Neurocomputing 139 (2014) 3–14

Contents lists available at ScienceDirect

Neurocomputing

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

Learning local factor analysis versus mixture of factor analyzers with automatic model selection

Lei Shi [a], Zhi-Yong Liu [b], Shikui Tu [a], Lei Xu [a,*]

[a] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[b] The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 3 June 2013
Received in revised form
29 August 2013
Accepted 15 September 2013
Available online 31 March 2014

Keywords:
Automatic model selection
Mixture of factor analyzers
Local factor analysis
Variational Bayes
Bayesian Ying-Yang
Dirichlet–Normal–Gamma

ABSTRACT

Considering Factor Analysis (FA) for each component of Gaussian Mixture Model (GMM), clustering and local dimensionality reduction can be addressed simultaneously by Mixture of Factor Analyzers (MFA) and Local Factor Analysis (LFA), which correspond to two FA parameterizations, respectively. This paper investigates the performance of Variational Bayes (VB) and Bayesian Ying-Yang (BYY) harmony learning on MFA/LFA for the problem of automatically determining the component number and the local hidden dimensionalities (i.e., the number of factors of FA in each component). Similar to the existing VB learning algorithm on MFA, we develop an alternative VB algorithm on LFA with a similar conjugate Dirichlet–Normal–Gamma (DNG) prior on all parameters of LFA. Also, the corresponding BYY algorithms are developed for MFA and LFA. A wide range of synthetic experiments shows that LFA is superior to MFA in model selection under either VB or BYY, while BYY outperforms VB reliably on both MFA and LFA. These empirical findings are consistently observed from real applications on not only face and handwritten digit images clustering, but also unsupervised image segmentation.

1. Introduction

Mixture models [1,2], such as Gaussian Mixture Model (GMM) [3,4], have been widely used in many applications. By exploiting the Factor Analysis (FA) [5] in each Gaussian component, the correlated high dimensional data can be represented by fewer latent factors without requiring $\mathcal{O}(d^2)$ parameters for each Gaussian covariance matrix, where d is the dimensionality of the data. The mixture model can be regarded as a constrained GMM, and has been studied under the name of Mixture of Factor Analyzers (MFA) [2,6] or Local Factor Analysis (LFA) [7,8] in the literature. MFA and LFA separately employ two parameterizations of FA, shortly called as FA-a that takes the form of a free factor loading matrix and an identity covariance matrix for the latent factors, and FA-b that constrains the factor loading matrix to be a rectangular orthogonal matrix, and allows a diagonal covariance matrix for the latent variables, respectively in [9].

Learning MFA/LFA includes parameter learning for estimating all the unknown parameters and model selection for determining the component number k and the hidden dimensionalities $\{h_i\}_{i=1}^{k}$. Parameter learning is usually implemented under the maximum likelihood principle by an Expectation–Maximization (EM) algorithm [1,10,11]. A conventional model selection approach is featured by a two-stage implementation. The first stage conducts parameter learning for each $\mathbf{k} \in \mathcal{M}$ to get a set of candidate models, where $\mathbf{k} = \{k, \{h_i\}\}$ for MFA/LFA. The second stage selects the best candidate by a model selection criterion, e.g., Akaike's Information Criterion (AIC) [12]. However, this two-stage implementation suffers from a huge computation because it requires parameter learning for each $\mathbf{k} \in \mathcal{M}$. Moreover, a larger $\mathbf{k}$ often implies more unknown parameters, and then parameter estimation becomes less reliable so that the criterion evaluation reduces its accuracy (see Section 2.1 in [13] for a detailed discussion).

To reduce the computation, an Incremental Mixture of Factor Analyzers (IMoFA) algorithm was proposed on MFA in [14] with the validation likelihood as the criterion to judge whether to split a component, or add a hidden dimension, or terminate. Although such an incremental procedure can save the costs to some extent, it usually leads to a suboptimal solution [13,15].

Another road is referred to as automatic model selection, which starts from a large enough $\mathbf{k}$, and has an intrinsic force to drive extra structures diminished, and thus automatically determines $\mathbf{k}$ during parameter learning. An early effort is Rival Penalized Competitive Learning (RPCL) on GMM [16,17]. Two Bayesian related approaches can be implemented with a nature of automatic model selection. One is Bayesian Ying-Yang (BYY) learning,

* Corresponding author.
E-mail address: lxu@cse.cuhk.edu.hk (L. Xu).

http://dx.doi.org/10.1016/j.neucom.2013.09.061
0925-2312/

proposed in [18] and systematically developed in the past decade and a half [13,15,19,20], which provides a general statistical learning framework that can handle both parameter learning and model selection under a best harmony principle. BYY is capable of automatic model selection even without imposing any priors on the parameters, and its performance can be further improved with appropriate priors incorporated according to a general guideline. The other is Variational Bayes (VB) [6,21]. It tackles the difficulty in computing the marginal likelihood with a lower bound by means of variational method, and an EM-like algorithm is employed to optimize this lower bound. The model selection of VB is realized by incorporating an appropriate prior distributions on the parameters.

Recently, a comparative study [4] was delivered on automatic model selection by BYY, VB and MML (Minimum Message Length) for GMM with priors over the parameters. Also in [9], FA-b shows better model selection performance than FA-a under BYY and VB, although FA-a and FA-b have equivalent likelihood functions.

This paper is motivated for an empirical investigation on the automatic model selection performances of BYY and VB, based on MFA and LFA, which actually correspond to Mixture of FA-a and Mixture of FA-b, respectively. There exists a VB algorithm [6] for MFA with a Dirichlet prior on the mixing weights, Normal priors on the columns of the factor-loading matrix, and Gamma priors on precision parameters. Following [4], we consider a full prior on all parameters and adopt a Normal prior over the mean vector in each component of MFA. For short, DNG is referred to the above Dirichlet, Normal, Gamma priors. By slightly modifying the one in [6], we obtain a VB learning algorithm with the DNG prior, shortly denoted as VB-MFA. Also, a similar conjugate DNG prior is considered on the parameters of LFA.

Moreover, we develop three automatic model selection algorithms, namely the VB algorithm on LFA, or VB-LFA for short, and the BYY algorithms on MFA and LFA, shortly denoted as BYY-MFA and BYY-LFA respectively. With the conjugate property of the priors, the BYY harmony measure is computed by directly integrating out the parameters with respect to the Yang posteriors, instead of using Taylor approximations as in [9]. The handled marginal density of observed variable in each component is tackled by a lower-bound approximation with the help of additional variables, leading to products of multiple Student's T-distributions.

The performances of automatic model selection are extensively compared on a wide range of randomly simulated data, via controlling the hardness of tasks by varying the dimension of data, the number of samples, the true number of components, and the overlap degree of components. The simulated results show the following empirical findings. First, LFA gets better performance than MFA under either VB or BYY, which echoes the advantages of FA-b over FA-a observed in [9]. Second, BYY outperforms VB on both MFA and LFA, and in most cases BYY-LFA performs the best. Also, we apply these algorithms to not only clustering face and handwritten digit images, but also unsupervised image segmentation on real world images. The results are consistent with the observations from simulated experiments.

The main contribution of this paper can be summarized in two-fold. First, three algorithms, i.e, the algorithm of VB based LFA with Dirichlet–Normal–Gamma (DNG) prior, denoted by VB-LFA, the algorithm of BYY based LFA with DNG prior, denoted by BYY-LFA, and the algorithm of BYY based MFA with DNG prior, denoted by BYY-MFA are derived in detail. Second, based on the algorithms, we empirically compared by extensive experiments the two types of clustering of factor analysis models, i.e., LFA and MFA, as well as two types of automatic model selection strategies, i.e., VB and BYY.

The remainder of this paper is organized as follows. Section 2 introduces MFA/LFA and their DNG priors. We introduce the automatic model selection algorithms with the DNG priors by BYY in Section 3, and by VB in Section 4. Experimental comparisons are conducted via a wide range of synthetic datasets and real applications in Section 5. Finally, concluding remarks are made in Section 6.

2. Models and priors

2.1. Model parameterizations

In a mixture model, the distribution $q(\mathbf{x}|\Theta)$ of a d-dimensional observed random variable $\mathbf{x}$ is a mixture of several local distributions $q(\mathbf{x}|i,\boldsymbol{\theta})$, with each named as a component:

$$q(\mathbf{x}|\Theta)=\sum_{i=1}^{k}\alpha_i q(\mathbf{x}|i,\boldsymbol{\theta}_i) \quad \text{with} \quad \Theta=\{\alpha_i\}\cup\{\boldsymbol{\theta}_i\}, \tag{1}$$

where k is the component number, $\{\alpha_i\}$ are mixing weights with $\sum_{i=1}^{k}\alpha_i=1$ and each $\alpha_i\geq 0$, and $\boldsymbol{\theta}_i$ denotes parameters of the ith component. Here and throughout this paper, $q(\cdot)$ is referred to as a generative distribution, likelihood or prior, while $p(\cdot)$ is referred to as a posterior distribution.

If each component is a Gaussian distribution, i.e., $q(\mathbf{x}|i,\Theta)=G(\mathbf{x}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_{x|i})$ with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_{x|i}$, $q(\mathbf{x}|\Theta)$ by Eq. (1) becomes the widely used Gaussian Mixture Model. For a full matrix $\boldsymbol{\Sigma}_{x|i}$, there are $0.5d(d+1)$ free parameters to be estimated, whose accuracy is difficult be guaranteed for a small sample size. One way for tackling this problem is to impose certain constraints on $\boldsymbol{\Sigma}_{x|i}$ with a Factor Analysis model, i.e.,

$$q(\mathbf{x}|\mathbf{y},i,\boldsymbol{\theta}_i)=G(\mathbf{x}|\mathbf{A}_i\mathbf{y}+\boldsymbol{\mu}_i,\boldsymbol{\Psi}_i),\quad q(\mathbf{y}|i,\boldsymbol{\theta}_i)=G(\mathbf{y}|\mathbf{0},\boldsymbol{\Sigma}_{y|i}),$$
$$q(\mathbf{x}|i,\boldsymbol{\theta}_i)=\int q(\mathbf{x}|\mathbf{y},i,\boldsymbol{\theta}_i)q(\mathbf{y}|i,\boldsymbol{\theta}_i)\,d\mathbf{y}=G(\mathbf{x}|\boldsymbol{\mu}_i,\mathbf{A}_i\boldsymbol{\Sigma}_{y|i}\mathbf{A}_i^T+\boldsymbol{\Psi}_i), \tag{2}$$

where we introduce a hidden factor $\mathbf{y}$ in an h_i-dimensional subspace with $h_i<d$, and constrain $\boldsymbol{\Psi}_i$ to be diagonal. FA actually factorizes $\boldsymbol{\Sigma}_{x|i}$ to be $\boldsymbol{\Sigma}_{x|i}=\mathbf{A}_i\boldsymbol{\Sigma}_{y|i}\mathbf{A}_i^T+\boldsymbol{\Psi}_i$ with fewer free parameters.

To reduce the indeterminacies of the FA by Eq. (2), two parameterizations of FA are typically used, called as Mixture of Factor Analyzers (MFA) [2,6] and Local Factor Analysis (LFA) [7,8] respectively, with their corresponding mixture models by Eq. (1) summarized in Table 1. The two FA parameterizations have equivalent likelihood functions by Eq. (2), and thus they have the same model selection performance in a two-stage implementation with AIC or BIC [22]. However, it was found that they result in different model selection performances under BYY [23], and a recent study [9] provided systematic empirical findings on how parameterizations affect model selection performance under not only BYY but also VB. Moreover, the differences of two parameterizations on model selection performance have been further analytically investigated in Section 2.2 of [20]. In this paper, we proceed to investigate the automatic model selection performances of MFA/LFA under BYY and VB.

Moreover, when each diagonal covariance $\boldsymbol{\Psi}_i$ in Table 1 is constrained to be spherical, i.e., $\boldsymbol{\Psi}_i=\psi_i\mathbf{I}_d$, MFA and LFA will degenerate to Mixture of PCA [11] and Local PCA [8], respectively.

Table 1
MFA v.s. LFA: similarity and difference. MFA and LFA are actually mixtures of FA-a and FA-b in [9], respectively.

Model:	MFA (mixture of FA-a)	LFA (mixture of FA-b)
Parameters $\boldsymbol{\theta}_i$:	$\{\mathbf{A}_i,\boldsymbol{\mu}_i,\boldsymbol{\Psi}_i\}$	$\{\mathbf{U}_i,\boldsymbol{\Lambda}_i,\boldsymbol{\mu}_i,\boldsymbol{\Psi}_i\}$
Same:	$\boldsymbol{\Psi}_i$ is $d\times d$ diagonal	$\boldsymbol{\Psi}_i$ is $d\times d$ diagonal
Different:	$\mathbf{A}_i$ is general $d\times h_i$	$\mathbf{U}_i$ is orthogonal, i.e., $\mathbf{U}_i^T\mathbf{U}_i=\mathbf{I}_{h_i}$, $\boldsymbol{\Lambda}_i$ is diagonal, $\boldsymbol{\Lambda}_i=\mathrm{diag}[\lambda_1,\ldots,\lambda_{h_i}]$
$q(\mathbf{y}\|i,\boldsymbol{\theta}_i)$:	$G(\mathbf{y}\|\mathbf{0},\mathbf{I}_{h_i})$	$G(\mathbf{y}\|\mathbf{0},\boldsymbol{\Lambda}_i)$
$q(\mathbf{x}\|\mathbf{y},i,\boldsymbol{\theta}_i)$:	$G(\mathbf{x}\|\mathbf{A}_i\mathbf{y}+\boldsymbol{\mu}_i,\boldsymbol{\Psi}_i)$	$G(\mathbf{x}\|\mathbf{U}_i\mathbf{y}+\boldsymbol{\mu}_i,\boldsymbol{\Psi}_i)$
$q(\mathbf{x}\|i,\boldsymbol{\theta}_i)$:	$G(\mathbf{x}\|\boldsymbol{\mu}_i,\mathbf{A}_i\mathbf{A}_i^T+\boldsymbol{\Psi}_i)$	$G(\mathbf{x}\|\boldsymbol{\mu}_i,\mathbf{U}_i\boldsymbol{\Lambda}_i\mathbf{U}_i^T+\boldsymbol{\Psi}_i)$