Applied Soft Computing 38 (2016) 373–383

Contents lists available at ScienceDirect

Applied Soft Computing

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

A Bayesian analysis of spherical pattern based on finite Langevin mixture

Ola Amayri, Nizar Bouguila*

Faculty of Engineering and Computer Science, Concordia University, Montreal, Qc, Canada H3G 2W1

ARTICLE INFO

Article history:
Received 11 February 2014
Received in revised form 1 October 2015
Accepted 12 October 2015
Available online 19 October 2015

Keywords:
Langevin mixture
Bayesian inference
MCMC
Spherical data
Topic detection and tracking
Image categorization

ABSTRACT

Parameter estimation is a cornerstone of most fundamental problems of statistical research and practice. In particular, finite mixture models have long been heavily relied on deterministic approaches such as expectation maximization (EM). Despite their successful utilization in wide spectrum of areas, they have inclined to converge to local solutions. An alternative approach is the adoption of Bayesian inference that naturally addresses data uncertainty while ensuring good generalization. To this end, in this paper we propose a fully Bayesian approach for Langevin mixture model estimation and selection via MCMC algorithm based on Gibbs sampler, Metropolis–Hastings and Bayes factors. We demonstrate the effectiveness and the merits of the proposed learning framework through synthetic data and challenging applications involving topic detection and tracking and image categorization.

1. Introduction

As various disciplines have witnessed integration of digital technologies, high-dimensional sparse data is becoming more prevalent in every field of human endeavor. In the particular case of machine learning, such problems have been tackled using statistical learning, providing a rich and flexible techniques that can be applied to model data randomness and uncertainty. In this context, often one tries to understand this mass of data through analyzing informative patterns and describing the best possible model which succeeds in capturing the regularities in the data generating process. Nonetheless selecting an appropriate model that solves all aspects of application at hand is a major challenge as different approaches are needed, for distinct aspects, often depend on different representational choices. For instance, although modeling based on Gaussian mixtures [1–4] has provided good performance in some applications, recent works have shown that Gaussian model is sensitive to noise and irresistible to outliers when dealing with high-dimensional data. Indeed, among the challenges when using finite mixture modeling, there is the choice of appropriate parametric form of the probability density functions to represent the components.

Compared to the Gaussian, Langevin distribution has been shown to be a good alternative [5–7]. Usually, it is adopted to model problems involving high-dimensional spherical (L_2-normalized) vectors [5]. Indeed, it implicitly uses cosine similarity that is easy to interpret and simple to compute for sparse vectors, and has been widely used in text mining [8], spam filtering [6], gene expression analysis [9], and topic detection [10–12]. Works about directional data in general and spherical ones in particular have been developed thanks to the efforts of Watson, Stephens and others [13–23]. In this work, we shall consider finite Langevin mixtures. A key step in mixture-based modeling of data is parameter estimation. Many methods have been proposed in the literature in order to estimate mixture parameters, including frequentist (a.k.a. deterministic) and Bayesian approaches [24]. In this paper, we focus in developing parameter estimation and model selection from Bayesian perspective. We are mainly motivated by the fact that Bayesian learning has several desirable properties that make it widely used in several applications. For instance, it does not suffer over-fitting and prior knowledge is incorporated naturally in a principled way. In this paper we shall not motivate further Bayesian learning which has been widely discussed in the past (interested reader may refer to [25–27] for further details and interesting discussions).

Rooted in the early work of [28] Bayesian inference for the von Mises Fisher (vMF) distribution (3-dimensional case of the Langevin distribution) was proposed. This work was based on the development of a conjugate prior for the mean (Jeffreys prior was also developed for the polar coordinates) when concentration parameter is known. In the area of radio signals, authors in [29] applied Bayesian approach for finding the location of an emergency transmitter signal based on the von Mises (vM) distribution (2-dimensional case of the Langevin distribution) by developing conjugate priors using the canonical parameterization. A Gibbs sampler for vM distribution was introduced in [30] by developing conjugate priors for the polar coordinates. In [31] authors provide a full Bayesian analysis of directional data using the vMF distribution, again using standard conjugate priors and obtaining samples from the posterior using a sampling-importance-resampling method. Compared to these methods, our work is not restricted to low dimensional data (i.e. von Mises (2D) or von Mises Fisher (3D)) which is a limited solution for many real-world problems. On contrary, we extend previous models to high dimensional data using Langevin mixture (for $D > 3$) where both the concentration and mean parameters are unknown. In particular, we propose a Markov Chain Monte Carlo (MCMC) algorithm that relies on Gibbs sampler and Metropolis–Hastings (M-H) for the estimation of the parameters. To this end, we develop a conjugate prior for the Langevin distribution taking into account the fact that it belongs to the exponential family. As well as considering the estimation over model parameters, we also wish to consider the optimal number of components that best describe data at hand. One common approach is integrated likelihood [32] which we shall adopt for Langevin mixture in this paper. Note that, despite various efforts to use Bayesian inference

* Corresponding author. Tel.: +1 514 848 2424.
E-mail addresses: o_amayri@ece.concordia.ca (O. Amayri), nizar.bouguila@concordia.ca (N. Bouguila).

http://dx.doi.org/10.1016/j.asoc.2015.10.024
1568-4946/

to learn mixtures [33,34], to the best of our knowledge, none of the recent works has considered the case where the feature vectors to model are spherical so far.

The rest of this paper is organized as follows. In Section 2 we briefly introduce Langevin finite mixture model. In Section 3 we present previous parameter estimation approaches and then we propose a pure Bayesian algorithm for the estimation and selection of Langevin mixture model. Experimental results in vital and challenging problems, namely topic detection and tracking and image categorization, are presented in Section 4. Finally, Section 5 concludes the paper.

2. Finite Langevin mixture model

Let $\vec{X} = (X_1, \ldots, X_D)$ be a random unit vector in $\mathbb{R}^D$. $\vec{X}$ has D-variate Langevin distribution if its probability density function is given by [35]:

$$p_D(\vec{X}|\vec{\mu}, \kappa) = \frac{\kappa^{\frac{D}{2}-1}}{(2\pi)^{\frac{D}{2}} I_{\frac{D}{2}-1}(\kappa)} \exp\{\kappa \vec{\mu}^{tr}\vec{X}\} \quad (1)$$

on the $(D-1)$-dimensional unit sphere $\mathbb{S}^{D-1} = \{\vec{X}|\vec{X} \in \mathbb{R}^D : \|\vec{X}\| = \sqrt{\vec{X}^{tr}\vec{X}} = 1\}$, with mean direction unit vector $\vec{\mu} \in \mathbb{S}^{D-1}$, where $\vec{\mu}^{tr}$ denotes the transpose of $\vec{\mu}$ and non-negative real concentration parameter $\kappa \geq 0$. Furthermore, $I_D(\kappa)$ denotes the modified Bessel function of first kind [35]. Let $p(\vec{X}_i|\Theta)$ be a mixture of M Langevin distributions (i.e. a linear weighted combination of M distributions). The probability density function $p(\vec{X}_i|\Theta)$ is then given by

$$p(\vec{X}_i|\Theta) = \sum_{j=1}^{M} p_D(\vec{X}_i|\theta_j)p_j \quad (2)$$

where $\Theta = \{\vec{P} = (p_1, \ldots, p_M), \vec{\theta} = (\theta_1, \ldots, \theta_M)\}$ denotes all the parameters of the mixture model such that $\theta_j = (\mu_j, \kappa_j)$ and $\vec{P}$ represents the vector of clusters probabilities (i.e. mixing weights) such that $p_j \geq 0$ and $\sum_{j=1}^{M} p_j = 1$.

3. Parameter estimation

3.1. Likelihood estimation via expectation maximization (EM)

An efficient way to estimate the parameters of underlying mixture model is to optimize the associated likelihood function, which plays a key role in many estimation approaches such as EM and Bayesian. Assuming that the unit vectors to cluster, $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$, are independent and identically distributed, thus, the likelihood of Langevin mixture in Eq. (2) can be formulated as:

$$p(\mathcal{X}|\Theta) = \prod_{i=1}^{N} p(\vec{X}_i|\Theta) \quad (3)$$

One approach to estimate the Θ parameters of the mixture is to maximize the log likelihood

$$\log p(\mathcal{X}|\Theta) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{M} p_p(\vec{X}_i|\theta_j)p_j \right) \quad (4)$$

Maximum likelihood estimation is generally implemented via the EM framework [36] which generates a sequence of models with non-decreasing log-likelihood on the data. Following EM, consider the complete data to be $\{\vec{X}_i, \vec{Z}_i\}$, where $\vec{Z}_i = (Z_{i1}, \ldots, Z_{iM})$ denotes the missing vectors, such that $\sum_{j=1}^{M} Z_{ij} = 1$ with $Z_{ij} = 1$ if $\vec{X}_i$ belongs to class j and 0, otherwise. The E-step in EM computes the posterior probabilities given by the following equation:

$$\hat{Z}_{ij} = \frac{p(\vec{X}_i|\theta_j)p_j}{\sum_{j=1}^{M} p(\vec{X}_i|\theta_j)p_j} \quad (5)$$

where $\hat{Z}_{ij} \in [0, 1]$, $\sum_{j=1}^{M} \hat{Z}_{ij} = 1$ and denotes the degree of membership of $\vec{X}_i$ in the jth cluster. In the M-step, given the conditional expectation of complete log-likelihood, we update the parameters estimate by maximizing the complete data log likelihood from the E-step. A complete EM algorithm for Langevin mixture has been proposed in [5,6]. Unfortunately, maximum likelihood estimation does not provide tractable (closed form) solution for the parameters of Langevin mixture, especially that calculations include the ratio of Bessel function for the concentration parameter κ. Moreover, when the data at hand has high dimensionality and large number of components EM shows poor generalization and might lead to over- or under-fitting [33].

3.2. Bayesian estimation

As we previously discussed EM provides an elegant and simple way to estimate the parameters of a given model, yet, EM algorithm is sensitive to the initialization and generally converges to local solution in the best case. To avoid this problem, an alternative way is to use Bayesian estimation for Langevin mixture model.

Bayesian estimation is based on finding the conditional distribution $\pi(\Theta|\mathcal{X}, \mathcal{Z})$ of parameters vector Θ which is brought by complete data $(\mathcal{X}, \mathcal{Z})$, where $\mathcal{Z} = \{\vec{Z}_1, \ldots, \vec{Z}_N\}$. We therefore select a prior distribution $\pi(\Theta)$ and then develop posterior distribution $\pi(\Theta|\mathcal{X}, \mathcal{Z})$ which is derived from the joint distribution $p(\mathcal{Z}, \Theta, \mathcal{X})$ via Bayes formula $\pi(\Theta|\mathcal{X}, \mathcal{Z}) \propto p(\mathcal{Z}, \Theta, \mathcal{X})$. The joint distribution of all variables can be written as:

$$\pi(\Theta|\mathcal{X}, \mathcal{Z}) = p(\vec{\theta}, \vec{P}|\mathcal{X}, \mathcal{Z}) \propto p(\vec{P})p(\vec{\theta})p(\mathcal{Z}|\vec{P}) \prod_{Z_{ij}=1} p(\vec{X}_i|\theta_j) \quad (6)$$

where $p(\vec{\theta})$ and $p(\vec{P})$ are the priors of θ and $\vec{P}$ which we will describe in what follows.

3.2.1. Priors and posteriors

In order to derive our Bayesian algorithm we now turn to defining our priors over the parameters. Langevin distribution is a member of (curved)-exponential family of order D, whose shape is symmetric and unimodal. Thus, we can write it as the following [37]:

$$p(\vec{X}|\theta) = H(\vec{X}) \exp(G(\theta)^{tr}T(\vec{X}) + \Phi(\theta)) \quad (7)$$

where $G(\theta) = (G_1(\theta), \ldots, G_l(\theta))$, $T(\vec{X}) = (T_1(\vec{X}), \ldots, T_l(\vec{X}))$ where l is the number of parameters of the distribution and tr denotes transpose. The conjugate prior[1] on θ, in this case, can be written as [27]:

$$\pi(\theta) \propto \exp\left(\sum_{l=1}^{S} \rho_l G_l(\theta) + \lambda \Phi(\theta) \right) \quad (8)$$

where $\rho = (\rho_1, \ldots, \rho_S) \in \mathbb{R}^S$ and $\lambda > 0$ are referred as hyperparameters. To this end, Langevin distribution can be written as follows:

$$p_D(\vec{X}|\vec{\mu}, \kappa) = \exp\{\kappa \vec{\mu}^{tr}\vec{X} - a_D(\kappa)\} \quad (9)$$

[1] Ref. [38] contains an interesting discussion about the characteristics of conjugate priors and their induced posteriors in Bayesian inference for von Mises Fisher distributions, using either the canonical natural exponential family or the more commonly employed polar coordinate parameterizations.