Pattern Recognition 47 (2014) 3143–3157

Contents lists available at ScienceDirect

Pattern Recognition

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

Bayesian estimation of Dirichlet mixture model with variational inference

Zhanyu Ma [a,*], Pravin Kumar Rana [b], Jalil Taghia [b], Markus Flierl [b], Arne Leijon [b]

[a] Pattern Recognition and Intelligent Systems Lab., Beijing University of Posts and Telecommunications, Beijing, 100876, China
[b] KTH - Royal Institute of Technology, School of Electrical and Engineering, SE-100 44 Stockholm, Sweden

ARTICLE INFO

Article history:
Received 13 June 2013
Received in revised form
1 February 2014
Accepted 1 April 2014
Available online 12 April 2014

Keywords:
Bayesian estimation
Variational inference
Extended factorized approximation
Relative convexity
Dirichlet distribution
Gamma prior
Mixture modeling
LSF quantization
Multiview depth image enhancement

ABSTRACT

In statistical modeling, parameter estimation is an essential and challengeable task. Estimation of the parameters in the Dirichlet mixture model (DMM) is analytically intractable, due to the integral expressions of the gamma function and its corresponding derivatives. We introduce a Bayesian estimation strategy to estimate the posterior distribution of the parameters in DMM. By assuming the gamma distribution as the prior to each parameter, we approximate both the prior and the posterior distribution of the parameters with a product of several mutually independent gamma distributions. The extended factorized approximation method is applied to introduce a single lower-bound to the variational objective function and an analytically tractable estimation solution is derived. Moreover, there is only one function that is maximized during iterations and, therefore, the convergence of the proposed algorithm is theoretically guaranteed. With synthesized data, the proposed method shows the advantages over the EM-based method and the previously proposed Bayesian estimation method. With two important multimedia signal processing applications, the good performance of the proposed Bayesian estimation method is demonstrated.

1. Introduction

Statistical modeling plays an important role in various research areas [1–3]. It provides a way to connect the data with the statistics. An essential part in statistical modeling is to estimate the values of the parameters in the distribution or to estimate the distribution of the parameters, if we consider them as random variables. The maximum likelihood (ML) estimation method gives point estimates to the parameters and disregards the remaining uncertainty in the estimation. Rather than taking the point estimates, the Bayesian estimation method gives the posterior probability distributions over all model parameters, using the observed data together with the prior distributions [3]. In general, compared to the ML estimation, the Bayesian estimation of the parameters in a statistical model could yield a robust and stable estimate, by including the resulting uncertainty into the estimation, especially when the amount of the observed data is small [4].

The Gaussian distribution and the corresponding Gaussian mixture model (GMM) are widely used to model the underlying distribution of the data. However, not all data we would like to model can be safely assumed to be Gaussian distributed [5]. Recently, the studies of non-Gaussian statistical models have become popular for the purpose of modeling bounded or semi-bounded data (see *e.g.*, [6–9]). The non-Gaussian statistical models include, among others, the beta distribution, the gamma distribution, and the Dirichlet distribution.

The Dirichlet distribution and the corresponding Dirichlet mixture model (DMM) were frequently applied to model proportional data, for example, in image processing [10], in text analysis [11], and in data mining [12]. For speech processing, applications of Dirichlet distribution in the line spectral frequency (LSF) parameter quantization [13] were shown superior to conventional GMM based methods. Another usage of the Dirichlet distribution is to model the probabilities of the weighting factors in a mixture model [14,15]. In non-parametric Bayesian modeling, the Dirichlet process is actually an infinite-dimensional generalization of the Dirichlet distribution so that an infinite mixture model can be obtained [15–17]. Here, we study only the finite DMM and the work conducted can also be extended to the infinite mixture modeling case.

In this paper, we carry on our previous study of Bayesian analysis of BMM [8] and extend it to the Bayesian analysis of DMM. The parameters in a Dirichlet distribution are assumed mutually independent and each of them is assigned by a gamma prior.

* Corresponding author.
E-mail address: mazhanyu@bupt.edu.cn (Z. Ma).

http://dx.doi.org/10.1016/j.patcog.2014.04.002
0031-3203/

Although this assumption violates the correlation among the parameters, it captures the non-negative properties of those parameters. By this assumption, we can apply the factorized approximation (FA) method to carry out the Bayesian estimation. However, as the expectation of the multivariate log-inverse-beta (MLIB) function cannot be calculated explicitly, an analytically tractable solution to the posterior distribution is not feasible. To overcome this problem, we study some relative convex properties of the MLIB function. Using these convexities, we approximate the expectation of the MLIB function by a single lower-bound (SLB). With this derived SLB and by principles of the VI framework and the extended factorized approximation (EFA) method [8,18–24], we approximate the posterior distributions of the parameters in a Dirichlet distribution with a product of several mutually independent gamma distributions, which satisfies the conjugate match between the prior and posterior distributions. Finally, an analytically tractable solution for calculating the posterior distribution is obtained. This analytically tractable solution avoids the numerical calculations in the EM algorithm [25,10].

The proposed method, which is a full Bayesian framework, can automatically determine the model complexity (in terms of the number of necessary mixture components) based on the data. This task is also challenging in model estimation and the ML estimation itself cannot handle this issue. Moreover, the overfitting problem in the ML estimation can also be prevented due to the advantages of Occam's razor effect in Bayesian estimation. With synthesized data evaluation, the effectiveness and the accuracy of the proposed Bayesian estimation method over the ML estimation method [10,25] and the recently proposed Bayesian estimation method [12] are demonstrated. For the real life applications, we evaluate the proposed Bayesian estimation method with two important multimedia signal processing applications, namely (1) the LSF parameter quantization in speech coding [13] and (2) the multiview depth image enhancement in free-viewpoint television (FTV) [26]. For both applications, the proposed Bayesian method works well and shows improvement over the conventional methods.

The remaining parts of this paper are organized as follows: the DMM and the Bayesian analysis of a DMM are introduced in Sections 2 and 3, respectively. In Section 4, we show the efficiency and good performance of the proposed method with the synthesized data and the real life data. Some conclusions are drawn in Section 5.

2. Dirichlet mixture model

If a K-dimensional vector $\mathbf{x} = [x_1, \ldots, x_K]^{\mathrm{T}}$ contains only positive values and the summation of all the K elements is smaller than one, the underlying distribution of $\mathbf{x}$ could be modeled by a Dirichlet distribution. The probability density function (PDF) of a Dirichlet distribution is[1]

$$\mathrm{Dir}(\mathbf{x};\mathbf{u}) = \frac{\Gamma(\sum_{k=1}^{K+1} u_k)}{\prod_{k=1}^{K+1}\Gamma(u_k)} \prod_{k=1}^{K+1} x_k^{u_k-1}, \quad u_k > 0, \ \ 0 < x_k < 1, \tag{1}$$

where $x_{K+1} = 1 - \sum_{k=1}^{K} x_k$, $\mathbf{u} = [u_1, \ldots, u_{K+1}]^{\mathrm{T}}$ is the parameter vector, and $\Gamma(\cdot)$ is the gamma function defined as $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\,dt$. The shape of the Dirichlet distribution depends on the parameters. When $u_k > 1$, $k = 1, \ldots, K+1$, it is unimodally distributed. This is a typical case in practical applications. Thus in this paper, we study only the Dirichlet distribution with all its parameters greater than one.

[1] To prevent confusion, we use $f(x;a)$ to denote the PDF of x parameterized by parameter a. $f(x|a)$ is used to denote the conditional PDF of x given a, where both x and a are random variables. Both $f(x;a)$ and $f(x|a)$ have exactly the same mathematical expressions.

To model the multimodality of the data, the mixture modeling technique [14] can be applied to create a DMM. With I mixture components, the PDF of a DMM can be represented, given a set of N *i.i.d.* observations $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, as

$$f(\mathbf{X};\mathbf{\Pi},\mathbf{U}) = \prod_{n=1}^{N}\sum_{i=1}^{I} \pi_i\,\mathrm{Dir}(\mathbf{x}_n;\mathbf{u}_i), \quad \pi_i > 0, \quad \sum_{i=1}^{I}\pi_i = 1, \tag{2}$$

where $\mathbf{\Pi} = [\pi_1, \ldots, \pi_I]^{\mathrm{T}}$ is the mixture weights and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_I]$ is the parameter matrix.

3. Bayesian estimation with variational inference framework

For a distribution belonged to the exponential family, the conjugate prior and the corresponding posterior distribution always exist [3]. Similar to the beta distribution [8], the Dirichlet distribution has its conjugate prior and the corresponding posterior distributions. However, they are not tractable in practical use. Thus we follow the principle of VI framework [18,3] to approximate the prior and posterior distributions. With the proposed approximation, the obtained prior and posterior distributions can be easily calculated and used.

3.1. Conjugate prior to Dirichlet distribution

Since the Dirichlet distribution is a member of the exponential family, the conjugate prior of the Dirichlet distribution exists. If we assume that the parameter vector $\mathbf{u} = [u_1, \ldots, u_{K+1}]^{\mathrm{T}}$ is a vector random variable, then the prior distribution of $\mathbf{u}$ can be expressed as

$$f(\mathbf{u};\boldsymbol{\beta}_0,\nu_0) = \frac{1}{C(\boldsymbol{\beta}_0,\nu_0)}\left[\frac{\Gamma(\sum_{k=1}^{K+1}u_k)}{\prod_{k=1}^{K+1}\Gamma(u_k)}\right]^{\nu_0} e^{-\boldsymbol{\beta}_0^{\mathrm{T}}(\mathbf{u}-\mathbf{1}_{K+1})}, \tag{3}$$

where $\boldsymbol{\beta}_0 = [\beta_{1_0}, \ldots, \beta_{K+1_0}]^{\mathrm{T}}$ and ν_0 are the hyperparameters in the prior distribution. $C(\boldsymbol{\beta}_0,\nu_0)$ is the normalization factor. $\mathbf{1}_m$ denotes an m-dimensional vector with all elements equal to one. With Bayes' theorem and combining (1) and (3) together, we can obtain the posterior distribution of the parameters, given the observation $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, as

$$\begin{aligned} f(\mathbf{u}|\mathbf{X};\boldsymbol{\beta}_N,\nu_N) &= \frac{\mathrm{Dir}(\mathbf{X}|\mathbf{u})f(\mathbf{u};\boldsymbol{\beta}_0,\nu_0)}{\int \mathrm{Dir}(\mathbf{X}|\mathbf{u})f(\mathbf{u};\boldsymbol{\beta}_0,\nu_0)\,d\mathbf{u}} \\ &= \frac{1}{C(\boldsymbol{\beta}_N,\nu_N)}\left[\frac{\Gamma(\sum_{k=1}^{K+1}u_k)}{\prod_{k=1}^{K+1}\Gamma(u_k)}\right]^{\nu_N} e^{-\boldsymbol{\beta}_N^{\mathrm{T}}(\mathbf{u}-\mathbf{1}_{K+1})}, \end{aligned} \tag{4}$$

where $\boldsymbol{\beta}_N = \boldsymbol{\beta}_0 - \ln\mathbf{X}\times\mathbf{1}_N$, $\nu_N = \nu_0 + N$ are the hyperparameters in the posterior distribution. Since some statistics of $\mathbf{u}$, *e.g.*, the mean, the covariance, cannot be obtained directly (by an analytically tractable expression) from (3) or (4), it is not convenient to use them in practical problems. In the following paragraphs, we will apply the VI framework to approximate the prior and posterior distributions of the parameters in a DMM. These approximations can lead to an analytically tractable solution and would be easily used in practice.

3.2. Factorized approximation

In a DMM, the observations $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ are considered as the incomplete data and an I-dimensional indication vector $\mathbf{z}_n = [z_{n1}, \ldots, z_{nI}]^{\mathrm{T}}$ is assigned to each observation $\mathbf{x}_n$ to build a complete data set. Only one element in the indication vector is equal to 1 and the remaining elements are zeros. Thus $z_{ni} = 1$ indicates that the nth observation is generated from the ith mixture component. For N observations, we have N indication vectors denoted as $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]$. If we treat all the parameters in (2) as the random variables, the conditional PDF of the complete