Nonparametric maximum likelihood for mixture models: A convex optimization approach to fitting arbitrary multivariate mixing distributions

Long Feng Lee H. Dicker

Department of Statistics and Biostatistics
Rutgers University
Piscataway, NJ 08854
e-mail: long.feng@rutgers.edu; ldicker@stat.rutgers.edu

Abstract: Nonparametric maximum likelihood (NPML) for mixture models is a technique for estimating mixing distributions, which has a long and rich history in statistics going back to the 1950s (Kiefer and Wolfowitz, 1956; Robbins, 1950). However, NPML-based methods have been considered to be relatively impractical because of computational and theoretical obstacles. Recent work focusing on approximate NPML methods and leveraging modern computing power suggests, on the other hand, that these methods may have great promise for many interesting applications. Most of this recent work has focused on specific examples involving relatively simple statistical models and univariate mixing distributions. In this paper, we propose a general approach to fitting arbitrary multivariate mixing distributions with NPML-based methods via convex optimization. The proposed methods are highly flexible and easy to implement. We illustrate their performance in several applications involving estimation and classification.

1. Introduction

Consider a setting where we have iid observations from a mixture model. More specifically, let G_0 be a probability distribution on $\mathcal{T} \subseteq \mathbb{R}^d$ and let $\{F_0(\cdot|\theta)\}_{\theta \in \mathcal{T}}$ be a family of probability distributions on $\mathbb{R}^n$ indexed by the parameter $\theta \in \mathcal{T}$. Assume that $X_1, ..., X_p \in \mathbb{R}^n$ are observed iid random variables and that $\Theta_1, ..., \Theta_p \in \mathbb{R}^d$ are corresponding iid latent variables, which satisfy

$$X_j|\Theta_j \sim F_0(\cdot|\Theta_j) \text{ and } \Theta_j \sim G_0. \tag{1}$$

In (1), it is implicitly assumed that $F_0(\cdot|\theta)$ and G_0 are known (pre-specified) distributions. In this paper, we address issues that arise when the mixing distribution G_0 is unknown (we will assume that $F_0(\cdot|\theta)$ is known throughout). Bayesian nonparametrics provides one popular solution in problems where G_0 is unknown; in particular, one might assume that the distribution G_0 is itself random, e.g. that it is drawn from a Dirichlet process. Further pursuing this idea leads to Bayesian nonparametric mixture models and Dirichlet process mixture models, which have been widely studied (e.g. Antoniak, 1974; Blei and Jordan, 2006; McAuliffe et al., 2006; Neal, 2000). Here, we take a different approach in problems where G_0 is unknown: We fit G_0 by nonparametric maximum likelihood.

Nonparametric maximum likelihood methods for mixture models have been studied in statistics since the 1950s (Kiefer and Wolfowitz, 1956; Robbins, 1950). They provide a well-known and elegant approach to many problems. Moreover, there is a long line of research studying algorithms for computing nonparametric maximum likelihood estimators (NPMLEs) in mixture models and theory (e.g. Böhning, 1995; Ghosal and Van der Vaart, 2001; Jiang and Zhang, 2009; Laird, 1978; Lesperance and Kalbfleisch, 1992; Lindsay, 1995; Martin, 2009). However, implementing and analyzing NPMLEs for mixture models has historically been considered very challenging (e.g. p. 571 of DasGupta, 2008; Donoho and Reeves, 2013).

Recently, Koenker and Mizera (2014) studied convex approximations to NPMLEs for mixture models in relatively large-scale problems, with up to 100,000s of observations. Their focus on convexity and scalability is one of the key concepts for this paper, where we advocate a practical and flexible approach to NPMLEs for mixture models. Koenker and Mizera (2014) showed that in the Gaussian location model, where $X_j = \Theta_j + Z_j \in \mathbb{R}$ and $\Theta_j \sim G_0$, $Z_j \sim N(0,1)$ are independent, a good approximation to the NPMLE for G_0 can be accurately and rapidly computed using generic interior point methods.

While Koenker and Mizera (2014) focus on the Gaussian location model and some slight generalizations, in this paper we argue that their ideas apply much more broadly. We propose a class of approximate NPMLEs for G_0, under the the general mixture model (1), which may be found by convex optimization. These NPMLEs allow for multivariate distributions G_0 with arbitrary dependence structures. After computing the NPMLE, denoted $\hat{G}$, inference in (1) may be conducted via the posterior distribution $\Theta_j|X_j$, under the assumption that $G_0 = \hat{G}$. This is the empirical Bayes approach to data analysis (Efron, 2010; Robbins, 1956; Zhang, 2003). Computing the posterior in this setting is often very simple. By contrast, in the Bayesian nonparametric approach mentioned above, posterior inference involves an additional hierarchical level, since G_0 is modeled to be random, and is typically more computationally challenging (Blei and Jordan, 2006; Neal, 2000).

It should be noted that the class of problems addressed by NPML methods for mixture models differs somewhat from those often addressed by nonparametric Bayes. Dirichlet process mixture models in particular are often used in clustering or clustering-related applications, because the prior distribution on G_0 is supported on discrete measures (Antoniak, 1974). Clustering is a less natural application for our methods, because the distribution G_0 is arbitrary, e.g. G_0 may be continuous. However, given its flexibility, we expect that the NPML approach will be useful in a wide range of interesting applications. In this paper, we describe applications of multivariate NPMLEs involving estimation and classification problems, and show that they are effective in experiments with simulated and real data (Sections 5–6). Other nonparametric Bayesian models do support continuous distributions G_0 (e.g. Polya trees; Lavine, 1992; Mauldin et al., 1992); however, existing approaches for these methods seem less conveniently adapted to handling mixture models like (1).

2. NPMLEs for mixture models via convex optimization

2.1. NPMLEs

Let $\mathcal{G}_{\mathcal{T}}$ denote the class of all probability distribution on $\mathcal{T} \subseteq \mathbb{R}^d$ and suppose that $f_0(\cdot|\theta)$ is the probability density corresponding to $F_0(\cdot|\theta)$ (with respect to some given base measure). For $G \in \mathcal{G}_{\mathcal{T}}$, the (negative) log-likelihood given the data $X_1, ..., X_p$ is

$$\ell(G) = -\frac{1}{p}\sum_{j=1}^{p} \log\left\{\int_{\mathcal{T}} f_0(X_j|\theta)\; dG(\theta)\right\}.$$

The Kiefer-Wolfowitz (1956) NPMLE for G_0, denoted $\hat{G}$, solves the optimization problem

$$\min_{G \in \mathcal{G}_{\mathcal{T}}} \ell(G); \tag{2}$$

in other words, $\ell(\hat{G}) = \min_{G \in \mathcal{G}_{\mathcal{T}}} \ell(G)$.

Solving (2) and studying properties of $\hat{G}$ forms the basis for basically all of the existing research into NPMLEs for mixture models (including this paper). Two important observations have had significant, but somewhat countervailing, effects on this research (2):

(i) The optimization problem (2) is convex.
(ii) If $f_0(X_j|\theta)$ and $\mathcal{T}$ satisfy certain (relatively weak) regularity conditions, then $\hat{G}$ exists and may be chosen so that it is a discrete measure supported on at most p points.

The first observation above is obvious; the second summarizes Theorems 18–21 of (Lindsay, 1995). Among the more significant regularity conditions mentioned in (ii) is that the set $\{f_0(X_j|\theta)\}_{\theta\in\mathcal{T}}$ should be bounded for each $j = 1, ..., p$.

Observation (i) leads to KKT-like conditions that characterize $\hat{G}$ in terms of the gradient of ℓ and can be used to develop algorithms for solving (2), e.g. (Lesperance and Kalbfleisch, 1992). While this approach is somewhat appealing, (2) is typically an infinite-dimensional optimization problem (whenever $\mathcal{T}$ is infinite). Hence, there are infinitely many KKT conditions to check, which is generally impossible in practice.

On the other hand, observation (ii) reduces (2) to a finite-dimensional optimization problem. Indeed, (ii) implies that $\hat{G}$ can be found by restricting attention in (2) to $G \in \mathcal{G}_p$, where $\mathcal{G}_p$ is the set of discrete probability measures supported on at most p points in $\mathcal{T}$. Thus, finding $\hat{G}$ is reduced to fitting a finite mixture model with at most p components. This is usually done with the EM-algorithm (Laird, 1978), where in practice one may restrict to $G \in \mathcal{G}_q$ for some $q < p$. However, while (ii) reduces (2) to a finite-dimensional problem, we have lost convexity:

$$\min_{G \in \mathcal{G}_q} \ell(G) \tag{3}$$

is not a convex problem because $\mathcal{G}_q$ is nonconvex. When q is large (and recall that the theory suggests we should take $q = p$), well-known issues related to nonconvexity and finite mixture models become a significant obstacle (McLachlan and Peel, 2004).

2.2. A simple finite-dimensional convex approximation

In this paper, we take a very simple approach to (approximately) solving (2), which maintains convexity and immediately reduces (2) to a finite-dimensional problem. Consider a pre-specified finite grid $\Lambda \subseteq \mathcal{T}$. We study estimators $\hat{G}_\Lambda$, which solve

$$\min_{G \in \mathcal{G}_\Lambda} \ell(G). \tag{4}$$

The key difference between (3) and (4) is that $\mathcal{G}_\Lambda$, and hence (4), is convex, while $\mathcal{G}_q$ is nonconvex. Additionally, (4) is a finite-dimensional optimization problem, because Λ is finite.

To derive a more convenient formulation of (4), suppose that

$$\Lambda = \{t_1, ..., t_q\} \subseteq \mathcal{T} \tag{5}$$

and define the simplex $\Delta^{q-1} = \{w = (w_1, ..., w_q) \in \mathbb{R}^q;\ w_l \geq 0,\ w_1 + \cdots + w_q = 1\}$. Additionally, let δ_t denote a point mass at $t \in \mathbb{R}^d$. Then there is a correspondence between $G = \sum_{k=1}^q w_k \delta_{t_k} \in \mathcal{G}_\Lambda$ and points $w = (w_1, ..., w_q) \in \Delta^{q-1}$. It follows that (4) is equivalent to the optimization problem over the simplex,

$$\min_{w \in \Delta^{q-1}} -\frac{1}{p} \sum_{j=1}^{p} \log \left\{ \sum_{k=1}^{q} f_0(X_j | t_k) w_k \right\}. \tag{6}$$

Researchers studying NPMLEs have previously considered estimators like $\hat{G}_\Lambda$, which solve (4)–(6). However, most have focused on relatively simple models with univariate mixing distributions G_0 (Böhning et al., 1992; Jiang and Zhang, 2009; Koenker and Mizera, 2014). In a very recent preprint, Gu and Koenker (2014) study a bivariate NPMLE $\hat{G}_\Lambda$ for the Gaussian location-scale model (described in Example 2 below) and applications involving modeling income dynamics. The main distinction of this paper is its generality: We argue that the approach described here provides a very broad, yet practical framework for using NPMLEs to fit arbitrary multivariate mixing distributions via convex optimization.

3. Performance characteristics of $\hat{G}_\Lambda$

As remarked in Section 1, this paper takes an empirical Bayes approach to data analysis (more specifically, a nonparametric empirical Bayes approach, because the distribution G_0 is estimated nonparametrically). Inference is conducted using the posterior distributions $\Theta_j|X_j$, based on the assumption that $\Theta_j \sim \hat{G}_\Lambda$, where the estimator $\hat{G}_\Lambda$ takes the place of the unknown distribution G_0. The accuracy of $\hat{G}_\Lambda$ may substantially affect the quality of inference. Detailed theoretical results for nonparametric empirical Bayes methods (e.g. rates of convergence, optimality results) tend to be very challenging and existing results are largely limited to a few simple models (e.g. Gaussian or Poisson models) (Brown and Greenshtein, 2009; Brown et al., 2013; Ghosal and Van der Vaart, 2001; Jiang and Zhang, 2009).

We do not attempt to provide an exhuastive theoretical account of NPMLE-based empirical Bayes methods here. Instead, we aim to provide practical guidelines for choosing the grid Λ in (4) and for identifying applications where the estimator $\hat{G}_\Lambda$ is likely to be effective (e.g. prediction, classification). However, we believe that theoretical work on empirical Bayes and NPLMEs for mixture models is an important and promising area for future research.

3.1. Choosing Λ

Our perspective is that the estimator $\hat{G}_\Lambda$ is an approximation to $\hat{G}$ and that its performance characteristics are inherited from $\hat{G}$. In general, it is clear that $\hat{G}_\Lambda \neq \hat{G}$. However, as one selects larger and larger finite grids $\Lambda \subseteq \mathcal{T}$, which are more and more dense in $\mathcal{T}$, evidently $\hat{G}_\Lambda \to \hat{G}$. Thus, heuristically, as long as the grid Λ is "dense enough" in $\mathcal{T}$, $\hat{G}_\Lambda$ should perform similarly to $\hat{G}$.

In practice, we propose a two-step approach to identifying a sufficiently dense grid $\Lambda \subseteq \mathcal{T}$: (i) Find a compact subset $\mathcal{T}_0 \subseteq \mathcal{T}$ so that (2) is equivalent to

$$\inf_{G \in \mathcal{G}_{\mathcal{T}_0}} \ell(G); \tag{7}$$

(ii) choose $\Lambda \subseteq \mathcal{T}_0 \subseteq \mathcal{T}$ to be a regular grid with q points, for some sufficiently large q. Numerical results in Section 5 provide some practical guidance for choosing q. Additionally, existing theoretical results suggest that in some simple models where G is univariate, if $q = \sqrt{p}$, then $\hat{G}_\Lambda$ is statistically indistinguishable from $\hat{G}$ (Dicker and Zhao, 2014).

Proposition 1 below describes a subset $\mathcal{T}_0 \subseteq \mathcal{T}$ such that (2) and (7) are equivalent. The main usefulness of Proposition 1 is that $\mathcal{T}_0$ is often compact, even when $\mathcal{T}$ is not; this provides a foothold to identifying the grid Λ, as described in the previous paragraph. In fact, most of the work in Proposition 1 lies in precisely defining $\mathcal{T}_0$.

For $\theta = (\theta_1, ..., \theta_d) \in \mathcal{T}$ and $m = 1, ..., d$, define the subset of $\mathbb{R}$,

$$\mathcal{T}^m(\theta) = \{\vartheta \in \mathbb{R};\ (\theta_1, ..., \theta_{m-1}, \vartheta, \theta_{m+1}, ..., \theta_d) \in \mathcal{T}\}.$$

Additionally, define the single-observation univariate conditional likelihood $L_j^m(\cdot|\theta) : \mathcal{T}^m(\theta) \to \mathbb{R}$ by

$$L_j^m(\vartheta|\theta) = f_0(X_j|\theta_1, ..., \theta_{m-1}, \vartheta, \theta_{m+1}, ..., \theta_d),$$

and the corresponding univariate conditional MLE $T_j^m(\theta) = \operatorname{argmax}_{\vartheta \in \mathcal{T}^m(\theta)} L_j^m(\vartheta|\theta)$, for $j = 1, ..., p$, $m = 1, ..., d$, and $\theta \in \mathcal{T}$. Next, let

$$T^1_{(1)} = \inf\{T^1_j(\theta);\ \theta \in \mathcal{T},\ j = 1, ..., p\}$$
$$T^1_{(p)} = \sup\{T^1_j(\theta);\ \theta \in \mathcal{T},\ j = 1, ..., p\}$$

and define the subset

$$\mathcal{T}^1 = \left\{\theta \in \mathcal{T};\ \theta_1 \in [T^1_{(1)}, T^1_{(p)}]\right\} \subseteq \mathbb{R}^d.$$

Finally, for $1 < m \le d$, inductively define

$$\begin{aligned} T^m_{(1)} &= \inf\{T^m_j(\theta);\ \theta \in \mathcal{T}^{m-1},\ j = 1, ..., p\}, \\ T^m_{(p)} &= \sup\{T^m_j(\theta);\ \theta \in \mathcal{T}^{m-1},\ j = 1, ..., p\}, \\ \mathcal{T}^m &= \left\{\theta \in \mathcal{T}^{m-1};\ \theta_m \in [T^m_{(1)}, T^m_{(p)}]\right\} \subseteq \mathbb{R}^d. \end{aligned}$$

We take $\mathcal{T}_0 = \mathcal{T}^d$. Observe that if $\mathcal{T} = \prod_{m=1}^d I_m \subseteq \mathbb{R}^d$ is a rectangular region, with each $I_m \subseteq \mathbb{R}$ a (possibly unbounded) interval, then

$$\mathcal{T}_0 = \prod_{m=1}^d [T^m_{(1)}, T^m_{(p)}] \subseteq \mathcal{T} \subseteq \mathbb{R}^d. \tag{8}$$

Proposition 1. *Suppose that $\mathcal{T}$ is a rectangle and that the univariate single-observation likelihoods $L_j^m(\cdot|\theta)$ are bounded and unimodal, for $j = 1, ..., p$, $m = 1, ..., d$, and $\theta \in \mathcal{T}$. Then*

$$\ell(\hat{G}) = \inf_{G \in \mathcal{G}_{\mathcal{T}_0}} \ell(G),$$

where $\mathcal{T}_0$ is defined in (8).

The proof of Proposition 1 is relatively straightforward and may be found in the Appendix. Proposition 1 implies that, under the specified conditions, it suffices to consider distribution $G \in \mathcal{G}_{\mathcal{T}_0}$ supported on the reduced parameter space $\mathcal{T}_0$ in order to solve (2). The main practical implication is that in many examples, $\mathcal{T}_0$ is compact even when $\mathcal{T}$ is not; two examples are described below. However, before proceeding, we point out that $\mathcal{T}_0$ depends on the ordering of the parameters $\theta_1, ..., \theta_d$; for instance, if $d = 2$, then $\mathcal{T}_0$ may differ dramatically depending on whether $\theta = (\theta_1, \theta_2)$ or $\theta = (\theta_2, \theta_1)$. This provides some additional flexibility (and potential complexity) in identifying a convenient reduced parameter space.

Example 1. *Poisson-binomial model.* Suppose that $\Theta_j = (\lambda_j, \pi_j) \in \mathcal{T} = (0, \infty) \times (0, 1) \subseteq \mathbb{R}^2$ and that $X_j = (A_j, H_j)$, where

$$A_j|\Theta_j \sim \text{Poisson}(\lambda_j), \tag{9}$$

$$H_j|(A_j, \Theta_j) \sim \text{binomial}(A_j, \pi_j). \tag{10}$$

If one marginalizes over A_j, then $H_j|\Theta_j \sim \text{Poisson}(\pi_j\lambda_j)$ may be viewed as an observation from a thinned Poisson process (while A_j is an observation from the original Poisson process). This model will be encountered in Section 6 below, where A_j and H_j represent the number of at-bats and hits, respectively, for a Major League Baseball player and π_j represent's the player's underlying batting average.

The parameter space $\mathcal{T} = (0, \infty) \times (0, 1)$ is non-compact. On the other hand, the j-th observation likelihood is given by

$$\begin{aligned} f_0(X_j|\theta) &= f_0(A_j, H_j|\lambda, \pi) \\ &= \binom{A_j}{H_j} \pi_j^{H_j} (1 - \pi_j)^{A_j - H_j} \frac{\lambda^{A_j} e^{-\lambda}}{A_j!} \end{aligned} \tag{11}$$

and the univariate conditional MLEs for λ_j and π_j are $T_j^1(\lambda, \pi) = \hat{\lambda}_j = A_j$ and $T_j^2(\lambda, \pi) = \hat{\pi}_j = H_j/A_j$. Since $\hat{\lambda}_j$ is independent of π and since $\hat{\pi}_j$ is independent of λ, it follows that $\mathcal{T}_0 = [\hat{\lambda}_{(1)}, \hat{\lambda}_{(p)}] \times [\hat{\pi}_{(1)}, \hat{\pi}_{(p)}]$, where $\hat{\lambda}_{(1)} = T_{(1)}^1 = \min_{j=1,...,p} \hat{\lambda}_j$, $\hat{\lambda}_{(p)} = T_{(p)}^1 = \max_{j=1,...,p} \hat{\lambda}_j$, $\hat{\pi}_{(1)} = T_{(1)}^2 = \min_{j=1,...,p} \hat{\pi}_j$, and $\hat{\pi}_{(p)} = T_{(p)}^2 = \max_{j=1,...,p} \hat{\pi}_j$. Thus, $\mathcal{T}_0$ is compact. Moreover, in this example, the ordering of the parameters is insignificant; in particular, if we take $\theta = (\pi, \lambda)$ (as opposed to $\theta = (\lambda, \pi)$), then we simply have $\mathcal{T}_0 = [\hat{\pi}_{(1)}, \hat{\pi}_{(p)}] \times [\hat{\lambda}_{(1)}, \hat{\lambda}_{(p)}]$. □

Example 2. *Gaussian location-scale model.* Suppose that $\Theta_j = (\mu_j, \sigma_j^2) \in \mathcal{T} = \mathbb{R} \times (0, \infty) \subseteq \mathbb{R}^2$ and that $X_j = (X_{1j}, ..., X_{nj}) \in \mathbb{R}^n$, where

$$X_{1j}, ..., X_{nj}|\Theta_j \overset{\text{iid}}{\sim} N(\mu_j, \sigma_j^2). \tag{12}$$

Additionally assume that $n \geq 2$. Closely related models have been used in applications involving surveys with cluster-sampling (Xie et al., 2012) and DNA microarray classification problems (Dicker and Zhao, 2014). In cluster-sampling problems, p represents the number of clusters and $X_j = (X_{1j}, ..., X_{nj})$ is the n survey responses from each cluster (the model (12) can be trivially extended to handle unequal cluster sizes). In microarray classification problems, $(X_{i1}, ..., X_{ip}) \in \mathbb{R}^p$ represents a vector of p gene expressions from a DNA microarray experiment for the i-th individual in a study.

In this example, it is clear that the parameter space $\mathcal{T} = \mathbb{R} \times (0, \infty)$ is non-compact. The j-th observation likelihood is

$$\begin{aligned} f_0(X_j|\theta) &= f_0(X_{1j}, ..., X_{nj}|\mu, \sigma^2) \\ &= \sigma^{-n} \prod_{i=1}^{n} \phi\left(\frac{X_{ij} - \mu}{\sigma}\right), \end{aligned} \tag{13}$$

where $\phi(\cdot)$ is the standard normal density. Furthermore, the univariate conditional MLEs are $T_j^1(\theta) = \hat{\mu}_j = n^{-1}\sum_{i=1}^n X_{ij}$ and $T_j^2(\theta) = \hat{\sigma}_j^2(\mu) = n^{-1}\sum_{i=1}^n (X_{ij} - \hat{\mu}_j)^2 + (\hat{\mu}_j - \mu)^2$. It follows that $\mathcal{T}_0 = [\hat{\mu}_{(1)}, \hat{\mu}_{(p)}] \times [\hat{\sigma}^2_{(1)}, \hat{\sigma}^2_{(p)}]$, where $\hat{\mu}_{(1)} = T^1_{(1)} = \min_{j=1,...,p} \hat{\mu}_j$, $\hat{\mu}_{(p)} = T^1_{(p)} = \max_{j=1,...,p} \hat{\mu}_j$, $\hat{\sigma}^2_{(1)} = T^2_{(1)} = \min_{j,j'=1,...,p} \hat{\sigma}^2_j(\hat{\mu}_{j'})$, and $\hat{\sigma}^2_{(p)} = T^2_{(p)} = \max_{j,j'=1,...,p} \hat{\sigma}^2_j(\hat{\mu}_{j'})$. Notice that in this example, the univariate conditional MLE for σ^2, $\hat{\sigma}_j^2(\mu)$, depends on the parameter μ. Still, $\mathcal{T}_0$ is compact because we have chosen an appropriate ordering for the parameters, $\theta = (\mu, \sigma^2)$. If, on the other hand, we take $\theta = (\sigma^2, \mu)$, then $\mathcal{T}_0 = \mathcal{T}$ is non-compact. □

In each of the previous examples, it is easily seen that the conditions of Proposition 1 hold (for the Gaussian location-scale model, we require $n \geq 2$ so that the likelihoods are bounded). Since $\mathcal{T}_0$ is compact in each of these examples, Proposition 1 implies that we can take Λ in (4) to be a finite regular grid on $\mathcal{T}_0$. In general, if $\mathcal{T}_0$ is compact and $q_1, ..., q_d$ are positive integers, we define the regular grid $\Lambda_{q_1,...,q_d}$ as follows. For $m = 1, ..., d$, let $T^m_{(1)} = t^m_1 < \cdots < t^m_{q_m} = T^m_{(p)}$ be q_m equally-spaced points between $T^m_{(1)}$ and $T^m_{(p)}$. Then

$$\Lambda_{q_1,...,q_d} = \left\{ (t^1_{k_1}, ..., t^d_{k_d}); \begin{matrix} 1 \leq k_1 \leq q_1, ..., \\ 1 \leq k_d \leq q_d \end{matrix} \right\} \subseteq \mathcal{T}_0. \tag{14}$$

In all of the numerical experiments in this paper, we take $\Lambda = \Lambda_{q_1,...,q_d}$ with the indicated values of $q_1, ..., q_d$. Observe that with $\Lambda = \Lambda_{q_1,...,q_d}$, (6) is a $(q_1 \cdots q_d - 1)$-dimensional optimization problem.

3.2. *"F-modeling" vs. "G-modeling"*

The previous section provides some guidance on choosing Λ in the optimization problem (4) and in $\hat{G}_\Lambda$. In this section, we discuss classes of problems where it is likely that $\hat{G}_\Lambda$ can be used effectively in an empirical Bayes analysis.

By many common statistical metrics, accurately estimating the mixing distribution G_0 in (1) is very challenging. Estimating G_0 is a version of the mixture deconvolution problem, which often has extremely slow (logarithmic) convergence rates (e.g. Carroll and Hall, 1988). On the other hand, the objective in this paper is to perform posterior inference with $\hat{G}_\Lambda$ in place of G_0, rather than directly estimate G_0. While the effectiveness of this inference certainly depends on the quality of $\hat{G}_\Lambda$, it is not clear that it should necessarily suffer from the same poor rates that arise when attempting to estimate G_0.

Efron (2014) has argued that many mixture deconvolution problems can be categorized as "F-modeling" or "G-modeling" problems. Roughly speaking, in F-modeling problems, the task is to estimate some quantity that is determined by the marginal distribution of the observed data $X_j \sim F_0$ (i.e. the mixture distribution); on the other hand, G-modeling problems target quantities that are more fundamentally linked to the mixing distribution G_0. For example, estimating the density of G_0 (F_0) under integrated squared-error loss is a G (F)-modeling problem. Since mixing tends to smooth-out features of the marginal distribution, F-modeling problems are often substantially easier than G-modeling problems; theoretical results by Ghosal and Van der Vaart (2001), Brown and Greenshtein (2009), Jiang and Zhang (2009), and others have confirmed this in settings that involve empirical Bayes and NPMLEs.

It turns out that many interesting quantities that are useful for posterior inference under G_0 can be recast as F-modeling problems. For instance, the Tweedie rule (Efron, 2011; Robbins, 1956) implies that computing the posterior mean of $\mathbb{E}(\Theta_j|X_j)$ can be formulated as an F-modeling problem, when $\{F(\cdot|\theta)\}_{\theta\in\mathcal{T}}$ is an exponential family; additionally, the posterior likelihood ratio plays a key role in classification problems and, as formulated in Section 5.3 below, computing this likelihood ratio is essentially an F-modeling problem. In general, we expect that the nonparametric empirical Bayes approaches proposed in this paper will be most successful in problems that can be formulated as F-modeling problems. Identifying the extent to which various aspects of posterior inference under G_0 can be recast as F-modeling problems may be an interesting direction for future research.

4. Implementation: EM algorithm

A variety of methods are available for solving the optimization problem (6) with $\Lambda = \Lambda_{q_1,q_2}$ given in (14). All of the results reported in this paper are based on computations performed in R, using the EM algorithm (Jiang and Zhang, 2009; Laird, 1978). We emphasize that this is an especially simple (convex) version of the EM algorithm because the support points for G are fixed in Λ_{q_1,q_2}.

Before electing to use the EM algorithm in our experiments, we tested several other algorithms that have been suggested for finding $\hat{G}_\Lambda$, after adapting them to the present setting with multivariate G. We implemented the "vertex direction method," the "vertex exchange method" (Böhning, 1995), and an algorithm based on interior point methods. The vertex direction method is basically the Frank-Wolfe (1956) algorithm applied to the NPMLE problem (6) and the vertex exchange method may be viewed as a version of the Frank-Wolfe algorithm with away steps. For implementing the interior point methods, we used the R package `Rmosek`, following (Koenker and Mizera, 2014).

Overall, the different algorithms performed quite similarly in our preliminary tests, in terms of the key metrics for the numerical experiments described in Sections 5–6 below (detailed results from these tests are omitted in the interest of space). These key metrics were related to F-modeling objectives (see Section 3.2), such as estimation and classification. On the other hand, we found it noteworthy that the estimated mixing distributions $\hat{G}_\Lambda$ sometimes looked (visually) substantially different for the various algorithms; this *ad hoc* visual test seems much more aligned with a G-modeling objective. Computation times were also similar for the various algorithms in the settings that we considered (additional information on computation time for the EM algorithm is reported in Section 5.2 below). More significant differences could become apparent in larger-scale experiments and demand further study. Other algorithms, which have not been thoroughly studied in the NPMLE literature, may also be well-suited for solving (6) and finding $\hat{G}_\Lambda$ (e.g. entropic mirror descent; Beck and Teboulle, 2003).

5. Experiments with simulated data

5.1. Mixing distributions

Our experiments with simulated data are focused on the Gaussian location-scale model (12). We considered three different mixing distributions G_0 to highlight various features of the NPMLEs and our proposed methods.

Mixing Distribution 1: Sparse μ, dependent (μ, σ^2)

Mixing Distribution 1 is given by

$$G_0 : (\mu_j, \sigma_j^2) \sim 0.99\delta_{(0,s_0^2)} + 0.01\delta_{(1,s_1^2)},$$

with $s_0^2, s_1^2 > 0$ specified below. Observe that μ_j is sparse because $\Pr(\mu_j = 0) = 0.99$ and that μ_j and σ_j^2 are correlated (provided $s_0^2 \neq s_1^2$).

Mixing Distribution 2: Sparse μ, independent (μ, σ^2)

For Mixing Distribution 2, G_0 is chosen so that

$$\mu_j \sim 0.99\delta_0 + 0.01\delta_1,$$
$$\sigma_j^2 \sim 0.99\delta_{s_0^2} + 0.01\delta_{s_1^2}$$

independently, with $s_0^2, s_1^2 > 0$ specified below. The mean parameter is sparse because $\Pr(\mu_j = 0) = 0.99$. Note that the marginal distributions for μ_j and σ_j^2 are the same under Mixing Distributions 1 and 2.

Mixing Distribution 3: Non-sparse μ, dependent (μ, σ^2)

Mixing Distribution 3 is derived from the "baseball data," which has served as a popular benchmark for empirical Bayes methods (Brown, 2008; Jiang and Zhang, 2010; Koenker and Mizera, 2014; Muralidharan, 2010). The distribution is generated from the number of at-bats and hits for 486 Major League Baseball players (non-pitchers) during the first half of the 2005 season, by applying the variance-stabilizing transformation for a normal approximation to the binomial distribution. (More specifically, we applied the transformation (18), given below, to each of the non-pitchers in the baseball dataset and took G_0 to be the empirical distribution of the observations $\{(X_j, (4A_j)^{-1})\} \subseteq \mathbb{R}^2$.) In the baseball data, it is well-known that at-bats and batting average (hits per at-bat) are positively correlated; this translates to a negative correlation between μ_j and σ_j^2 under G_0. A two-dimensional histogram of Mixing Distribution 3 (G_0) may be found in Figure 1. Under this mixing distribution, $\mathbb{E}(\mu_j) = 0.53$, $\mathbb{E}(\sigma_j^2) = 0.0033$, and $\mathrm{Cor}(\mu_j, \sigma_j^2) = -0.46$. Additionally, it is reasonably clear from Figure 1 that μ_j is non-sparse under Mixing Distribution 3.

5.2. Estimation

We simulated 100 independent datasets from (12), with each of the three mixing distributions G_0 described in Section 5.1. For Mixing Distributions 1–2, we took $s_0^2 = 0.1$, $s_1^2 = 1$ and $n = 10$, $p = 1,000$; for Mixing Distribution 3, we took $n = 20$, $p = 500$. For each simulated dataset, we computed the NPMLE $\hat{G}_\Lambda$ with $\Lambda = \Lambda_{q_1,q_2}$ and $(q_1, q_2) = (20, 20), (30, 30), (100, 100)$.

The objective of this experiment was to illustrate the usefulness of $\hat{G}_\Lambda$ for estimation problems. After generating each dataset $\mathcal{D} = \{X_j\}$ and estimating $\hat{G}_\Lambda$, we hypothesized a single additional observation $X = \mu + \sigma Z$, with independent $(\mu, \sigma^2) \sim G_0$, $Z \sim N(0, 1)$, and attempted to estimate μ using various estimators $\hat{\mu} = \hat{\mu}(X)$. For each estimator $\hat{\mu}$, we recorded the mean squared error

$$\mathrm{MSE}(\hat{\mu}) = \mathbb{E}\left[\{\hat{\mu}(X) - \mu\}^2 | \mathcal{D}\right], \tag{15}$$

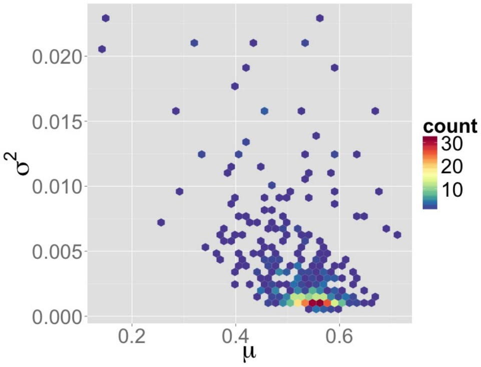

Fig 1. *Two-dimensional histogram for Mixing Distribution 3.*

where the expectation was computed numerically over μ, σ^2, and Z, conditional on the training data $\mathcal{D} = \{X_j\}$.

Overall, we considered six different estimators for μ. The *NPMLE estimator* that we propose in this paper is the posterior mean under $\hat{G}_\Lambda$,

$$\hat{\mu}(X) = \frac{\int \mu\sigma^{-1}\phi\left(\frac{X-\mu}{\sigma}\right)\ d\hat{G}_\Lambda(\mu,\sigma^2)}{\int \sigma^{-1}\phi\left(\frac{X-\mu}{\sigma}\right)\ d\hat{G}_\Lambda(\mu,\sigma^2)}. \tag{16}$$

Th NPMLE estimator is meant to approximate the *Bayes (optimal) estimator*, which is the posterior mean of μ under G_0 (i.e. replace $\hat{G}_\Lambda$ in (16) with G_0) and was also computed for each simulated dataset. Additionally, we computed the *naive estimator* $\hat{\mu}(X) = X$ and the *grand mean* $\hat{\mu}(X) = \bar{X} = (np)^{-1}\sum_{i,j} X_{ij}$. Observe that the naive estimator does not depend on the training data $\{X_{ij}\}$, while the grand mean does not depend on the testing data X. Finally, we computed a shrinkage estimator $\hat{\mu}(X) = (1+t)^{-1}(X-\bar{X})+\bar{X}$ and a soft-thresholding estimator $\hat{\mu}(X) = s_t(X-\bar{X})+\bar{X}$, where $t \geq 0$ is a constant and $s_t(x) = \text{sign}(x)\max\{|x|-t, 0\}$, $x \in \mathbb{R}$. For each of these estimators, t was chosen to minimize the average MSE; choosing t in this way is typically not possible in real applications, hence, we refer to these estimators as the *oracle shrinkage* and *oracle soft-thresholding* estimators. Shrinkage and soft-thresholding estimators have a long and illustrious history in statistics and nonparametric estimation problems (e.g. Donoho, 1995; Pinsker, 1980; Stein, 1956).

For each mixing distribution and each estimator $\hat{\mu}$, we calculated the average (mean) MSE over all 100 simulated datasets, relative to the average MSE of the Bayes optimal estimator. Results are reported in Table 1. In each setting, it is clear that the NPMLE estimator performs almost as well as the Bayes estimator and out-performs the other estimators that we have considered (substantially so for Mixing Distributions 1–2). Note that the naive estimator

TABLE 1

Average MSE *of various estimators for μ, based on 100 simulated datasets, relative to the average* MSE *of the Bayes (optimal) estimator. For NPMLE, (q_1, q_2) indicates the grid points used to fit $\hat{G}_\Lambda$, as in (14).*

	Mixing Dist. 1	Mixing Dist. 2	Mixing Dist. 3
Bayes	1	1	1
Naive	33.39	29.79	1.78
Grand mean	1.36	1.16	2.41
Oracle shrinkage	1.31	1.11	1.03
Oracle soft-thresh.	1.12	1.13	1.30
NPMLE			
$(q_1, q_2) = (20, 20)$	1.02	1.02	1.02
$(q_1, q_2) = (30, 30)$	1.01	1.02	1.01
$(q_1, q_2) = (100, 100)$	1.01	1.02	1.01

performs very poorly for Mixing Distributions 1–2 (the mixing distributions with sparse μ_j); this reflects the fact that the signal strength is very low in these simulation settings. It also seems noteworthy that the oracle shrinkage estimator performs very well under Mixing Distribution 3; almost as well as the NPMLE. Additionally, for the NPMLE, the MSEs are quite insensitive to (q_1, q_2), which determine the grid $\Lambda = \Lambda_{(q_1,q_2)}$. For Mixing Distributions 1–2, the average time to compute $\hat{G}_{\Lambda_{20,20}}$, $\hat{G}_{\Lambda_{30,30}}$, and $\hat{G}_{\Lambda_{100,100}}$ for each dataset was approximately 1 sec., 4 sec., and 2 min., respectively (on a MacBook Pro laptop with 2.8GHz Intel processor and 16GB RAM); for Mixing Distribution 3, the times were 0.2 sec., 0.6 sec., and 23 sec. In the sequel, we restrict our attention to $(q_1, q_2) = (30, 30)$ and take $\Lambda = \Lambda_{(30,30)}$ for all NPMLEs.

5.3. Classification

Dicker and Zhao (2014) showed how a univariate NPMLE for the Gaussian location model could be used in high-dimensional classification problems. In this section, we show how their method can be adapted to the Gaussian location-scale model with correlated mean and variance parameters, μ, σ^2.

We simulated training data from two groups, $\mathcal{D}^1 = \{X_j^1\}$ and $\mathcal{D}^2 = \{X_j^2\}$, with each X_j^k generated from the Gaussian location-scale model (12). In more detail, we fixed $p = 10,000$ and $n = 20$; then, for $k = 1, 2$ and $j = 1, ..., p$, we took $X_j^k = (X_{1j}^k, ..., X_{nj}^k)$, where

$$X_{ij}^k | (\mu_{jk}, \sigma_{jk}^2) \sim N(\mu_{jk}, \sigma_{jk}^2).$$

and $(\mu_{jk}, \sigma_{jk}^2)$ was drawn from Mixing Distribution k (described in Section 5.1) with $s_0^2 = 0.64$ and $s_1^2 = 1$. Moreover, μ_{j1} and μ_{j2} were generated so that they were correlated with $\Pr(\mu_{j1} =$

$1|\mu_{j2} = 1) = 0.9505$. We fit bivariate NPMLEs $\hat{G}^1$ and $\hat{G}^2$ for groups 1 and 2, separately, using the training data $\mathcal{D}^1$ and $\mathcal{D}^2$ (with grid $\Lambda = \Lambda_{30,30}$).

The classification task is to build a classifier that determines the group membership of a new observation $\mathbf{X}^{\text{test}} = (X_1^{\text{test}}, ..., X_p^{\text{test}})$, which is drawn from either group 1 or group 2, i.e. $X_j^{\text{test}} = \mu_{j1} + \sigma_{j1} Z_j^{\text{test}}$ or $\mu_{j2} + \sigma_{j2} Z_j^{\text{test}}$, with independent $Z_j^{\text{test}} \sim N(0,1)$. Our *bivariate NPMLE* classification rule is

$$\hat{\delta}(\mathbf{X}^{\text{test}}) = I\left\{\prod_{j=1}^{p} \frac{\phi \star \hat{G}_j^1(X_j^{\text{test}})}{\phi \star \hat{G}_j^2(X_j^{\text{test}})} < 1\right\},$$

where, for $k = 1, 2$, $\phi \star \hat{G}_j^k(x)$ is the posterior density

$$\phi \star \hat{G}_j^k(x) = \frac{\int \sigma^{-(n+1)} \phi\left(\frac{x-\mu}{\sigma}\right) \phi\left(\frac{\bar{X}_j^k - \mu}{\sigma/\sqrt{n}}\right) \, d\hat{G}^k(\mu, \sigma^2)}{\int \sigma^{-n} \phi\left(\frac{\bar{X}_j^k - \mu}{\sigma/\sqrt{n}}\right) \, d\hat{G}^k(\mu, \sigma^2)}$$

and $\bar{X}_j^k = n^{-1} \sum_{i=1}^n X_{ij}^k$. The NPMLE classifier $\hat{\delta}$ is an approximation to the Bayes classifier, which has G^k in place of $\hat{G}^k$ throughout.

Other classifiers that we tested in this experiment were: (i) The *univariate NPMLE* classifier of Dicker and Zhao (2014); (ii) another nonparametric empirical Bayes classifier proposed by Greenshtein and Park (2009), which uses nonparametric smoothing to fit a univariate density to the μ_j and then employs a version of linear discriminant analysis (referred to as *EBayes LDA*); (iii) a support vector machine (*SVM*) classifier with radial basis kernel; (iv) the *logistic lasso* fit with the `glmnet` R package; (v) a quadratic discriminant analysis (*QDA*) classifier (Hastie et al., 2009), fit under the assumption that $\text{Cov}(\mathbf{X}_i^k)$ is diagonal and $\mathbf{X}_i^k = (X_{i1}^k, ..., X_{ip}^k)$; and (vi) an ℓ^1-regularized version of linear discriminant analysis (*ℓ^1-regularized LDA*) proposed by Mai et al. (2012).

We calculated the misclassification rate for each classifier on a test dataset with $n = 20$ observations from each group. Mean misclassification rates computed over 100 independent datasets are reported in Table 2. The bivariate NPMLE dramatically outperforms the other classifiers considered in this experiment. Under these simulation settings, $(\mu_{11} - \mu_{12}, ..., \mu_{p1} - \mu_{p2})$ has only about 10 nonzero coordinates on average. However, the bivariate NPMLE apparently leverages the different correlation structures in Mixing Distribution 1 (μ_{j1}, σ_{j1}^2 positively correlated) and Mixing Distribution 2 (μ_{j2}, σ_{j2}^2 independent) to obtain a powerful classifier.

6. The baseball data

The baseball dataset contains the number of at-bats and hits for all of the Major League Baseball players during the 2005 season. The goal of the analysis is to use the data from

Table 2
Average misclassification rates for various classifiers.

	Misclassification Rate
SVM	0.42
Logistic lasso	0.24
QDA	0.37
ℓ^1-regularized LDA	0.24
EBayes LDA	0.27
Univariate NPMLE	0.33
Bivariate NPMLE	0.03

the first half of the season to predict each player's batting average (hits/at-bats) during the second half of the season. Overall, there are 929 players in the baseball dataset; however, following Brown (2008) and others, we restrict attention to the 567 players with more than 10 at-bats during the first half of the season (we follow the other preprocessing steps described in (Brown, 2008) as well).

Let A_j and H_j denote the number of at-bats and hits, respectively, for player j during the first half of the season. We assume that (A_j, H_j) follows the Poisson-binomial model (9)–(10) and propose to estimate each player's batting average for the second half of the season by the posterior mean of π, computed under $(\lambda, \pi) \sim \hat{G}_{\Lambda_{30,30}}$:

$$\hat{\pi}_j = \frac{\int \pi f_0(A_j, H_j|\lambda, \pi)\ d\hat{G}_{\Lambda_{30,30}}(\lambda, \pi)}{\int f_0(A_j, H_j|\lambda, \pi)\ d\hat{G}_{\Lambda_{30,30}}(\lambda, \pi)}, \tag{17}$$

where $f_0(A_j, H_j|\lambda, \pi)$ is defined in (11).

Most previously published analyses of the baseball data transform the data via the variance stabilizing transformation

$$X_j = \arcsin\sqrt{\frac{H_j + 1/4}{A_j + 1/2}}. \tag{18}$$

Under this transformation, X_j is approximately distributed as $N\{\mu_j, (4A_j)^{-1}\}$, where $\mu_j = \arcsin\sqrt{\pi_j}$. Methods for Gaussian observations may be applied to the transformed data, with the objective of estimating μ_j. Following this approach, a variety of methods based on shrinkage, the James-Stein estimator, and empirical Bayes methodology for Gaussian data have been proposed and studied (Brown, 2008; Jiang and Zhang, 2010; Xie et al., 2012).

Under the transformation (18), many authors (including those cited in the previous paragraph) have used the total squared error to measure the performance of estimators $\hat{\mu}_j$. The

Table 3

TSE *relative to that of the naive estimator. Minimum error estimator is in* **bold** *for each analysis.*

	All	Non- Pitchers	Pitchers
Naive	1	1	1
Grand mean	0.85	0.38	0.13
James-Stein	0.53	0.36	0.16
Weighted GMLE	0.30	0.26	0.14
Semiparametric SURE	0.41	0.26	**0.08**
Binomial mixture	0.59	0.31	0.16
Poisson-binomial NPMLE	**0.27**	**0.25**	0.13

total squared error is defined as

$$\mathrm{TSE} = \sum_j \left\{ (\hat{\mu}_j - \tilde{X}_j)^2 - \frac{1}{4\tilde{N}_j} \right\}^2,$$

where $\tilde{X}_j = \arcsin\sqrt{\frac{\tilde{H}_j + 1/4}{\tilde{A}_j + 1/2}}$, and $\tilde{A}_j$ and $\tilde{H}_j$ denote the at-bats and hits from the second half of the season, respectively. For convenience of comparison, we used TSE to measure the performance of our estimates $\hat{\pi}_j$, after applying the transformation $\hat{\mu}_j = \arcsin\sqrt{\hat{\pi}_j}$.

Results from the baseball analysis are reported in Table 3. Following the work of others, we have analyzed all players from the dataset together, and then the pitchers and non-pitchers from the dataset separately. In addition to our *Poisson-binomial NPMLE*-based estimators (17), we considered the following estimators.

1. The *naive* estimator: $\hat{\mu}_j = X_j$.
2. The *grand mean*: $\hat{\mu}_j = p^{-1}\sum_{k=1}^{p} X_k$.
3. The *James-Stein* estimator described in (Brown, 2008).
4. The weighted generalized MLE (*weighted GMLE*), which uses at-bats as a covariate (Jiang and Zhang, 2010). This is essentially a univariate NPMLE-method for Gaussian models with covariates.
5. The *semiparametric SURE* estimator (Xie et al., 2012) is a flexible shrinkage estimator that may be viewed as a generalization of the James-Stein estimator.
6. The *binomial mixture* method from (Muralidharan, 2010) is another empirical Bayes method, which does not require the data to be transformed and estimates π_j directly (Muralidharan works conditionally on the at-bats A_j). TSE is computed after applying the $\arcsin\sqrt{\cdot}$ transformation.

The values reported Table 3 are the TSEs of each estimator, relative to the TSE of the

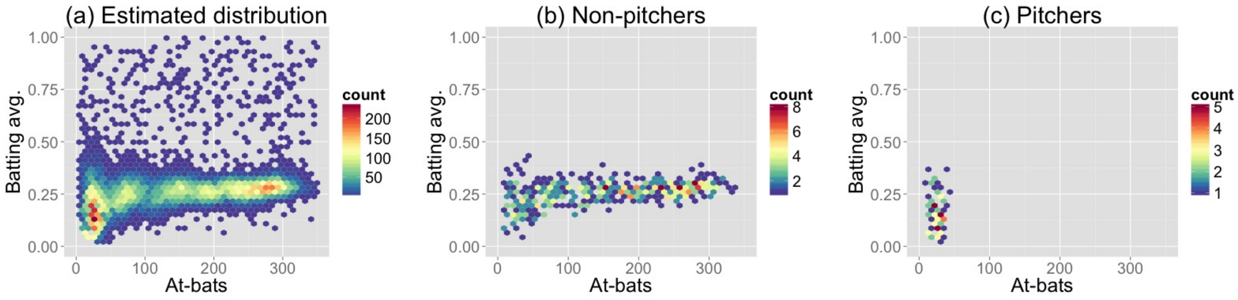

Fig 2. *(a) Histogram of 20,000 independent draws from the estimated distribution of* $(A_j, H_j/A_j)$*, fitted with the Poisson-binomial NPMLE to all players in the baseball dataset; (b) histogram of non-pitcher data from the baseball dataset; (c) histogram of pitcher data from the baseball dataset.*

naive estimator. Our Poisson-binomial method performs very well, recording the minimum TSE when all of the data (pitchers and non-pitchers) are analyzed together and for the non-pitchers. Moreover, the Poisson-binomial NPMLE $\hat{G}_{\Lambda_{30,30}}$ works on the original scale of the data (no normal tranformation is required) and may be useful for other purposes, beyond just estimation/prediction. Figure 2 (a) contains a histogram of 20,000 independent draws from the estimated distribution of $(A_j, H_j/A_j)$, fitted with the Poisson-binomial NPMLE to all players in the baseball dataset. Observe that the distribution appears to be bimodal. By comparing this histogram with histograms of the observed data from the non-pitchers and pitchers separately (Figure 2 (b)–(c)), it appears that the mode at the left of Figure 2 (a) represents a group of players which includes the pitchers and the mode at the right represents the bulk of the non-pitchers.

7. Discussion

We have proposed a flexible, practical approach to fitting arbitrary multivariate mixing distributions with NPMLEs via convex optimization. We have illustrated the effectiveness of this approach through a variety of numerical experiments and applications involving estimation and classification.

Maximum likelihood methods are natural candidates for estimating distributions nonparametrically in other interesting statistical models, beyond the mixture model (1). However, nonconvexity quickly becomes an obstacle. For instance, McAuliffe et al. (2006) estimated the base measure in a Dirichlet process mixture model by nonparametric smoothing. Nonparametric maximum likelihood would be another natural approach to estimating this distribution, but the associated NPML optimization problem is nonconvex. While this seems to be a significant hurdle, variational methods and other modern computational techniques might be useful for addressing these problems.

Appendix A: Proof of Proposition 1

Suppose that $G \in \mathcal{G}_{\mathcal{T}}$. To prove the proposition, we show that there is a probability measure $G' \in \mathcal{G}_{\mathcal{T}_0}$ such that

$$\ell(G') \leq \ell(G). \tag{19}$$

By observation (ii) from Section 2.1, we may assume without loss of generality that G is a discrete measure supported on $q \leq p$ points; more specifically, we assume that $G = \sum_{k=1}^{q} w_k \delta_{t_k}$, where $w = (w_1, ..., w_q) \in \Delta^{q-1}$ and $t_k = (t_{1k}, ..., t_{dk})$.

Let

$$\tilde{m} = \min\left\{m;\ t_{mk} \notin [T_{(1)}^{m}, T_{(N)}^{m}] \text{ for some } k = 1, ..., q\right\}.$$

Then $t_k \in \mathcal{T}^{m-1}$ for each $k = 1, ..., q$ (we take $\mathcal{T}^0 = \mathcal{T}$); in other words, the support of G is contained in $\mathcal{T}^{m-1}$. For $k = 1, ..., q$, define $\tilde{t}_k = (\tilde{t}_{1k}, ..., \tilde{t}_{dk}) \in \mathcal{T}$ so that $\tilde{t}_{mk} = t_{mk}$ for $m \neq \tilde{m}$ and

$$\tilde{t}_{\tilde{m}k} = \begin{cases} T_{(1)}^{\tilde{m}} & \text{if } t_{\tilde{m}k} < T_{(1)}^{\tilde{m}}, \\ t_{\tilde{m}k} & \text{if } T_{(1)}^{\tilde{m}} \leq t_{\tilde{m}k} \leq T_{(p)}^{\tilde{m}}, \\ T_{(p)}^{\tilde{m}} & \text{if } T_{(p)}^{\tilde{m}} < t_{\tilde{m}k}. \end{cases}$$

We claim that

$$f_0(X_j|t_k) \leq f_0(X_j|\tilde{t}_k), \tag{20}$$

for $j = 1, ..., p$, $k = 1, ..., q$. If $t_{\tilde{m}k} \in [T_{(1)}^{\tilde{m}}, T_{(p)}^{\tilde{m}}]$, then $t_k = \tilde{t}_k$ and (20) is obvious. Suppose now that $t_{\tilde{m}k} \notin [T_{(1)}^{\tilde{m}}, T_{(p)}^{\tilde{m}}]$. Then $\tilde{t}_{\tilde{m}k} = T_{\tilde{j}}^{\tilde{m}}(\theta)$ for some $\tilde{j} \in \{1, ..., p\}$ and $\theta \in \mathcal{T}^{\tilde{m}-1}$. Furthermore, since $L_j^m(\cdot|\theta)$ is unimodal,

$$f_0(X_j|t_k) = L_j^{\tilde{m}}(t_{\tilde{m}k}|t_k) \leq L_j^{\tilde{m}}(\tilde{t}_{\tilde{m}k}|t_k) = f_0(X_j|\tilde{t}_k).$$

It follows that (20) holds for all $j = 1, ..., p$ and $k = 1, ..., q$.

Now let $\tilde{G} = \sum_{k=1}^{q} w_k \delta_{\tilde{t}_k}$ and observe that the support of $\tilde{G}$ is contained in $\mathcal{T}^m$. By (20),

$$\begin{aligned} \ell(\tilde{G}) &= -\frac{1}{p}\sum_{j=1}^{p} \log\left\{\sum_{k=1}^{q} w_k L_j(\tilde{t}_k)\right\} \\ &\leq -\frac{1}{p}\sum_{j=1}^{p} \log\left\{\sum_{k=1}^{q} w_k L_j(t_k)\right\} = \ell(G). \end{aligned}$$

Thus, beginning with a probability measure G, supported on $\mathcal{T}^{m-1}$, we have obtained a probability measure $\tilde{G}$, supported on $\mathcal{T}^m$ such that $\ell(\tilde{G}) \leq \ell(G)$. Repeating this process finitely many times, we obtain a probability measure G', supported on $\mathcal{T}^d = \mathcal{T}_0$ (i.e. $G' \in \mathcal{G}_{\mathcal{T}_0}$) satisfying (19), as was to be shown.

References

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Stat.* **2** 1152–1174.

Beck, A. and Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.* **31** 167–175.

Blei, D. and Jordan, M. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **1** 121–143.

Böhning, D. (1995). A review of reliable maximum likelihood algorithms for semiparametric mixture models. *J. Stat. Plan. Infer.* **47** 5–28.

Böhning, D., Schlattmann, P. and Lindsay, B. (1992). Computer-assisted analysis of mixtures (CA MAN): Statistical algorithms. *Biometrics* **48** 283–303.

Brown, L. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *Ann. Appl. Stat.* **2** 113–152.

Brown, L. and Greenshtein, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Stat.* **37** 1685–1704.

Brown, L., Greenshtein, E. and Ritov, Y. (2013). The Poisson compound decision problem revisited. *J. Am. Stat. Assoc.* **108** 741–749.

Carroll, R. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Am. Stat. Assoc.* **83** 1184–1186.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability.* Springer.

Dicker, L. and Zhao, S. (2014). Nonparametric empirical Bayes and maximum likelihood estimation for high-dimensional data analysis. ArXiv preprint arXiv:1407.2635.

Donoho, D. (1995). De-noising by soft-thresholding. *IEEE T. Inform. Theory* **41** 613–627.

Donoho, D. and Reeves, G. (2013). Achieving Bayes MMSE performance in the sparse signal + Gaussian white noise model when the noise level is unknown. In *IEEE Int. Symp. Inf. Theory (ISIT).* IEEE.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Cambridge University Press.

Efron, B. (2011). Tweedie's formula and selection bias. *J. Am. Stat. Assoc.* **106** 1602–1614.

Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statist. Sci.* **29** 285–301.

Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Nav. Res. Log.* **3** 95–110.

Ghosal, S. and Van der Vaart, A. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Stat.* **29** 1233–1263.

Greenshtein, E. and Park, J. (2009). Application of non parametric empirical Bayes estimation to high dimensional classification. *J. Mach. Learn. Res.* **10** 1687–1704.

Gu, J. and Koenker, R. (2014). Unobserved heterogeneity in income dynamics: An empirical Bayes perspective. Cemmap working paper, doi:10.1920/wp.cem.2014.4314.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. Springer.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Stat.* **37** 1647–1684.

Jiang, W. and Zhang, C.-H. (2010). Empirical Bayes in-season prediction of baseball batting averages. In *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown.* Institute of Mathematical Statistics, 263–273.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.* **27** 887–906.

Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Am. Stat. Assoc.* **109** 674–685.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Am. Stat. Assoc.* **73** 805–811.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *Ann. Stat.* **20** 1222–1235.

Lesperance, M. and Kalbfleisch, J. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *J. Am. Stat. Assoc.* **87** 120–126.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry, and Applications.* Institute of Mathematical Statistics.

Mai, Q., Zou, H. and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99** 29–42.

Martin, R. (2009). *Fast nonparametric estimation of a mixing distribution with application to high-dimensional inference.* Ph.D. thesis, Purdue University.

Mauldin, R., Sudderth, W. and Williams, S. (1992). Polya trees and random distributions. *Ann. Stat.* **20** 1203–1221.

McAuliffe, J., Blei, D. and Jordan, M. (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Stat. Comput.* **16** 5–14.

McLachlan, G. and Peel, D. (2004). *Finite Mixture Models.* John Wiley & Sons.

Muralidharan, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *Ann. Appl. Stat.* **4** 422–438.

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.* **9** 249–265.

Pinsker, M. (1980). Optimal filtration of square-integrable functions in Gaussian noise. *Probl. Inf. Transm.* **16** 52–68.

Robbins, H. (1950). A generalization of the method of maximum likelihood: Estimating a mixing distribution (abstract). *Ann. Math. Stat.* **21** 314–315.

Robbins, H. (1956). The empirical Bayes approach to statistical decision problems. In *Proc. Third Berkeley Symp. on Math. Statist. and Prob.*, vol. 1.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. on Math. Statist. and Prob.*

Xie, X., Kou, S. and Brown, L. (2012). SURE estimates for a heteroscedastic hierarchical model. *J. Am. Stat. Assoc.* **107** 1465–1479.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods: Invited paper. *Ann. Stat.* **31** 379–390.