ARTICLE IN PRESS

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

Addressing overfitting and underfitting in Gaussian model-based clustering

Jeffrey L. Andrews

Department of Statistics, Irving K. Barber School of Arts and Sciences, University of British Columbia, Okanagan Campus, 1177 Research Rd, Kelowna, British Columbia, V1V 1V7, Canada

HIGHLIGHTS

- Overfitting and underfitting (convergence to local maxima) are illustrated using the EM algorithm for model-based clustering.
- A nonparametric bootstrap augmented EM-style algorithm is proposed and contrasted with other approaches.
- It is shown in both simulations and real applications to simultaneously address both overfitting and underfitting.

ARTICLE INFO

Article history:
Received 31 March 2017
Received in revised form 17 May 2018
Accepted 18 May 2018
Available online xxxx

Keywords:
EM algorithm
Bootstrap
Cluster analysis
Mixture models

ABSTRACT

The expectation–maximization (EM) algorithm is a common approach for parameter estimation in the context of cluster analysis using finite mixture models. This approach suffers from the well-known issue of convergence to local maxima, but also the less obvious problem of overfitting. These combined, and competing, concerns are illustrated through simulation and then addressed by introducing an algorithm that augments the traditional EM with the nonparametric bootstrap. Further simulations and applications to real data lend support for the usage of this bootstrap augmented EM-style algorithm to avoid both overfitting and local maxima.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Finite mixture models (cf. McLachlan and Peel, 2004) are often used as an attractive approach to cluster analysis, providing a wholly statistical handling of the unsupervised classification problem (cf. McNicholas, 2016a). Recent advances within the field of model-based clustering can be found by consulting (McNicholas, 2016b), while Bouveyron and Brunet-Saumard (2014) provide overview of the state-of-the-art in a high-dimensional context. A random vector $\mathbf{X}$ can be said to arise from a parametric finite mixture model if its probability density function is of the form $f(\mathbf{x}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x} \mid \Theta_g)$ where G is the number of mixture components, π_g are positive mixing proportions such that $\sum_{g=1}^{G} \pi_g = 1$, and $f_g(\mathbf{x} \mid \Theta_g)$ is a parametric density function with parameter space Θ_g. Notably, finite mixture models provide a probabilistic view of clustering, wherein observations are assigned a probability of having arisen from each component via the conditional expectation of their membership given the data and parameters. Specifically, a component indicator variable Z_{ig} is introduced such that $Z_{ig} = 1$ if observation i arises from component g and 0 otherwise — in the context of clustering, these Z_{ig} are missing data. Assuming the true model $f(\mathbf{x})$ is known with multivariate densities $f_1(\cdot), \ldots, f_G(\cdot)$ and parameter space $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_G, \Theta_1, \ldots, \Theta_G)$, then for any observation vector $\mathbf{x}_i$, we can find

$$\mathbb{E}[Z_{ig} \mid \mathbf{x}_i, \boldsymbol{\vartheta}] = \frac{\pi_g f_g(\mathbf{x}_i \mid \Theta_g)}{\sum_{j=1}^{G} \pi_j f_j(\mathbf{x}_i \mid \Theta_j)}. \tag{1}$$

E-mail address: jeff.andrews@ubc.ca.

https://doi.org/10.1016/j.csda.2018.05.015
0167-9473/© 2018 Elsevier B.V. All rights reserved.

UNCORRECTED PROOF

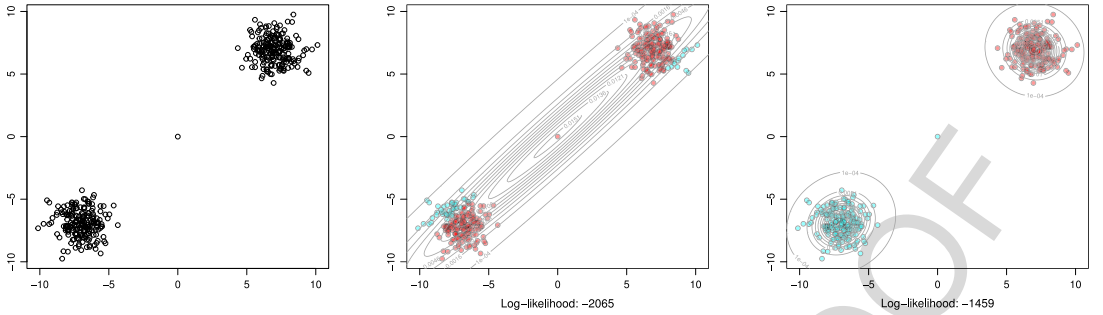

Fig. 1. From left to right: (A.) Mirror image groups with a central outlier (B.) Common solution from random z_{ig} initialization (C.) Common solution from 'good' z_{ig} and parameter initializations.

Unfortunately, in realistic applications, nearly all elements of $f(\mathbf{x})$ are unknown — specifically, the component densities $f_g(\cdot)$ are generally unknown and any associated free parameters from $\boldsymbol{\vartheta}$ require estimation. Often, the multivariate densities $f_g(\cdot)$ are taken to be from the same parametric family, with the multivariate Gaussian being a common choice (cf. Banfield and Raftery, 1993; Celeux and Govaert, 1995; Fraley and Raftery, 2002; McNicholas and Murphy, 2008). Herein, we will focus on the unconstrained Gaussian model, where the parameter space is $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G)$ with $\boldsymbol{\mu}_g$ defined as the mean vector for group g and $\boldsymbol{\Sigma}_g$ defined as the covariance matrix for group g — all $\boldsymbol{\vartheta}$ unknown and requiring estimation. The most common approach for model fitting is the expectation–maximization (EM) algorithm (cf. Dempster et al., 1977; McLachlan and Krishnan, 2008) which suffers drawbacks such as converging to local optima and even local minima (Titterington et al., 1985; McLachlan and Krishnan, 2008) — the EM algorithm will be further detailed in Section 2.1 of this manuscript. Many have approached these issues by seeking 'good' initializations (Biernacki et al., 2003; Karlis and Xekalaki, 2003), imposing constraints on the EM (Ingrassia, 2004; Greselin and Ingrassia, 2010), exploring alternative estimation techniques that rely on differing optimization strategies (Martınez and Vitria, 2000; Pernkopf and Bouchaffra, 2005; Andrews and McNicholas, 2013), or implementing techniques arising from the Bayesian paradigm (Robert and Titterington, 1998; Attias, 1999; Frühwirth-Schnatter, 2006; McGrory and Titterington, 2007).

While achieving the global maximum likelihood parameter space for a fitted model is an important task to work towards, it is essential to recognize that achieving the global maxima while fitting a model to observed data will generally result in some amount of overfitting. This fact is sometimes overlooked in unsupervised scenarios, and has large ramifications in the context of mixture models because it calls into question the validity of the expected values calculated from Eq. (1). To illustrate this point, we provide a toy simulation by generating one bivariate Gaussian cluster centred at $(7, 7)$, introducing a mirror image of that cluster reflected about $Y = -X$, and placing a point directly at the origin (see Fig. 1A). While this simulation is wholly unrealistic, it could be expected that a Gaussian mixture model should give a satisfactory fit to it. It turns out that two problems arise when fitting an unconstrained Gaussian mixture model with two components in this scenario. Firstly, if provided random initialization via cluster memberships, the solution provided in Fig. 1B is one of several stable local maxima that can be found due to the exact mirror image nature of the data. While randomly generated, or real, data is unlikely to exhibit this particular model-fitting behaviour, this still illustrates the complexity of model-fitting in the context of the EM algorithm. Secondly, provided a 'good' initialization, such as providing an initialization from another clustering algorithm (k-means is used for this example), a solution similar to Fig. 1C is found. While the contours look reasonable to the naked eye, the centre point has its expected cluster memberships calculated as

$$\mathbb{E}[Z_{01} \mid \boldsymbol{\vartheta}] = .9999999706 \quad \text{and} \quad \mathbb{E}[Z_{02} \mid \boldsymbol{\vartheta}] = .0000000294.$$

Keeping in mind that the data is specifically manufactured such that the centre point is exactly equidistant to two mirrored groups, this is a grossly high probability of membership to the first group.

The estimated log-likelihood value for the solution from Fig. 1B is approximately -2065. For the solution in Fig. 1C, the log-likelihood is computed as approximately -1459. And yet, the 'appropriate' solution (with the caveat that we are restricted to using a two-group Gaussian mixture model), wherein the point at the origin is given equal probabilistic classification into both groups, has a computed log-likelihood of approximately -1462. This solution is, in fact, a local maxima, however it is a maxima that is practically impossible to find through standard initializations of the EM algorithm — one has to initialize the EM very close to the solution in order to converge there. And so, we have illustrated a paradox for this toy data — we do not actually want to find the global maxima (leads to overfitting), but we also do not want the vast majority of local maxima (lead to underfitting).

In the sections that follow, we introduce and apply a model-fitting approach that combines the EM algorithm with the nonparametric bootstrap, and in doing so addresses this paradox. We begin by detailing background material and previous work in Section 2. Then, in Section 3 we introduce methodology and pseudocode for the proposed model-fitting algorithm. Next, in Section 4 we revisit the simulation from Fig. 1 and show, via further simulations and real applications, the effectiveness of the new algorithm. Finally, Section 5 concludes the manuscript with a summary and comments on future avenues for this work.

UNCORRECTED PROOF

Please cite this article in press as: Andrews J.L., Addressing overfitting and underfitting in Gaussian model-based clustering. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.05.015.

Download English Version:

https://daneshyari.com/en/article/6868610

Download Persian Version:

https://daneshyari.com/article/6868610

Daneshyari.com