Pattern Recognition Letters 69 (2016) 22–27

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

# Improved Bayesian information criterion for mixture model selection☆

Arash Mehrjou[a,b,*], Reshad Hosseini[a,**], Babak Nadjar Araabi[a,b]

[a] Control and Intelligent Processing Center of Excellence, School of ECE, College of Engineering, University of Tehran, Tehran, Iran
[b] School of Cognitive Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

ARTICLE INFO

*Article history:*
Received 21 January 2015
Available online 23 October 2015

*Keywords:*
Model selection
Bayesian method
Minimum message length
Finite mixture models
Laplace approximation
Clustering

ABSTRACT

In this paper, we propose a mixture model selection criterion obtained from the Laplace approximation of marginal likelihood. Our approximation to the marginal likelihood is more accurate than Bayesian information criterion (BIC), especially for small sample size. We show experimentally that our criterion works as good as other well-known criteria like BIC and minimum message length (MML) for large sample size and significantly outperforms them when fewer data points are available.



## 1. Introduction

Mixture modeling is a powerful statistical technique for unsupervised density estimation especially for high-dimensional data. Because of its usefulness as an extremely flexible method of modeling, finite mixture models have received increased attention over years with applications on pattern recognition, computer vision, signal and image processing, machine learning, and so forth [4,17,20,23].

Mixture models divide the entire space of data to several regions and each region is modeled by a probability density which is usually chosen from a class of similar parametric distributions. In mixture density estimation this division is soft, meaning that each datum may belong to several components [15]. In clustering applications the division is hard, that is each datum is assigned to only one cluster. Therefore, mixture models are applicable to clustering applications through probabilistic model-based approaches [14,20,21].

An important issue in mixture modeling is the selection of the number of mixture components [20]. Too many components may over-fit the observations, meaning that it can fit the training data accurately but it may not be a good model for underlying data-generating process. On the other hand, too few components may not be flexible enough to approximate the underlying model.

Different approaches have been proposed in the literature for determining the number of mixture components [21]. Some criteria select the number of components based on the generalization performance of the model. This is done either by having a separate validation set for testing performance [29], or by deriving asymptotic bias for goodness-of-fit as done in AIC measure [1]. Some other criteria like BIC use Bayesian framework in model selection, meaning that they try to find a model that has the maximum posterior probability or maximum marginal likelihood under some regularity conditions [11,28].

*Integrated Complete Likelihood* (ICL) is another criterion that like BIC approximates the marginal likelihood. ICL performs poorly when mixture components overlap and tends to underestimate the number of components [10]. As predicted from the theory behind ICL and shown in experiments, this criterion works well for the cases where each datum is assigned to only one cluster, that is in clustering applications [2].

Mixture model selection has remained a topic of active research until recently. Xie et al. [33] used an adaptive method that investigates the stability of log characteristic function versus number of components to find the true model. Their method proved to be suitable for large sample sizes. Zeng and Cheung [34] suggested a model-based clustering algorithm which uses a modified version of MML for model selection. Their method is tailored to clustering applications and also requires sufficiently large sample sizes. Maugis and Michel [19] suggested a non-asymptotic penalized criterion for mixture model selection. Their method needs compute a quantity named *bracketing entropy* which is not easily obtainable for non-Gaussian components.

We propose a criterion for determining the number of components of mixture models that is based on Laplace approximation to the marginal likelihood. BIC criterion can be also viewed as the

☆ This paper has been recommended for acceptance by Egon L. van den Broek.

* Corresponding author. Tel.: +98 936 122 8008.
** Corresponding author. Tel.: +98 21 6111 9799.
*E-mail addresses:* a.mehrjou@ut.ac.ir, mehrjou.arash@gmail.com, a_mehrjou@yahoo.com (A. Mehrjou), reshad.hosseini@ut.ac.ir (R. Hosseini).

http://dx.doi.org/10.1016/j.patrec.2015.10.004
0167-8655/

asymptotic Laplace approximation neglecting the terms that does not grow with the number of components. Instead of neglecting those terms, we assume the components are well-separated and derive a different approximation. It turns out that our approximation works better in many simulations.

We summarize the contributions of this paper as follows:

1. Under the assumption of non-overlapping components, we derive a new model selection criterion.
2. We show through different experiments that violating well-separateness assumption of components does not have a detrimental effect on the performance. Our criterion actually works significantly better than other criteria even for the case of overlapping components.

The rest of the paper is organized as follows: We first review the concepts related to mixture model parameter estimation in Section 2. Popular model selection techniques are summarized in Section 3. We describe the derivation of our proposed method in Section 4. The simulation results and comparisons are given in Section 5. Finally, we finish the paper by a short conclusion and envisioning future directions of research in Section 6.

## 2. Mixture models

In this section, we discuss some basics of mixture models, for in depth treatment of this subject, see [20,21]. The density of a mixture of $K$ components assumes the form

$$\forall \mathbf{x} \in \mathbb{R}^c, \quad p(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{m=1}^{K} \pi_m p(\mathbf{x}|\boldsymbol{\theta}_m), \tag{1}$$

where $\pi_1, \ldots, \pi_K$ are the mixing probabilities coming from the $K$-dimensional probability simplex, i.e. $\sum_{m=1}^{K} \pi_m = 1$, and $\boldsymbol{\theta}_m$ is the set of parameters defining $m$th component. The variable $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \pi_1, \ldots, \pi_K\}$ indicates the complete set of parameters of the mixture model.

Let $\mathcal{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ be the set of $n$ i.i.d samples from the underlying distribution, the log-likelihood over this set is given by:

$$\log p(\mathcal{X}|\boldsymbol{\Theta}) = \log \prod_{i=1}^{n} p(\mathbf{x}^{(i)}|\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log \sum_{m=1}^{K} \pi_m p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_m)\,.$$

The set of parameters that maximizes the log-likelihood function is called *maximum likelihood* (ML) estimate and is given by

$$\hat{\boldsymbol{\Theta}}_{\text{ML}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \{\log p(\mathcal{X}|\boldsymbol{\Theta})\}\,. \tag{2}$$

Assuming a prior $p(\boldsymbol{\Theta})$ over the parameter set and maximizing the posterior likelihood over the parameters results in *maximum a posteriori* (MAP) estimate which takes the form

$$\hat{\boldsymbol{\Theta}}_{\text{MAP}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \{\log p(\mathcal{X}|\boldsymbol{\Theta}) + \log p(\boldsymbol{\Theta})\}\,. \tag{3}$$

### 2.1. Maximum likelihood solution

The common procedure for solving the optimization problems in (2) and (3) is expectation-maximization (EM) algorithm [20,24]. Despite its fast convergence, simple EM suffers from one main drawback and it is the problem of converging to a local maximum. We observed that simple strategies like *multiple initialization* are not able to solve the local maxima problem. To this end, we implemented the well-known split and merge EM (SMEM) algorithm of [31] that nicely addresses this problem.[1]

[1] The toolbox developed by our group can be downloaded from http://visionlab.ut.ac.ir/mixest .

ML estimation procedure does not return reasonable estimate of the parameters when the log-likelihood of mixture model is unbounded. Intuitively, this happens when one component gets small number of data but its log-likelihood becomes infinite. Several possible remedies have been proposed in the literature [6,12]. One of the simplest and most powerful of them is using a suitable prior on the parameter space, that is estimating MAP instead of ML [11]. Accordingly, in all of our experimental results, the parameters of mixture models are estimated using MAP estimator.

## 3. Previous model selection criteria

The aim of model selection is selecting a model in the hypothesis space that best describes the underlying distribution of the observed data. For the case of mixture models, each hypothesis corresponds to a mixture with specific number of components. There are two main classes of model selection procedures: deterministic and stochastic. Deterministic methods are commonly used for determining the number of components in mixture models and are the main focus of the current paper.

### 3.1. Deterministic methods

Given a set of models in the hypothesis space, deterministic procedures select the model with the optimum information criterion. Information criterion is a function of data log-likelihood at the ML solution and the model complexity represented as

$$\mathcal{IC}(\hat{\boldsymbol{\Theta}}, \mathcal{X}) = -\alpha \log p\big(\mathcal{X}|\hat{\boldsymbol{\Theta}}\big) + \beta\, \mathcal{F}(\hat{\boldsymbol{\Theta}}),$$

where function $\mathcal{F}(\hat{\boldsymbol{\Theta}})$ represents model complexity and is independent of the observed data. Also, $\alpha, \beta \geq 0$ are the weights determining the influence for each of these opposing terms. Normally, in the case of mixture models $\mathcal{F}(\hat{\boldsymbol{\Theta}})$ increases by increasing the number of components penalizing complex mixture models with more components.

*3.1.1. Akaike Information Criterion.* Let $q(.)$ be the true data-generating density and $p(.|\boldsymbol{\Theta})$ be a parametric model density. The expected bias between the log-likelihood of the training data evaluated at ML solution and expected logarithm of the model density evaluated at its maximum $\boldsymbol{\Theta}_0$ is written as

$$\mathrm{E}_q\Big[\, \log p\big(\mathcal{X}|\hat{\boldsymbol{\Theta}}\big) - \mathrm{E}_q\big(\log p\big(\mathbf{x}|\boldsymbol{\Theta}_0\big)\big)\Big]\,,$$

and can be used as a measure of complexity. If the number of data points go to infinity, Akaike [1] derived an analytic form for this bias. He proved that the bias term is equal to the number of parameters. Thus, the bias corrected log-likelihood called *Akaike Information Criterion* (AIC) which is defined as

$$\text{AIC} = -2[\log p(\mathcal{X}|\hat{\boldsymbol{\Theta}}) - b] = -2 \log p(\mathcal{X}|\hat{\boldsymbol{\Theta}}) + 2\,d\,,$$

can be used for model selection. Here, $d$ is the dimensionality of the parameter space.

*3.1.2. Corrected Akaike Information Criterion.* The bias used in AIC is not accurate in the case of finite number of data points. However, in practice it has been used and has been an effective criterion for model selection. The only special case that the bias can be calculated analytically for finite sample size is the linear regression model. In this case, the corrected version of the information criterion becomes

$$\text{AIC}_c = -2 \log p(\mathcal{X}|\hat{\boldsymbol{\Theta}}) + 2\frac{n}{n-d-1}d$$

that takes the number of data points $n$ into account as proposed in [13].