(z) := 1 - \exp\left(-\exp(z)\right)$, $z \in \mathbb{R}$. Observe that $|g'(z)| \leq 1$. Therefore, Corollary 3.1 applies to all the three link functions above.

Remark 3.5. If L, M, R, and T are all constant, then the rate ϵ_n given by Corollary 3.1 is parametric up to logarithmic factors in n.

In the following section, we describe an iterative approach based on the EM algorithm (Dempster, Laird, and Rubin 1977; McLachlan and Peel 2000; Lange 2016) to compute the MLE described in (7). We had also looked into an alternative maximization based approach for solving (7). Our simulations revealed that the EM algorithm significantly and consistently outperformed the alternative maximization scheme. Hence, we only describe the details of the EM based algorithm.

4. EM Algorithm for Joint Likelihood Maximization

Let us first recall a familiar setting from Section 1. Consider n independent but unobserved (latent) Bernoulli random variables $Z_1, Z_2, \ldots, Z_n$ such that $\mathbb{P}(Z_i = 1|X_i) = \pi^*(X_i)$ for some $\pi^*(\cdot) \in \Pi$ and suppose that the conditional densities of $(Y_i|Z_i = 1, X_i)$ and $(Y_i|Z_i = 0, X_i)$ are f_1^* and f_0, respectively. The EM algorithm then, proceeds as follows. We first write down the "complete data" likelihood which involves the joint density of our observed data $\{(Y_i, X_i)\}_{i=1}^n$ and the latent variables $Z_1, \ldots, Z_n$. Observe that the joint (complete) average

Algorithm 3.1 EM implementation of (7)

Input $\{(Y_i, X_i)\}_{i=1}^n$ and initial estimates $\pi^{(0)}, f_1^{(0)}$
$k \leftarrow 1$
repeat

$$\textbf{E-step: } w_i^{(k)} \leftarrow \frac{\pi^{(k-1)}(X_i) f_1^{(k-1)}(Y_i)}{\pi^{(k-1)}(X_i) f_1^{(k-1)}(Y_i) + \big(1 - \pi^{(k-1)}(X_i)\big) f_0(Y_i)}, \quad i = 1, 2, \ldots, n.$$

$$\textbf{M-step: } \pi^{(k)} \leftarrow \hat{\pi}_{\mathrm{EM}}(\mathbf{w}^{(k)}, \Pi) \qquad \text{and} \qquad f_1^{(k)} \leftarrow \hat{f}_{\mathrm{EM}}(\mathbf{w}^{(k)}, \mathfrak{F})$$

$k \leftarrow k + 1$

until convergence of $\mathbf{w}^{(k)} = (w_1^{(k)}, \ldots, w_n^{(k)})$.

log-likelihood of (X_i, Y_i, Z_i), for $i = 1, \ldots, n$, equals

$$\frac{1}{n} \sum_{i=1}^{n} \Big\{ Z_i \log\big[\pi(X_i) f_1(Y_i)\big] + (1 - Z_i) \log\big[(1 - \pi(X_i)) f_0(Y_i)\big] \Big\},$$

where we have ignored some terms that do not depend on the parameters of interest. Observe that the conditional expectation of Z_i given the data can be expressed as

$$\mathbb{E}_{\pi^*, f_1^*}[Z_i | Y_i = y, X_i = x] = \frac{\pi^*(x) f_1^*(y)}{\pi^*(x) f_1^*(y) + (1 - \pi^*(x)) f_0(y)}, \quad \text{for } i = 1, \ldots, n. \tag{12}$$

As the random variables Z_i's are unobserved, we replace them in the log-likelihood in the E-step of the algorithm by their conditional expectations evaluated as in (12) with $\pi^*(\cdot)$ and f_1^* replaced by their estimates from the previous iteration; see Algorithm 3.1 for details. The obtained expected log-likelihood function is then maximized in the M-step of the algorithm with respect to both the parameters $\pi \in \Pi$ and $f_1 \in \mathfrak{F}$. We provide the corresponding pseudo-code for the EM algorithm below.

In Algorithm 3.1, for any $\mathbf{w} = (w_1, \ldots, w_n) \in [0, 1]^n$,

$$\hat{\pi}_{\mathrm{EM}}(\mathbf{w}, \Pi) := \operatorname*{argmax}_{\pi \in \Pi} \frac{1}{n} \sum_{i=1}^{n} \big[w_i \log \pi(X_i) + (1 - w_i) \log(1 - \pi(X_i)) \big], \quad \text{and} \tag{13}$$

$$\hat{f}_{\mathrm{EM}}(\mathbf{w}, \mathfrak{F}) := \operatorname*{argmax}_{f_1 \in \mathfrak{F}} \frac{1}{n} \sum_{i=1}^{n} \big[w_i \log f_1(Y_i) + (1 - w_i) \log f_0(Y_i) \big]. \tag{14}$$

When the classes Π and $\mathfrak{F}$ are convex (e.g., $\Pi := \Pi_{\uparrow}$, $\mathfrak{F} := \mathfrak{F}_{\mathrm{Gauss}}$ or $\mathfrak{F}_{\downarrow}$), the optimization problems (13) and (14) are also convex in π and f_1, respectively. Further, due to the particular form of the expected log-likelihood, this joint maximization breaks into two isolated maximization problems, that is, problems (13) and (14) are decoupled. Hence, solving (13) (or (14)) requires no knowledge of $\mathfrak{F}$ (or Π). Therefore, both of the above problems are usually more tractable than (7). we defer the more specific details about the implementations of (13) and (14) to the Appendix (see Sections A.1 and A.2). However, as (7) is a nonconvex problem we cannot guarantee the convergence of our EM algorithm to the global maximizer. Moreover, we need proper initial estimates of (π^*, f_1^*) to start the iterative scheme in the EM algorithm (see Section 3.2). In Sections 5.1 and 5.2, we describe two easily implementable procedures that can be used as starting points for the EM algorithm.

5. Marginal Methods

Maximizing the joint likelihood (of (X, Y); see (6)) can be computationally expensive, especially when dealing with nonparametric classes for Π or $\mathfrak{F}$. Further, the EM algorithm proposed in Section 4 to find the MLEs is iterative in nature and can get stuck at a local maxima, different from the global maximizer (as the underlying optimization problem is nonconvex). In this subsection, we propose two novel *marginal* methods that bypass the joint estimation of $\pi^*(\cdot)$ and f_1^*. As the name suggests, these methods do not deal with a joint maximization problem; instead they use properties of model (5) to isolate each of the parameters and estimate them separately. Both the proposed methods are conceptually simple and easy to implement. They also provide good estimates for the true parameters in model (5); in Section 6, we compare their performance to FDR regression (see Scott et al. 2015). Our marginal methods can also be used to obtain preliminary estimators of $\pi^*(\cdot)$ and f_1^* which can then be chosen as starting points for the EM algorithm outlined in Section 4 (see Algorithm 3.1).

5.1. Marginal Method—I

To motivate this decoupled approach, first observe that the marginal distribution of Y in model (2) has the form (1) where $\bar{\pi} := \mathbb{E}_{X \sim m}[\pi(X)]$, which is the standard two-groups model with unknown F_1^* and $\bar{\pi}$. The above observation can be used to directly estimate f_1^* (the density of F_1^*), bypassing the estimation of $\pi^*(\cdot)$. Observe that, if $\bar{\pi} \equiv \alpha$ were known (assume $\overline{\pi} > 0$), estimation of $f_1^* \in \mathfrak{F}$ could be accomplished by maximizing the marginal likelihood of the Y_i's, that is,

$$\hat{f}_1^{(\alpha)} := \operatorname*{argmax}_{f_1 \in \mathfrak{F}} \frac{1}{n} \sum_{i=1}^{n} \log\Big(\alpha f_1(Y_i) + (1 - \alpha) f_0(Y_i)\Big). \tag{15}$$

The above optimization problem is indeed computationally more tractable—note that for function classes $\mathfrak{F}$ that are convex (e.g., $\mathfrak{F}_{\mathrm{Gauss}}$ and $\mathfrak{F}_{\downarrow}$) (15) is a convex program and can be solved efficiently. For instance, we may directly use the convex optimization technique outlined in Section A.2.1 to solve (15) if $\mathfrak{F} = \mathfrak{F}_{\mathrm{Gauss}}$.

Once we obtain an estimator $\hat{f}_1^{(\alpha)}$ of f_1^*, we can maximize the joint log-likelihood just as a function of $\pi(\cdot) \in \Pi$ to obtain

$$\hat{\pi}^{(\alpha)} := \operatorname*{argmax}_{\pi\in\Pi} \frac{1}{n}\sum_{i=1}^{n} \log \left(\pi(X_i)\hat{f}_1^{(\alpha)}(Y_i) + (1-\pi(X_i))f_0(Y_i)\right). \quad (16)$$

Problem (16) is also tractable for a variety of choices of Π. In particular, if $\Pi := \Pi_{\uparrow}$, one can once again use the convex optimization strategy discussed in Section A.5 in the supplementary materials, whereas if $\Pi := \Pi_{\text{logit}}$, we can use the BFGS method discussed in Section A.1.1. Based on the above discussion, we end up with one-step estimators $\hat{\pi}^{(\alpha)}$ and $\hat{f}_1^{(\alpha)}$ of π^* and f_1^* (respectively), if we knew the value of $\bar{\pi} \equiv \alpha$.

In practice $\bar{\pi}$ may not be known, in which case we will need to estimate $\bar{\pi}$ from the data to estimate f_1^* using (15). As we are now in the well-known two-groups model, there are many estimators available for $\bar{\pi}$ (see, e.g., Zhang 1990; Storey 2002; Langaas, Lindqvist, and Ferkingstad 2005; Tang and Zhang 2005; Efron 2010; Patra and Sen 2016). However, the estimation of $\bar{\pi}$ is a difficult problem when $\mathfrak{F}$ is nonparametric (e.g., when $\mathfrak{F} = \mathfrak{F}_{\text{Gauss}}$ or $\mathfrak{F}_{\downarrow}$) and there is no known $\sqrt{n}$-consistent estimator of $\bar{\pi}$ with finite variance (see, e.g., Nguyen and Matias 2014). Note that, when $f_0 \in \mathfrak{F}$ and $\mathfrak{F}$ is convex (e.g., $f_0(\cdot) = \phi(\cdot)$, $\mathfrak{F} = \mathfrak{F}_{\text{Gauss}}$), we cannot obtain a consistent estimator of $\bar{\pi}$ by maximizing (15) jointly with respect to f_1 and α (as the likelihood in such a case will always be maximized at $\alpha = 1$). In fact, $\bar{\pi}$ is a parameter for which a lower (honest) confidence bound can be provided easily (see, e.g., Genovese and Wasserman 2004; Meinshausen and Rice 2006; Patra and Sen 2016) but an upper confidence bound is difficult to obtain (see, e.g., Donoho 1988 for a unified treatment of such "one-sided" parameters).

The methods for estimating $\bar{\pi}$ cited above do not use the covariate information available in our model. Based on extensive simulation studies (see Section 6), we believe that incorporating covariate information in the estimation of $\bar{\pi}$ can lead to a better estimator. In the following display, we propose a possible strategy to estimate $\bar{\pi}$ that uses the joint likelihood of the available data. Note that as defined, both (15) and (16), depend on $\alpha \in (0, 1]$. We can now consider the "profiled" one-dimensional MLE of $\bar{\pi}$:

$$\hat{\bar{\pi}} = \arg\max_{\alpha\in(0,1]} \frac{1}{n}\sum_{i=1}^{n} \log \left(\hat{\pi}^{(\alpha)}(X_i)\hat{f}_1^{(\alpha)}(Y_i) + (1-\hat{\pi}^{(\alpha)}(X_i))f_0(Y_i)\right), \quad (17)$$

where $\hat{f}_1^{(\alpha)}(\cdot)$ is defined in (15), and $\hat{\pi}^{(\alpha)}(\cdot)$ is defined in (16). To solve problem (17), we recommend a grid search over the unit interval $(0, 1]$. One may also start with a standard estimator of $\bar{\pi}$ (using any of the methods from the references cited above), and restrict the grid search to a suitably small neighborhood of the initial estimate.

Below we state a theoretical result which gives finite sample risk bounds for the estimated marginal density of Y. In fact, the following can be interpreted as an estimation accuracy result in the two-groups model (without covariates), that is, model (1) when an upper bound for the signal proportion ($\bar{\pi}$) is known.

Theorem 5.1. Suppose that the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ are drawn from model (5) for some $\pi^* \in \Pi$ and $f_1^* \in \mathfrak{F} = \mathfrak{F}_{\text{Gauss}}$ which can be written as $f_1^*(x) = \int \phi(x-u)dG^*(u)$, for $x \in \mathbb{R}$, and for some probability measure G^* supported on $[-M, M]$ (for some $M > 0$). If $\overline{\pi} \le \alpha \le 1$, we have

$$\mathbb{E}\left[h^2\left(\left(\alpha, \hat{f}_1^{(\alpha)}\right), \left(\overline{\pi}, f_1^*\right)\right)\right] \le \frac{C}{n}\max\left(M^*(\log n)^{3/2}, M^* + \log n\right),$$

where C is a universal constant, $M^* = \max(M, \sqrt{\log n})$ and

$$h^2\left(\left(\alpha, \hat{f}_1^{(\alpha)}\right), \left(\overline{\pi}, f_1^*\right)\right) := \int \left(\sqrt{(1-\alpha)f_0(y) + \alpha\hat{f}_1^{(\alpha)}(y)} - \sqrt{(1-\overline{\pi})f_0(y) + \overline{\pi}f_1^*(y)}\right)^2 dy.$$

Consequently, if $\bar{\pi}$ was known, that is, $\alpha = \bar{\pi}$, then we further have

$$\mathbb{E}\left[\int \left|\hat{f}_1^{(\bar{\pi})}(y) - f_1^*(y)\right| dy\right] \le \frac{4C\bar{\pi}^{-1}}{\sqrt{n}} \cdot \sqrt{\max\left(M^*(\log n)^{3/2}, M^* + \log n\right)}.$$

Remark 5.1. If M does not change with n in Theorem 5.1, then, for $n \ge 3$, we have $\mathbb{E}\left[h^2\left(\left(\alpha, \hat{f}_1^{(\alpha)}\right), \left(\overline{\pi}, f_1^*\right)\right)\right] \le C'(\log n)^2/n$, where C' is a constant free of n but depending on M. In particular, C' can be taken as $C(M+1)$. A similar conclusion holds for $\mathbb{E}\left[\int \left|\hat{f}_1^{(\bar{\pi})}(y) - f_1^*(y)\right| dy\right]$.

5.2. Marginal Method—II

In the previous marginal procedure, we isolated the effect of the unknown density f_1^* and used the marginal distribution of Y to estimate f_1^*. In this subsection we describe a procedure that targets the estimation of $\pi^*(\cdot)$ first. Observe that the regression function of Y on X is

$$\mathbb{E}(Y|X = x) = (1-\pi^*(x))\mu_0 + \pi^*(x)\mu^*, \quad (18)$$

where $\mu_0 := \mathbb{E}_{Y\sim F_0}[Y]$ and $\mu^* := \mathbb{E}_{Y\sim F_1^*}[Y]$. Here μ_0 is known (as F_0 is known) but μ^* is unknown. Thus, the regression function isolates the effect of $\pi^*(\cdot)$, modulo the estimation of μ^*. If $\mu^* \ne \mu_0$ and $\pi^*(\cdot)$ is not a constant function, (18) poses a nonlinear regression problem and we can use the method of least squares to estimate (π^*, μ^*):

$$(\hat{\pi}, \hat{\mu}) := \operatorname*{argmin}_{\pi\in\Pi, \mu\in\mathbb{R}} \sum_{i=1}^{n} \left(Y_i - \mu_0 - \pi(X_i)(\mu - \mu_0)\right)^2. \quad (19)$$

An application of van der Vaart (1998, Theorem 5.23) then yields the following result. For the sake of completeness, we present a proof of the above result in Appendix I.7 (in the supplementary materials).

Theorem 5.2. Suppose that (X, Y) has a joint distribution described by (5) where $\pi^*(\cdot) \in \Pi_g$, that is, $\pi^* \equiv \pi^*_{\beta^*}(x) = g(x^\top\beta^*)$. Also assume that Y and each component of X has a finite fourth moment. Let $g(\cdot)$ be thrice differentiable and the ith derivative of $g(\cdot)$ satisfy $\sup_{\lambda\in\mathbb{R}} |g^{(i)}(\lambda)| \le c_i$ for some constants c_i, $i = 0, 1, 2, 3$. (Note that c_0 can be chosen as

1.) Further assume $\Theta \subset \mathbb{R}^{p+1}$ is a fixed compact set and $\theta^* \equiv (\beta^*, \mu^*) \in \text{int}(\Theta)$ is identifiable from (18) in the sense that $\theta \neq \theta^*$ implies that $\mu g(X^\top \beta) \neq \mu^* g(X^\top \beta^*)$ with positive probability under the measure m. Then, the LSE $\hat{\theta}_n$, defined in (19), is $\sqrt{n}$-consistent, and has an asymptotically normal limit given by $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, V_{\theta^*}^{-1}(\mathbb{E}[\dot{m}_{\theta^*}\dot{m}_{\theta^*}^\top])V_{\theta^*}^{-1})$ as $n \to \infty$. Here $m_\theta(X, Y) := -(Y - \mu\pi_\beta^*(X))^2$, $\dot{m}_\theta = \nabla_\theta m_\theta$ and $V_\theta := \mathbb{E}[\nabla_\theta^2 m_\theta(X, Y)]$ is assumed to be invertible at θ^*.

Corollary 5.1. Recall the choices for $g(\cdot)$ in Remark 3.4: $g(z) = (1 + \exp(-z))^{-1}$, $g(z) = \Phi(z)$, and $g(z) = 1 - \exp(-\exp(z))$. It is straight-forward to check that all these three functions satisfy the assumptions on $g(\cdot)$ in Theorem 5.2. As a result, the asymptotic normality of the LSE $\hat{\theta}_n$ (stated in Theorem 5.2) holds for these three choices of $g(\cdot)$.

Once $\hat{\pi}(\cdot)$ is estimated, we can use the joint likelihood of (X, Y) to estimate f_1^* (plugging in the value of $\hat{\pi}(\cdot)$): $\hat{f}_1 := \text{argmax}_{f_1 \in \mathfrak{F}} \sum_{i=1}^n \log\left[\hat{\pi}(X_i)f_1(Y_i) + (1 - \hat{\pi}(X_i))f_0(Y_i)\right]$. The optimization problems discussed in this section can be solved based on the methods discussed in Appendix A.1 in the supplementary materials. As the least squares problem in (19) can be nonconvex, we recommend fixing μ and optimizing over $\pi(\cdot)$ followed by a grid search in the space of μ. In the following we discuss in detail the estimation of π^*.

6. Simulations

In this section, we discuss the implementation of our methods and compare their performances with the closest existing method in the literature, namely FDRreg in Scott et al. (2015). We use version 0.2 of the FDRreg package (see Scott 2015). We have additionally compared our methods to AdaPT (see Lei and Fithian 2016) and the method proposed in Boca and Leek (2018). However, due to space constraints, we defer these additional simulations to the supplementary materials (see Appendix G). Further, see Appendix D where we compare our methods with the above competitors under model misspecifications.

In our simulations here, we confine ourselves to $\pi^*(\cdot) \in \Pi_{\text{logit}}$ and $f_1^* \in \mathfrak{F}_{\text{Gauss}}$ (as in Scott et al. (2015)). In fact, most of our simulation settings are borrowed from Scott et al. (2015). Additional simulations that highlight the usefulness of (5) over the two-groups model can be found in the supplementary materials (see Appendix B).

6.1. Estimation of Parameters and Multiple Hypotheses Testing

We now document an extensive set of simulations investigating the performance of all our proposed methods: (i) the first marginal method based on profile likelihood maximization (Marginal-I), (ii) the second marginal method based on nonlinear regression (Marginal-II), and (iii) the full MLE (fMLE) implemented via the EM algorithm (see the end of this section for a discussion on the initialization scheme). We also compare our methods to FDRreg, proposed in Scott et al. (2015).

To evaluate the performance of these methods we compute six different metrics, described below. We use $(\check{\pi}, \check{f}_1)$ to denote any generic estimator of (π^*, f_1^*). We also use $\check{\mathfrak{l}}_i$ to denote any generic estimator of $\mathfrak{l}_i^*$, where $\mathfrak{l}_i^*$, the lFDR of the ith observation, is defined as one minus the right hand side in (12) (we discuss the importance of the vector $(\mathfrak{l}_1^*, \ldots, \mathfrak{l}_n^*)$ in greater detail in Appendix E of our supplementary materials). The first three metrics below are directly aimed at understanding the accuracy in the estimation of π^*, f_1^* and the lFDR's, respectively.

(a) Root mean squared error (RMSE) in estimating the vector $(\pi^*(X_1), \ldots, \pi^*(X_n))$: $\left[\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\check{\pi}(X_i) - \pi^*(X_i))^2\right]^{1/2}$.

(b) RMSE in estimating the vector $(f_1^*(Y_1), \ldots, f_1^*(Y_n))$: $\left[\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\check{f}_1(Y_i) - f_1^*(Y_i))^2\right]^{1/2}$.

(c) RMSE in estimating the vector $(\mathfrak{l}_1^*, \ldots, \mathfrak{l}_n^*)$: $\left[\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\check{\mathfrak{l}}_i - \mathfrak{l}_i^*)^2\right]^{1/2}$. Here $\check{\mathfrak{l}}_i$'s are evaluated as one minus the right hand side of (12) with $(\pi^*(\cdot), f_1^*)$ replaced by $(\check{\pi}(\cdot), \check{f}_1)$.

Further, we consider three more measures that are aimed at understanding the efficacy of these methods for the purpose of post-estimation multiple testing.

(d) Underestimation in the vector of lFDRs $(\mathfrak{l}_1^*, \ldots, \mathfrak{l}_n^*)$: $\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\mathfrak{l}_i^* - \check{\mathfrak{l}}_i)_+$. In multiple testing problems, such underestimation may result in too many hypotheses being rejected which may lead to inflated measures of Type I error, such as FDR. Thus, for an efficient multiple testing procedure, we would expect this underestimation metric to be large.

(e) FDR: $\mathbb{E}\left[\dfrac{\text{Number of false rejections}}{\text{Total number of rejections}}\right]$.

(f) True positive rate (TPR): $\mathbb{E}\left[\dfrac{\text{Number of true rejections}}{\text{Total number of nonnull hypotheses}}\right]$.

Measures (e) and (f) can be interpreted as analogs of Type I error and power, respectively. Note that, methods that yield higher values of TPR while keeping FDR under a certain specified threshold, should be considered more effective.

We consider the following choices for $\pi^*(x) := [1 + \exp(-s(x))]^{-1}$: (A) $s(x_1, x_2) = -2 + 3.5x_1^2 - 3.5x_2^2$; (B) $s(x_1, x_2) = -3 + 1.5x_1 + 1.5x_2$; (C) $s(x_1, x_2) = -1 + 9(x_1 - 0.5)^2 - 5|x_2|$; (D) $s(x_1, x_2) = 20(x_1 - 0.75)$.

For the nonnull density f_1^* we choose the following: (i) $f_1^* = 0.4\mathcal{N}(-1.25, 3) + 0.2\mathcal{N}(0, 5) + 0.4\mathcal{N}(1.25, 3)$; (ii) $f_1^* = 0.3\mathcal{N}(0, 1.1) + 0.4\mathcal{N}(0, 2) + 0.3\mathcal{N}(0, 10)$; (iii) $f_1^* = 2^{-1}\mathcal{N}(0.5, 1) + 3^{-1}\mathcal{N}(1, 1.1) + 6^{-1}\mathcal{N}(1.5, 2)$; (iv) $f_1^* = 0.48\mathcal{N}(-2, 2) + 0.04\mathcal{N}(0, 17) + 0.48\mathcal{N}(2, 2)$.

Most of the settings mentioned above, in particular, (A) and (B) for $s(\cdot, \cdot)$, and (i), (ii) and (iv) for f_1^*, have been borrowed from Scott et al. (2015). The settings (A)–(D) capture a broad spectrum of relationships between the covariates and the response: for instance, the graph of $\pi^*(\cdot)$ corresponding to scenario (B) seems relatively flat as (x_1, x_2) varies, whereas the graph of $\pi^*(\cdot)$ from scenario (D) shows a steep change in $\pi^*(\cdot)$ as x_1 exceeds 0.6. Scenarios (A) and (C) are in between these two extremes. Through Figures 3 and H.10 (in the supplementary materials), and Figures H.8 and H.9 (see Appendix H.8 in

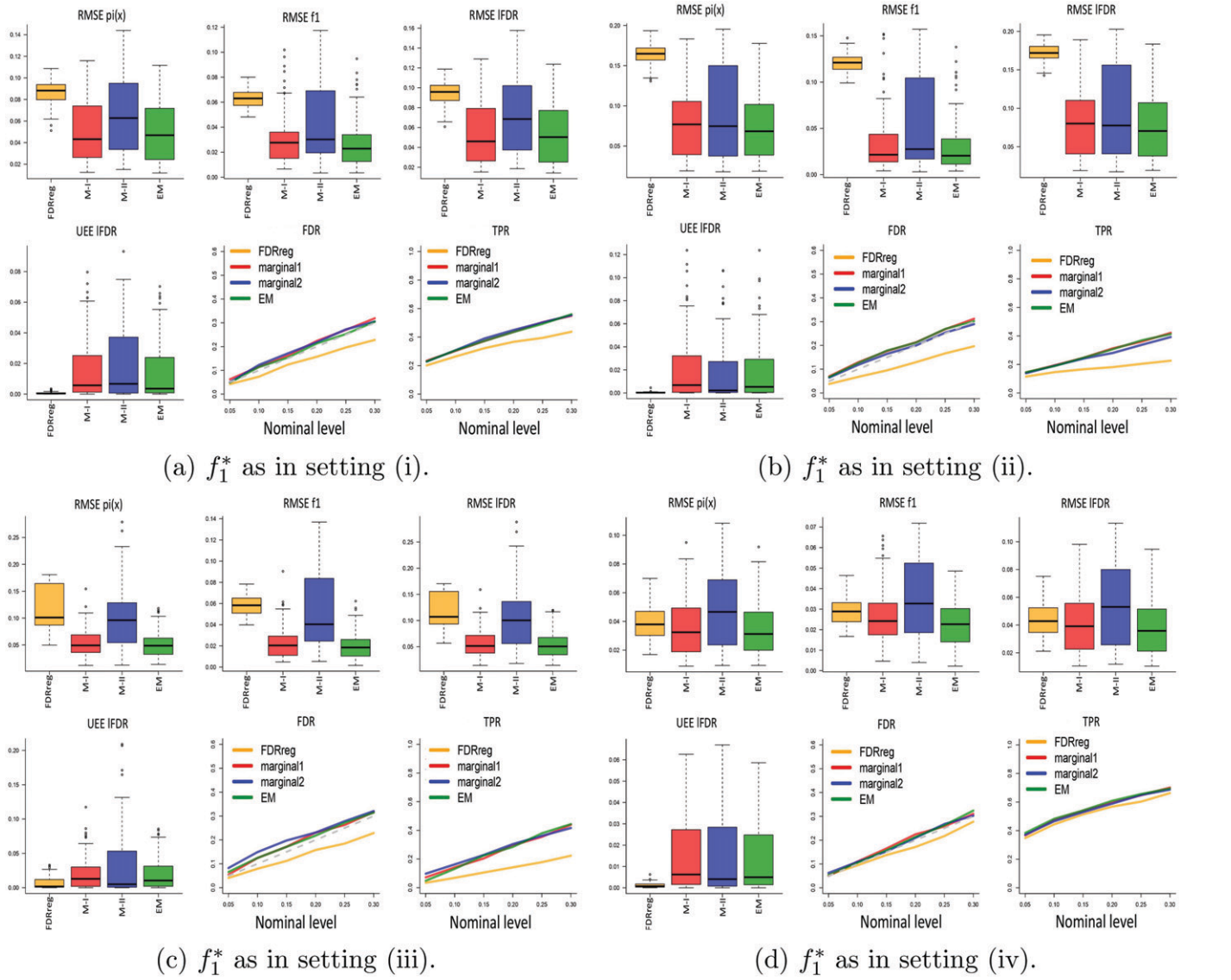

Figure 3. Each subplot shows the performances of FDRreg (in yellow), Marginal-I or M-I (in red), Marginal-II or M-II (in blue) and fMLE (in green) based on the six metrics (a)–(f) in row-major order. The four subplots are obtained for the four different choices of f_1^*, namely (i)–(iv) (in row-major order); UEE there stands for underestimation error from metric (d); the choice of $s(\cdot)$ was fixed at setting (A). For metrics (a)–(d), boxplots are constructed based on 200 replicates. For metrics (e) and (f), the plots show average false discovery and true positive rates computed over 200 replicates for a grid of nominal levels {0.05, 0.10, 0.15, 0.20, 0.25, 0.30}. In the plot depicting FDR (metric (e)), the gray dashed line indicates the nominal level.

the supplementary materials) we illustrate the performance of FDRreg, Marginal-I, Marginal-II, and fMLE in these diverse simulation settings. We observe that our proposed methods consistently outperform FDRreg, in terms of most of the metrics (a)–(f) as discussed above, more so when $\pi^*(\cdot)$ varies significantly with (x_1, x_2).

For each pair of parameters (π^*, f_1^*), we implement the methods—Marginal-I, Marginal-II, fMLE, and FDRreg—on 200 independent replicates each with sample size $n = 10^4$. In each replicate, two-dimensional covariates $X_i = (X_{i1}, X_{i2})$, $i = 1, \ldots, n$, are drawn uniformly at random from the unit square, that is, $[0, 1]^2$. Then $\{Y_i\}_{i=1}^n$ are drawn independently from the mixture density $\pi^*(X_i)f_1^* + (1 - \pi^*(X_i))f_0$. In our simulations we model the covariates, expanded from two dimensions to six dimensions, via basis splines with three degrees of freedom (using a logistic link) as in Scott et al. (2015).

Recall that to compute the fMLE, one has to solve a nonconvex optimization problem and a good starting point is necessary. We initialize this iterative method by choosing the estimate with the highest likelihood value obtained from the other procedures, namely, Marginal-I, Marginal-II, and FDRreg. The EM algorithm is then run for 500 iterations or until convergence (i.e., the iterative change in the norm of the vector of the estimated lFDRs falls below 10^{-6}). Our results are illustrated in Figure 3 (and in Figures H.8, H.9, and H.10 in Appendix H.8). In Table H.4 (see Appendix H.8), we show that Marginal-I most often has the highest likelihood value and thus serves as the initializer for fMLE. With the exception of setting (D)(i), in the same table, we also note that FDRreg was rarely used to initialize fMLE. This shows that, across our simulation settings, estimates from Marginal-I and Marginal-II consistently yield higher likelihoods than those from FDRreg.

6.1.1. Estimation of Model Parameters

We begin our discussion by considering the RMSEs in estimating the unknowns $\pi^*(\cdot)$, f_1^*, and $\mathfrak{l}_i^*$'s, as defined in the metrics (a)–(c); see Figures 3 and H.10 (in the supplementary materials). Note that, fMLE is almost always the most accurate estimator as it results in lower RMSEs (except in Figure H.8(d) in the supplementary materials where FDRreg performs the best). Even Marginal-I and Marginal-II yield better estimates than FDRreg in most settings; except in Figures H.8(d) and H.9(d) for Marginal-I, and in Figures 3(d), H.8(a), H.8(c), and H.9(d) for Marginal-II (see Appendix H.8 in the supplementary materials).

In the interest of fairness however, we point out two specific caveats. First, Figure H.8(d) shows an example where fMLE is outperformed by FDRreg. However, a closer inspection reveals that by slightly tweaking the above simulation setting we observe a completely different outcome, that is, fMLE performs much better than FDRreg (see Appendix H.7 in the supplementary materials for more details). Second, although our methods outperform FDRreg in almost all the settings, they are in general more time consuming to compute than FDRreg. This is expected because FDRreg does not fully use the covariate information while estimating f_1^*, while our methods utilize this covariate information, solving a more complex optimization problem in the process (see Appendix H.8 in the supplementary materials for details).

6.1.2. Multiple Hypotheses Testing

Having established the superiority of fMLE for the purposes of estimating the model parameters, we now move our attention to the application of each of these methods for the purpose of multiple hypotheses testing. As described in Appendix E in the supplementary materials, multiple hypotheses testing is conducted in these settings by estimating the lFDR of each observation and then constructing a set of rejections based on these lFDRs. An overwhelming observation based on the metrics (d)–(f) is the conservatism of FDRreg. In this context, conservatism refers to whether a method leads to substantially lower false rejections than the nominal FDR level it has been set to, and consequently suffers a loss in power. Indeed, in most of the simulations, the underestimation corresponding to FDRreg is almost zero, implying that it regularly overestimates the true lFDRs. As such, false null hypotheses are often accepted by FDRreg, leading to low power (TPR). Thus, based on these simulations it is evident that the FDRreg method frequently produces heavily biased estimates of lFDR, with the bias directed such that FDR control is satisfied but TPR is low.

In contrast, fMLE and the marginal methods do not exhibit such a behavior. Indeed, in all figures except in Figures H.8(a), H.8(b), and H.8(c) in the supplementary materials, Marginal-I, Marginal-II, and fMLE maintain (or only marginally exceed) the nominal level in FDR and are further able to correctly reject more false hypotheses (higher TPR) as compared to FDRreg. We reiterate that one of our goals in the investigation of likelihood based methodology in model (5), beyond the estimation of model parameters, is to construct more powerful multiple testing procedures utilizing the information present in the covariates. As such, we conclude that in most settings, fMLE provides a valid, more powerful multiple testing procedure than FDRreg.

6.2. Related Discussions and Recommendations

In addition to the discussions in this section so far, there are two important observations which we believe augment the utility and reliability of our methods. First, recall the statement of Theorem 3.1. The near-parametric rates that we derive there, for estimating the conditional distribution of Y given X, hold for all AMLEs. A natural question arises: "Do our proposed methods yield AMLEs in practice?" In Appendix H.5, we show results from extensive simulations that illustrate the consistency with which all of our methods (particularly fMLE) result in AMLEs. Second, recall the statement of Remark 3.2 which highlights that the fMLE method solves a nonconvex optimization problem. Therefore, a natural question to ask during implementation is whether the proposed iterative (EM) algorithm is sensitive to the proposed starting points (Marginal-I, Marginal-II, or FDRreg). In Appendix H.6 (see the supplementary materials), our simulations demonstrate that the fMLE approach yields estimates which are mostly stable across the suggested initializations.

Based on our detailed simulation studies (and theoretical results), we would recommend the fMLE method to estimate the unknowns in (5) and consequently address the multiple hypotheses testing problem, especially for moderate sample sizes (at least up to $n = 10^5$). We believe that Marginal-I is possibly the most reliable candidate for producing estimates that may be used to initialize the EM algorithm for computing the fMLEs. It must be pointed out though that we expect the estimates from Marginal-I and Marginal-II initializations of the EM algorithm to be pretty similar; see Appendix H.6 for details. For very large datasets (n over a million), we suggest using Marginal-I instead of fMLE.

Note that, if the two-groups model (1) is adequate for the data, the estimates produced by fMLE (and also FDRreg) can be unreliable, due to identifiability issues (as discussed in Section 2). Therefore, we recommend using the distance covariance based method (see Appendix B in the supplementary materials) first, to understand whether model (1) is adequate, before proceeding with our proposed methodology. However even under identifiability, estimates from model (5) (based on Marginal-I, Marginal-II, fMLE, FDRreg) may turn out to be highly variable (unless n is very large) if the model is nearly nonidentifiable (see Appendix H.7 in the supplementary materials for some discussion on this issue).

7. Real Data Example

7.1. Neuroscience Application

Recall the multiple testing problem discussed in Example 1.1 where we have data arising from the firing rate of 128 V1 neurons in an anesthetized monkey in response to a visual stimulus (see *https://github.com/jgscott/FDRreg/data*). The data consists of 7004 test statistics, each one corresponding to a test of the null hypothesis of no interaction between a neuron pair. The dataset also includes two interesting covariates which capture the spatial and functional relationships among neurons: (a) distance between units, and (b) tuning curve correlation between units; for a more detailed understanding of this experiment (see Kelly

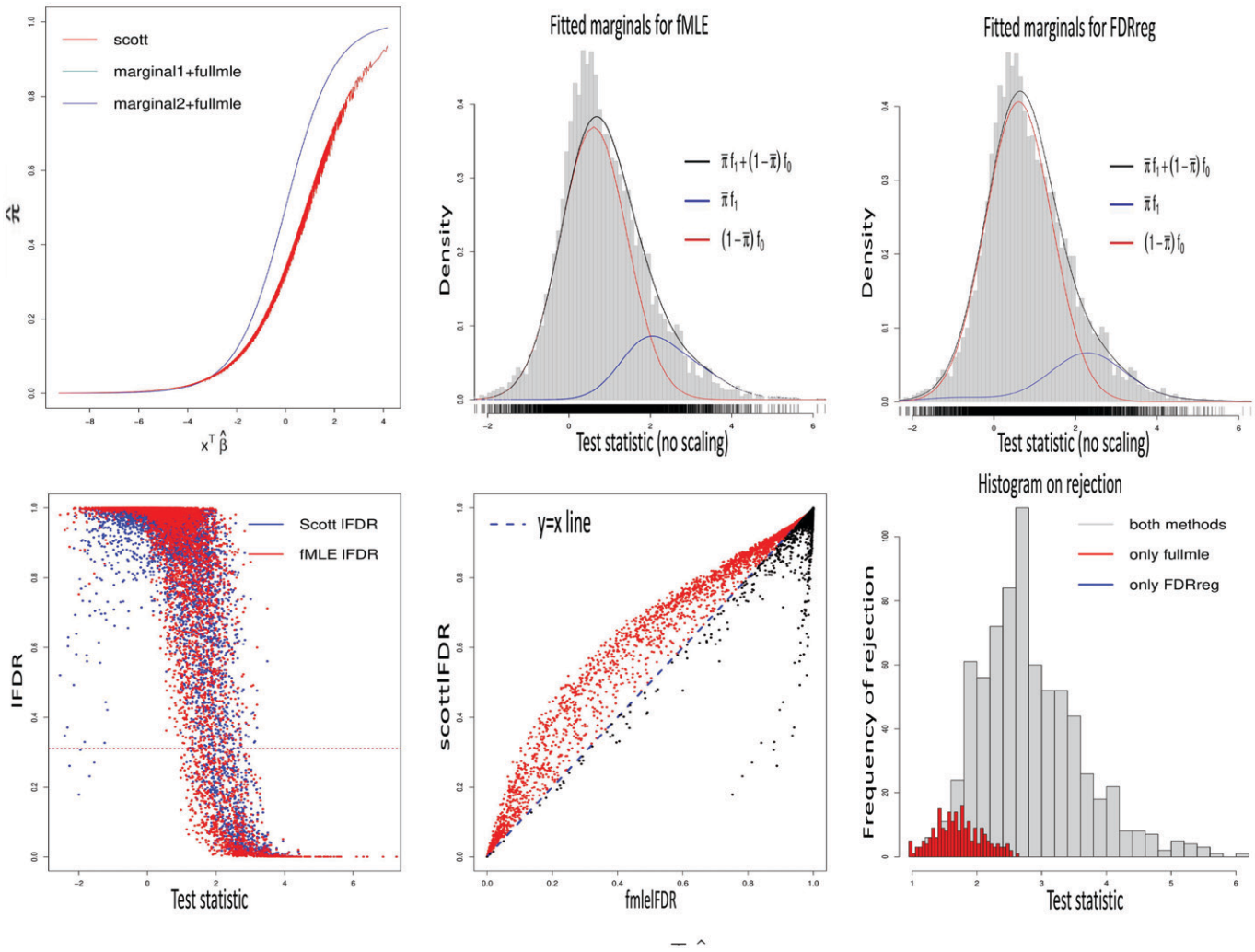

Figure 4. Top left panel: The plot of $\hat{\pi}^*(\cdot)$ against $x^\top \hat{\beta}$ (obtained from fMLE) for the two methods: FDRreg and fMLE (with initializations Marginal-I and Marginal-II) where $\hat{\beta}$ is computed using fMLE. Top center and top right panels: Plots of fitted marginal densities for fMLE and FDRreg, respectively. Bottom left panel: The plot of lFDRs from FDRreg and fMLE with the test statistics plotted along the x-axis. The horizontal line indicates the threshold for rejection for the two methods (which are essentially the same ≈ 0.31). Bottom center panel: Plot of lFDRs from FDRreg versus the same from fMLE (points above and below the $y = x$ line have been colored using red and black, respectively). Bottom right panel: Plot shows the rejection sets plotted across the test statistic for fMLE and FDRreg.

et al. 2007). The primary goal of this study was to detect spiking synchrony among neuron pairs.

For our analysis, we will use the same data processing as has been thoroughly outlined in Scott et al. (2015, secs. 4.2 and 5). In particular, we use a basis spline expansion on the covariates and model the null distribution as a Gaussian with mean and variance estimated using Efron's method of maximum likelihood (see, e.g., Efron 2004). The estimates turn out to be μ (mean) = 0.61 and σ^2 (variance) = 0.66. We model the joint distribution of (Z, X) (here $Z := (Y-\mu)/\sigma$ denotes the centered and scaled test statistic and X denotes the covariate) as in (2) with $F_0 = \Phi(\cdot)$, $\pi^*(\cdot) \in \Pi_{\text{logit}}$ and $F_1^* \in \mathcal{F}_{\text{Gauss}}$. This is slightly different from the approach in Scott et al. (2015) where the authors directly model Y—they take F_0 as $\mathcal{N}(\mu, \sigma^2)$ and $F_1^*(y) := \int \Phi\left(\frac{y-\mu-\theta}{\sigma}\right) dG(\theta)$, where μ and σ are the same as above and G is an unknown DF. To estimate the parameters in our model, we use the methods discussed in Sections 4, 5.1, and 5.2. We then apply the multiple testing proposal from Appendix E with these estimates. We use a nominal level of $\alpha = 0.1$ in our analysis (same as in Scott et al. (2015)). Figure 4 illustrates our findings.

The top left panel in Figure 4 shows that the estimate of $\pi^*(\cdot)$ from fMLE is in general higher than that from FDRreg. From the top center and right panels it looks as though the marginally fitted density from FDRreg fits the data slightly better. However, on observing the test statistic values between -1 and -2, we find that FDRreg estimates a nontrivial contribution of the signal density in that region (see the blue solid line in that region). This leads to smaller lFDRs corresponding to Y values between -1 and -2 (see the bottom left panel) which seems rather counterintuitive. The bottom center panel offers more insight into this observation. Among the 7004 test statistics, the lFDR estimates corresponding to fMLE actually turn out to be higher in over 4000 cases compared to those from FDRreg. However, almost all these cases correspond to points in the top right corner of the bottom center panel (below the $y = x$ line). So, the

fMLE procedure essentially yields higher lFDRs for test statistics which are highly unlikely to be signals.

Correspondingly, the lFDRs based on fMLE are smaller (than FDRreg) in the more critical regions (i.e., where both lFDR estimates are small). In the same plot, observe a sparse cluster near the lower right corner. These points correspond to test statistics in $[-2,-1]$ for which FDRreg yields much lower lFDRs as compared to fMLE. The plot in the bottom right panel illustrates the rejection sets from the two methods. Observe that fMLE admits more rejections than FDRreg. In particular, fMLE rejects 220 more hypotheses (all in the range of test statistics values between 1 and 3). FDRreg rejects 5 more hypotheses all of which correspond to Y values in $[-2,-1]$, which as we mentioned before, seems somewhat counterintuitive. Overall, the fMLE procedure rejects 970 hypotheses out of 7004, at a nominal level of 0.1, whereas FDRreg rejects 755.

Supplementary Materials

The supplementary material, which is available online, contains proofs of our main results, detailed discussions on some of the algorithms we propose in the paper, and additional computational studies.

Acknowledgments

We would like to thank the associate editor and the two anonymous reviewers for their constructive comments that helped improve the quality of this article.

Funding

Adityanand Guntuboyina was sSupported by NSF CAREER grant DMS-16-54589 and Bodhisattva Sen was supported by NSF grants DMS-17-12822 and AST-16-14743.

References

Barber, R. F., and Candès, E. J. (2015), "Controlling the False Discovery Rate via Knockoffs," *The Annals of Statistics*, 43, 2055–2085. [1823]

Basu, P., Cai, T. T., Das, K., and Sun, W. (2018), "Weighted False Discovery Rate Control in Large-Scale Multiple Testing," *Journal of the American Statistical Association*, 113, 1172–1183. [1823]

Benjamini, Y., and Bogomolov, M. (2014), "Selective Inference on Multiple Families of Hypotheses," *Journal of the Royal Statistical Society*, Series B, 76, 297–318. [1823]

Boca, S. M., and Leek, J. T. (2018), "A Direct Approach to Estimating False Discovery Rates Conditional on Covariates," *PeerJ*, 6, e6035. [1829]

Cai, T. T., and Jin, J. (2010), "Optimal Rates of Convergence for Estimating the Null Density and Proportion of Nonnull Effects in Large-Scale Multiple Testing," *The Annals of Statistics*, 38, 100–145. [1820,1822]

Cao, H., Chen, J., and Zhang, X. (2020), "Optimal False Discovery Rate Control for Large Scale Multiple Testing With Auxiliary Information," *The Annals of Statistics* (submitted). [1826]

Dai, H., and Charnigo, R. (2007), "Inferences in Contaminated Regression and Density Models," *Sankhyā*, 69, 842–869. [1820]

Deb, N., Saha, S., Guntuboyina, A., and Sen, B. (2020), "NPMLEmix: Two-Groups Mixture Model with Covariates," R Package Version 1.1. [1823]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Series B, 39, 1–38. [1822,1826]

Dobriban, E. (2016), "A General Convex Framework for Multiple Testing With Prior Information," arXiv no. 1603.05334. [1823]

Donoho, D. L. (1988), "One-Sided Inference About Functionals of a Density," *The Annals of Statistics*, 16, 1390–1420. [1828]

Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104. [1823,1832]

——— (2008), "Microarrays, Empirical Bayes and the Two-Groups Model," *Statistical Science*, 23, 1–22. [1820,1823]

——— (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Institute of Mathematical Statistics (IMS) Monographs (Vol. 1), Cambridge: Cambridge University Press. [1820,1821,1823,1828]

Genovese, C. R., Roeder, K., and Wasserman, L. (2006), "False Discovery Control With p-Value Weighting," *Biometrika*, 93, 509–524. [1823]

Genovese, C. R., and Wasserman, L. (2004), "A Stochastic Process Approach to False Discovery Control," *The Annals of Statistics*, 32, 1035–1061. [1821,1822,1823,1824,1828]

Hu, J. X., Zhao, H., and Zhou, H. H. (2010), "False Discovery Rate Control With Groups," *Journal of the American Statistical Association*, 105, 1215–1227. [1823]

Ignatiadis, N., and Huber, W. (2017), "Covariate Powered Cross-Weighted Multiple Testing," arXiv no. 1701.05179. [1823]

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016), "Data-Driven Hypothesis Weighting Increases Detection Power in Genome-Scale Multiple Testing," *Nature Methods*, 13, 577. [1823]

Ignatiadis, W. (2018), "Covariate Powered Cross-Weighted Multiple Testing," Unpublished manuscript, available at *https://arxiv.org/pdf/1701.05179v3.pdf*. [1823]

Jenkins, S. P. (1995), "Easy Estimation Methods for Discrete-Time Duration Models," *Oxford Bulletin of Economics and Statistics*, 57, 129–136. [1826]

Johnstone, I. M., and Silverman, B. W. (2004), "Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences," *The Annals of Statistics*, 32, 1594–1649. [1823]

Kelly, R. C., Smith, M. A., Samonds, J. M., Kohn, A., Bonds, A., Movshon, J. A., and Lee, T. S. (2007), "Comparison of Recordings From Microelectrode Arrays and Single Electrodes in the Visual Cortex," *Journal of Neuroscience*, 27, 261–264. [1820,1832]

Langaas, M., Lindqvist, B. H., and Ferkingstad, E. (2005), "Estimating the Proportion of True Null Hypotheses, With Application to DNA Microarray Data," *Journal of the Royal Statistical Society*, Series B, 67, 555–572. [1822,1826,1828]

Lange, K. (2016), *MM Optimization Algorithms*, Philadelphia, PA: Society for Industrial and Applied Mathematics. [1822,1826]

Lei, L., and Fithian, W. (2016), "Adapt: An Interactive Procedure for Multiple Testing With Side Information," *Journal of the Royal Statistical Society*, Series B, 80, 649–679. [1823,1829]

Lemdani, M., and Pons, O. (1999), "Likelihood Ratio Tests in Contamination Models," *Bernoulli*, 5, 705–719. [1820]

Li, A., and Barber, R. F. (2016), "Multiple Testing With the Structure Adaptive Benjamini-Hochberg Algorithm," *Journal of the Royal Statistical Society*, Series B, 81, 45–74. [1822,1823,1824]

——— (2017), "Accumulation Tests for FDR Control in Ordered Hypothesis Testing," *Journal of the American Statistical Association*, 112, 837–849. [1821,1822]

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, Wiley Series in Probability and Statistics: Applied Probability and Statistics, New York: Wiley-Interscience. [1820,1826]

McLachlan, G. J., Bean, R., and Peel, D. (2002), "A Mixture Model-Based Approach to the Clustering of Microarray Expression Data," *Bioinformatics*, 18, 413–422. [1823]

McLachlan, G. J., and Wockner, L. (2010), "Use of Mixture Models in Multiple Hypothesis Testing With Applications in Bioinformatics," in *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization, eds. H. Locarek-Junge and C. Weihs, Berlin: Springer, pp. 177–184. [1823]

Meinshausen, N., and Rice, J. (2006), "Estimating the Proportion of False Null Hypotheses Among a Large Number of Independently Tested Hypotheses," *The Annals of Statistics*, 34, 373–393. [1823,1828]

Müller, P., Parmigiani, G., and Rice, K. (2007), "FDR and Bayesian Multiple Comparisons Rules," in *Bayesian Statistics* (Vol. 8), eds. J. M. Bernardo,

M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford: Oxford University Press, pp. 349–370. [1823]

Nguyen, V. H., and Matias, C. (2014), "On Efficient Estimators of the Proportion of True Null Hypotheses in a Multiple Testing Setup," *Scandinavian Journal of Statistics*, 41, 1167–1194. [1828]

Patra, R. K., and Sen, B. (2016), "Estimation of a Two-Component Mixture Model With Applications to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 78, 869–893. [1822,1823,1824,1828]

Robin, A. C., Reylé, C., Derrière, S., and Picaud, S. (2003), "A Synthetic View on Structure and Evolution of the Milky Way," *Astronomy and Astrophysics*, 409, 523–540. [1821]

Robin, S., Bar-Hen, A., Daudin, J.-J., and Pierre, L. (2007), "A Semi-Parametric Approach for Mixture Models: Application to Local False Discovery Rate Estimation," *Computational Statistics & Data Analysis*, 51, 5483–5493. [1823]

Schildknecht, K., Tabelow, K., and Dickhaus, T. (2016), "More Specific Signal Detection in Functional Magnetic Resonance Imaging by False Discovery Rate Control for Hierarchically Structured Systems of Hypotheses," *PLOS ONE*, 11, e0149016. [1821]

Schweder, T., and Spjøtvoll, E. (1982), "Plots of *p*-Values to Evaluate Many Tests Simultaneously," *Biometrika*, 69, 493–502. [1822,1826]

Scott, J. G., and Berger, J. O. (2006), "An Exploration of Aspects of Bayesian Multiple Testing," *Journal of Statistical Planning and Inference*, 136, 2144–2162. [1823]

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P., and Kass, R. E. (2015), "False Discovery Rate Regression: An Application to Neural Synchrony Detection in Primary Visual Cortex," *Journal of the American Statistical Association*, 110, 459–471. [1820,1821,1822,1824,1825,1827,1829,1830,1832]

Scott, J. G., with contributions from Rob Kass and Jesse Windle (2015), "FDRreg: False Discovery Rate Regression," R Package Version 0.2-1. [1829]

Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Series B, 64, 479–498. [1820,1823,1828]

——— (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the *q*-Value," *The Annals of Statistics*, 31, 2013–2035. [1820,1823]

Tang, W., and Zhang, C. (2005), "Bayes and Empirical Bayes Approaches to Controlling the False Discovery Rate," Technical Report, Technical Report# 2005-004, Department of Statistics, Rutgers University. [1828]

Teicher, H. (1961), "Identifiability of Mixtures," *The Annals of Mathematical Statistics*, 32, 244–248. [1824]

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, Chichester: Wiley. [1824]

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics (Vol. 3), Cambridge: Cambridge University Press. [1828]

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Series in Statistics, New York: Springer-Verlag. [1825]

Walker, M. G., Mateo, M., and Olszewski, E. W. (2009), "Stellar Velocities in the Carina, Fornax, Sculptor, and Sextans dSph Galaxies: Data From the Magellan/MMFS Survey," *The Astronomical Journal*, 137, 3100. [1821]

Walker, M. G., Mateo, M., Olszewski, E. W., Sen, B., and Woodroofe, M. (2009), "Clean Kinematic Samples in Dwarf Spheroidals: An Algorithm for Evaluating Membership and Estimating Distribution Parameters When Contamination Is Present," *The Astronomical Journal*, 137, 3109. [1820,1821,1823]

Zhang, C.-H. (1990), "Fourier Methods for Estimating Mixing Densities and Distributions," *The Annals of Statistics*, 18, 806–831. [1828]