Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Approximate 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, United States

ARTICLE INFO

Article history:
Received 14 May 2017
Received in revised form 10 January 2018
Accepted 12 January 2018
Available online xxxx

Keywords:
Nonparametric maximum likelihood
Kiefer–Wolfowitz estimator
Multivariate mixture models
Convex optimization

ABSTRACT

Nonparametric maximum likelihood (NPML) for mixture models is a technique for estimating mixing distributions that has a long and rich history in statistics going back to the 1950s, and is closely related to empirical Bayes methods. Historically, NPML-based methods have been considered to be relatively impractical because of computational and theoretical obstacles. However, recent work focusing on approximate NPML methods suggests that these methods may have great promise for a variety of modern applications. Building on this recent work, a class of flexible, scalable, and easy to implement approximate NPML methods is studied for problems with multivariate mixing distributions. Concrete guidance on implementing these methods is provided, with theoretical and empirical support; topics covered include identifying the support set of the mixing distribution, and comparing algorithms (across a variety of metrics) for solving the simple convex optimization problem at the core of the approximate NPML problem. Additionally, three diverse real data applications are studied to illustrate the methods' performance: (i) A baseball data analysis (a classical example for empirical Bayes methods), (ii) high-dimensional microarray classification, and (iii) online prediction of blood-glucose density for diabetes patients. Among other things, the empirical results demonstrate the relative effectiveness of using multivariate (as opposed to univariate) mixing distributions for NPML-based approaches.

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 $\{F_0(\cdot \mid \theta)\}_{\theta\in\mathcal{T}}$ be a family of probability distributions on $\mathbb{R}^n$ indexed by the parameter $\theta \in \mathcal{T}$. Throughout the paper, we assume that $\mathcal{T}$ is closed and convex. Assume that $X_1, \ldots, X_p \in \mathbb{R}^n$ are observed iid random variables and that $\Theta_1, \ldots, \Theta_p \in \mathbb{R}^d$ are corresponding iid latent variables, which satisfy

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

In (1), it may be the case that $F_0(\cdot \mid \theta)$ and G_0 are both known (pre-specified) distributions; more frequently, this is not the case. In this paper, we will study problems where the mixing distribution G_0 is unknown, but will assume $F_0(\cdot \mid \theta)$ is known throughout. Problems like this arise in applications throughout statistics, and various solutions have been proposed.

* Correspondence to: 110 Frelinghuysen Rd, Piscataway, NJ 08854, United States.
E-mail addresses: long.feng@rutgers.edu (L. Feng), ldicker@stat.rutgers.edu (L.H. Dicker).

https://doi.org/10.1016/j.csda.2018.01.006
0167-9473/

The distribution G_0 can be modeled parametrically, which leads to hierarchical modeling and parametric empirical Bayes methods (e.g. Efron, 2010). Another approach is to model G_0 as a discrete distribution supported on finitely- or infinitely-many points; this leads to the study of finite mixture models or nonparametric Bayes, respectively (McLachlan and Peel, 2004; Ferguson, 1973). This paper focuses on another method for estimating G_0: Nonparametric maximum likelihood.

Nonparametric maximum likelihood (NPML) methods for mixture models – and closely related empirical Bayes methods – have been studied in statistics since the 1950s (Robbins, 1950; Kiefer and Wolfowitz, 1956; Robbins, 1956). They make virtually no assumptions on the mixing distribution G_0 and provide an elegant approach to problems like (1). The general strategy is to first find the nonparametric maximum likelihood estimator (NPMLE) for G_0, denoted by $\hat{G}$, then perform inference via empirical Bayes (Robbins, 1956; Efron, 2010); that is, inference in (1) is conducted via the posterior distribution $\Theta_j \mid X_j$, under the assumption $G_0 = \hat{G}$. Research into NPMLEs for mixture models has included work on algorithms for computing NPMLEs and theoretical work on their statistical properties (e.g. Laird, 1978; Böhning et al., 1992; Lindsay, 1995; Ghosal and van der Vaart, 2001; Jiang and Zhang, 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). In this paper, we study a computationally convenient approach involving approximate NPMLEs, which sidesteps many of these difficulties and is shown to be effective in a variety of applications.

Our approach is largely motivated by recent work initiated by Koenker and Mizera (2014)[1] and further pursued by others, including Gu and Koenker (2016, 2017a, b) and Dicker and Zhao, (2016). Koenker & Mizera studied convex approximations to NPMLEs for mixture models in relatively large-scale problems, with up to 100,000s of observations. In Koenker and Mizera (2014), they showed that for 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.

Koenker and Mizera (2014)'s focus on convexity and scalability is one of the key concepts for this paper. Here, we show how a simple convex approximation to the NPMLE can be used effectively in a broad range of problems with nonparametric mixture models; including problems involving (i) multivariate mixing distributions, (ii) discrete data, (iii) high-dimensional classification, and (iv) state–space models. Backed by new theoretical and empirical results, we provide concrete guidance for efficiently and reliably computing approximate multivariate NPMLEs. Our main theoretical result (Proposition 1) suggests a simple procedure for finding the support set of the estimated mixing distribution. Many of our empirical results highlight the benefits of using multivariate mixing distributions with correlated components (Sections 6.2, 7, and 8), as opposed to univariate mixing distributions, which have been the primary focus of previous research in this area (notable exceptions include theoretical work on the Gaussian location-scale model in Ghosal and van der Vaart (2001) and applications in Gu and Koenker (2017a, b) involving estimation problems with Gaussian models). In Sections 7–9, we illustrate the performance of the methods described here in real-data applications involving baseball, cancer microarray data, and online blood-glucose monitoring for diabetes patients. In comparison with other recent work on NPMLEs for mixture models, this paper distinguishes itself from Gu and Koenker (2016) in that it focuses on more practical aspects of fitting general multivariate NPMLEs. Additionally, in this paper we consider a substantially broader swath of applications than Gu and Koenker (2017a, b) and Dicker and Zhao (2016), where the focus is estimation in Gaussian models and classification with a univariate NPMLE, respectively, and show that the same fundamental ideas may be effectively applied in all of these settings.

2. NPMLEs for mixture models via convex optimization

2.1. NPMLEs

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

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

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

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

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

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:

[1] Koenker & Mizera's work was itself partially inspired by relatively recent theoretical work on NPMLEs by Jiang and Zhang (2009).

Please cite this article in press as: Feng L., Dicker L.H., Approximate nonparametric maximum likelihood for mixture models: A convex optimization approach to fitting arbitrary multivariate mixing distributions. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.01.006.