Computational Statistics and Data Analysis 54 (2010) 1930–1941

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A computational strategy for doubly smoothed MLE exemplified in the normal mixture model

Byungtae Seo [a,*], Bruce G. Lindsay [b]

[a] Department of Mathematics and Statistics, Texas Tech University, Broadway and Boston, Lubbock, TX 79409, USA
[b] Department of Statistics, Pennsylvania State University, 326 Thomas building, University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 14 May 2009
Received in revised form 26 February 2010
Accepted 26 February 2010
Available online 10 March 2010

Keywords:
DS-MLE
Normal mixture
Kernel smoothing
Monte Carlo
Bandwidth selection
Spectral degrees of freedom

ABSTRACT

A typical problem for the parameter estimation in normal mixture models is an unbounded likelihood and the presence of many spurious local maxima. To resolve this problem, we apply the doubly smoothed maximum likelihood estimator (DS-MLE) proposed by Seo and Lindsay (in preparation). We discuss the computational issues of the DS-MLE and propose a simulation-based DS-MLE using Monte Carlo methods as a general computational tool. Simulation results show that the DS-MLE is virtually consistent for any bandwidth choice. Moreover, the parameter estimates in the DS-MLE are quite robust to the choice of bandwidths, as the theory indicates. A new method for the bandwidth selection is also proposed.

1. Introduction

Seo and Lindsay (in preparation) introduced a new method for doing likelihood-style inference in models with problematic likelihoods. Their new method, which they called doubly smoothed maximum likelihood estimator (DS-MLE), was shown to be universally consistent in the random sampling framework and highly efficient in nonparametric models. This companion paper has two main themes: First, we wish to provide some general guidance on how the estimators can be calculated. Secondly, we wish to investigate their performance in an important parametric model.

The problematic likelihood we consider is the normal mixture model. In a finite mixture of location-scale distributions, it is well known that the likelihood is not bounded (Kiefer and Wolfowitz, 1956). In such cases, the likelihood diverges as a scale parameter goes to zero at a certain location parameter value. As an example, let us consider the two-component normal mixture density,

$$\frac{p}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right)+\frac{1-p}{\sigma_2}\phi\left(\frac{x-\mu_2}{\sigma_2}\right),$$

where $\phi(\cdot)$ is the standard normal density and p is the mixing proportion of the first component. The likelihood based on IID sample $X_1,\ldots,X_n$ is then

$$L(\theta;\boldsymbol{x})=\prod_i\left[\frac{p}{\sigma_1}\phi\left(\frac{x_i-\mu_1}{\sigma_1}\right)+\frac{1-p}{\sigma_2}\phi\left(\frac{x_i-\mu_2}{\sigma_2}\right)\right]. \tag{1.1}$$

* Corresponding author. Tel.: +1 806 742 2580; fax: +1 806 742 1112.
E-mail address: byungtae.seo@ttu.edu (B. Seo).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2010.02.026

If we let $\mu_2 = x_1$ and let σ_2 go to zero, (1.1) goes to infinity. Consequently, the MLE of parameters on the parameter space

$$\Omega = \left\{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p) : -\infty < \mu_1, \mu_2 < \infty, \sigma_1^2 > 0, \sigma_2^2 > 0, 0 < p < 1\right\}$$

always occurs on the boundary of Ω and this MLE is clearly inconsistent.

This infinite likelihood problem has been known since Kiefer and Wolfowitz (1956). It has been studied for many years, and a variety of solutions proposed. Our goal in this paper is to show that the completely general DS-MLE methodology solves this problem, and does so in a manner competitive with other methods that were specifically designed for the normal mixture.

For the unbounded likelihood problem in normal mixture problems, Hathaway (1985) suggested the constrained MLE that arises from restricting the parameter space to

$$\Omega' = \Omega \cap \left\{(\sigma_1^2, \sigma_2^2) : \min\left(\frac{\sigma_1^2}{\sigma_2^2}, \frac{\sigma_2^2}{\sigma_1^2}\right) \geq c > 0\right\}$$

for a fixed $c > 0$. In a similar vein, Tanaka and Takemura (2006) suggested the sequentially constrained MLE by letting $c = c_n$ go to zero with a rate e^{-n^d}, $0 < d < 1$. This simple amendment enables us to exclude singular solutions if a constant c or sequence c_n is appropriately selected. However, the proper choice of c or c_n in a finite sample raises yet further problems. If too large c is selected, we might exclude the true parameters and if not chosen large enough, the maximum can occur on the boundary of Ω'.

Ridolfi and Idier (2000) proposed a penalized likelihood which gives a penalty for small scale parameters to avoid infinite spikes. The consistency of the penalized MLE was studied in Ciuperca et al. (2003), Chen et al. (2008), Chen and Tan (2009). This can be also seen as Bayesian methods using prior information for the scale parameters (Marin et al., 2005), as the penalty term plays role of a prior law. Although the penalized likelihood method works well with properly chosen penalty functions, the choice of a penalty function in a finite sample is still problem. We will show that our methods compete favorably with penalized likelihood in a simulation study.

The sequentially constrained and penalized MLE modify either the parameter space or the likelihood function. In both methods, the effect of this modification is designed to disappear as a sample size increases. This design enables us to prove the strong consistency of both estimators. However, in a finite sample study, this modification could lead to nonignorable biases.

As a general tool to minimize this bias caused by artificial modification, Seo and Lindsay (in preparation) recently proposed the *doubly smoothed maximum likelihood estimator* (DS-MLE) which can produce consistent estimators for an arbitrary estimation problem. Although this method still requires bandwidth selection, they showed that the DS-MLE is consistent with any fixed bandwidth. This implies that the DS-MLE is robust to the choice of bandwidths even with a small sample.

In this article, we propose a general computational strategy to find the DS-MLE for any given model. Then we discuss how one can implement their methodology in practice with normal mixture examples. This paper is organized as follows: We first briefly review the DS-MLE in Section 2. Then we propose a computational strategy for the DS-MLE in Section 3 and bandwidth selection problem is discussed in Section 4. In Section 5, we present some simulation studies and a real data analysis, and then a further discussion is given in Section 6.

2. The DS-MLE

In this section, we briefly summarize results from Seo and Lindsay (in preparation). In general, the inconsistency of the MLE is caused by one or several violations of regularity conditions on a given model. In this case, one might want to modify the given model to be regular. However, there will be multiple choices to modify the model and this would be another practical problem. Moreover, any artificial modification could distort the scientific meaning of the parameters in the model.

To make any irregular estimation problem regular, Seo and Lindsay (in preparation) proposed a kernel smoothing method on both data and model. That is, they used a kernel smoothed model instead of the original model so that the usual regularity conditions are satisfied. Since this could cause the distortion of the original estimation problem, they also used kernel smoothed data instead of raw data.

To explain this, suppose $X_1, \ldots, X_n$ is a random sample from an unknown probability measure $M_{\theta_\tau}(x)$ on $\mathbb{R}^d$. The smoothed model density from $M_\theta(x)$ is constructed as

$$m^*(t; M_\theta) = \int K_h(x, t)\mathrm{d}M_\theta(x), \tag{2.1}$$

where K_h is a kernel with bandwidth h. Then, using the same kernel and bandwidth, the smoothed kernel density is defined as

$$\hat{f}_n^*(t) = \int K_h(x, t)\mathrm{d}\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i, t),$$

where $\hat{F}_n(x)$ is the empirical distribution based on a given sample.