Neurocomputing 134 (2014) 149–158

Contents lists available at ScienceDirect

Neurocomputing

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

Learning binary factor analysis with automatic model selection

Shikui Tu, Lei Xu *

Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

ARTICLE INFO

Article history:
Received 4 June 2012
Received in revised form
12 December 2012
Accepted 31 December 2012
Available online 6 February 2014

Keywords:
Automatic model selection
Binary Factor Analysis
Variational Bayes
Bayesian Ying-Yang

ABSTRACT

Binary Factor Analysis (BFA) uncovers the independent binary information sources from observations with wide applications. BFA learning hierarchically nests three levels of inverse problems, i.e., inference of binary code for each observation, parameter estimation and model selection. Under Bayesian Ying-Yang (BYY) framework, the first level becomes an intractable Binary Quadratic Programming (BQP) problem, while model selection can be conducted automatically during parameter learning. We conduct extensive experiments to reveal that the performance order of four BQP methods is reversed from making BQP optimization to making BYY automatic model selection, which implies that learning is not merely optimization. Moreover, the BFA learning algorithm is further developed with priors over parameters to improve the performance. Finally, based on BFA, we empirically compare BYY with Variational Bayes (VB) and Bayesian information criterion (BIC).

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Binary Factor Analysis (BFA) explores latent binary structures of data. Unlike the conventional factor analysis where the latent factor is assumed to be Gaussian, BFA traces the observation to independent Bernoulli information sources. Research on BFA has been focused on analysis of binary data, such as social research questionnaires and market basket data, with the aid of Boolean algebra [1], and also on the discovery of binary factors in continuous data, [2–4], taking advantage of the representational capacity of the underlying binary structure. When considering all the random variables to be binary, factor analysis becomes the restricted Boltzmann machine which is the building block of the deep belief network [5]. This paper considers the same BFA model as in [4,2], under Bayesian Ying-Yang (BYY) harmony learning [6,7], in a comparison with Variation Bayes (VB) [8] and Bayesian information criterion (BIC) [9]. Rissanen's Minimum Description Length (MDL) stems from another viewpoint but coincides with BIC when it is simplified to a simple computable criterion [10].

The hierarchy of all unknowns in a learning system makes the learning process not just an optimization but a series of hierarchically nested continuous or discrete optimizations. As summarized in [7], there are three levels of inverse problems, i.e., inverse inference from observation to inner representation, parameter learning, and model selection. In terms of BFA, the first level of inverse problems in BFA is the inference of an m-bit inner binary code $\boldsymbol{y}(\boldsymbol{x})$ or a 2^m-point posterior distribution $p(\boldsymbol{y}|\boldsymbol{x})$ for each observation $\boldsymbol{x}$, given the parameters and the coding length of $\boldsymbol{y}$, i.e., $m=\dim(\boldsymbol{y})$. It is difficult due to its combinatorial complexity. Under BYY, maximizing the objective functional turns this problem into a Binary Quadratic Programming (BQP) problem that searches an optimal binary code $\boldsymbol{y}(\boldsymbol{x})$ for each training sample $\boldsymbol{x}$. A preliminary study in [11] compared four BQP methods and suggested that some amount of error in BQP optimization is not always a bad thing but instead provides a helpful regularization for the learning process. Conventionally, the second and the third level are implemented by a two-phase procedure, i.e., parameter learning (usually maximum likelihood learning) is conducted for each m in a candidate set $\mathcal{M}$, one of which is then selected by a model selection criterion, e.g., BIC [9]. However, this two-phase implementation suffers from a huge computation, because it requires parameter learning that is nested with a BQP for each $m\in\mathcal{M}$. Moreover, a larger m often implies more unknown parameters, and thus parameter estimation becomes less reliable so that the criterion evaluation reduces its accuracy, see Section 2.1 in [12] for a detailed discussion.

This paper further investigates the four BQP methods in [11] used for the BYY learning on BFA. One is the exact BQP solver by enumeration (shortly denoted as **enum**). The other three are approximate methods, i.e., the **greedy** method in [13], the **cdual** method derived from the canonical duality theory [14], and the **round** method by relaxing the binary $\boldsymbol{y}$ to a continuous one and rounding the optimal solution back to binary [15]. Their BQP optimization performances follow an order: **round** < **cdual** < **greedy** < **enum**. Extensive experiments show that **cdual** and **round** are fast and more effective in discarding extra factors, and lead to much better model selection

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

http://dx.doi.org/10.1016/j.neucom.2012.12.069
0925-2312 © 2014 Elsevier B.V. All rights reserved.

performances than **greedy** and **enum**. Actually, some amount of error in BQP provides a helpful learning regularization with a gain on both computational efficiency and model selection performance.

Moreover, automatic model selection is adopted to save the computation of two-phase implementation by starting from a large enough m and then discarding redundant binary factors during parameter learning. We further develop BFA learning algorithms by considering prior distributions over parameters, which play a role of Bayesian regularization. With the help of priors, **enum** and **greedy** improve their automatic model selection performances, but are still inferior to **cdual** and **round**.

Finally, we empirically investigate the performance between BYY, VB, and BIC. Such comparisons have been made on factor analysis in [16] and Gaussian mixture model in [17], but not on BFA yet. We simplify the VB-ICA algorithm [18,19] to obtain a VB algorithm on BFA. The results reveal that BYY is the best for most configurations, while BIC is more robust than VB. VB is good only when both training sample size N is large and noise is small, and declines drastically when N reduces and noise increases. Moreover, applied to the problem of blind binary image separation, the results again show that BYY outperforms VB.

The rest of this paper is organized as follows. BFA model is introduced in Section 2. BYY harmony learning is briefly reviewed in Section 3, and a BYY-BFA algorithm is derived with priors over the parameters. Section 4 introduces VB and BIC for an empirical analysis in Section 5, while concluding remarks are given in Section 6.

2. Binary Factor Analysis

In Binary Factor Analysis (BFA), an n-dimensional observed variable $\boldsymbol{x}$ is modeled as

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{c} + \boldsymbol{\varepsilon}, \tag{1}$$

where the hidden factor vector $\boldsymbol{y} \in \{-1, 1\}^m$ is an internal binary code with each element being either -1 or 1 drawn from a Bernoulli distribution, and $\boldsymbol{y}$ is independent of the Gaussian noise $\boldsymbol{\varepsilon}$. This model has been studied previously from different perspectives [15,4,2].

The BFA can also be mathematically formalized by the following probabilistic distributions:

$$q(\boldsymbol{y}|\boldsymbol{\Theta}) = \prod_{i=1}^{m} \beta_i^{(1+y_i)/2}(1-\beta_i)^{(1-y_i)/2}, \quad q(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\Theta}) = G(\boldsymbol{x}|\boldsymbol{A}\boldsymbol{y}+\boldsymbol{c}, \boldsymbol{\Sigma}_e), \tag{2}$$

where $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_m]$, $0 < \beta_i < 1$, $i = 1, 2, \ldots, m$, $\boldsymbol{\Sigma}_e$ is a positive definite diagonal matrix, and $G(\cdot|\boldsymbol{\mu}, \boldsymbol{\Psi})$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Psi}$, and $\boldsymbol{\Theta} = \{\boldsymbol{A}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\Sigma}_e\}$ is the set of parameters.

Similar to [20,18], we consider the joint prior distribution on the parameters $\boldsymbol{\Theta} = \{\boldsymbol{A}, \boldsymbol{\Sigma}_e, \boldsymbol{\beta}, \boldsymbol{c}\}$ to be a product of distributions on each parameter independently:

$$q(\boldsymbol{\Theta}|\Xi) = q(A)q(\boldsymbol{\beta})q(\boldsymbol{c})q(\boldsymbol{\Sigma}_e), \tag{3}$$

where Ξ is the set of hyperparameters. Each column $\mathbf{a}_i$ of A is independently distributed according to a Gaussian distribution with its covariance controlled by a precision parameter α_i which is further assumed to follow a Gamma distribution

$$q(A) = \prod_{i=1}^{m} G\left(\mathbf{a}_i|0, \frac{1}{\alpha_i}\mathbf{I}_n\right), \quad q(\alpha_i) = \Gamma(\alpha_i|a^\alpha, b^\alpha), \tag{4}$$

where $\Gamma(x|a,b) = (b^a/\Gamma(a))x^{a-1}e^{-bx}$ denotes the Gamma density. A Dirichlet distribution is appropriate for each β_i which satisfies $\beta \in [0,1]$:

$$q(\boldsymbol{\beta}) = \prod_{i=1}^{m} \mathcal{D}(\beta_i|\boldsymbol{\lambda}_i, \xi_i) = \prod_{i=1}^{m} \frac{\Gamma(\xi_i)\cdot \beta_i^{\xi_i\lambda_{i1}-1}(1-\beta_i)^{\xi_i\lambda_{i2}-1}}{\Gamma(\xi_i\beta_i)\Gamma(\xi_i(1-\beta_i))}. \tag{5}$$

Usually, $q(\boldsymbol{\mu})$ is assumed to be a Gaussian with zero mean, i.e., $G(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \lambda_0^\mu \mathbf{I}_n)$. Moreover, the case of isotropic noise is considered, i.e., $\boldsymbol{\Sigma}_e = \varphi^{-1}\mathbf{I}_n$, and a Gamma distribution is imposed on the noise precision parameter φ:

$$q(\boldsymbol{\Sigma}_e) = q(\varphi) = \Gamma(\varphi|a^\varphi, b^\varphi). \tag{6}$$

3. Bayesian Ying-Yang (BYY) harmony learning

Firstly proposed in [6] and systematically developed over a decade and half [12,21], the Bayesian Ying-Yang harmony learning theory is a unified statistical learning framework under a best harmony principle, which leads to a new family of algorithms that performs automatic model selection during parameter learning. The best harmony is mathematically to maximize the following general harmony functional [12,7]:

$$H(p\|q) = \int p(X)p(R|X)\ln[q(X|R)q(R)]dR\,dX \tag{7}$$

$$H(p\|q) = \int p(\boldsymbol{\Theta}|X)H(p\|q,\boldsymbol{\Theta})d\boldsymbol{\Theta}, \tag{8}$$

$$H(p\|q,\boldsymbol{\Theta}) = \int p(Y|X,\boldsymbol{\Theta})p(X)\ln[q(X|Y,\boldsymbol{\Theta})q(Y|\boldsymbol{\Theta})]dY\,dX + \ln q(\boldsymbol{\Theta}|\Xi), \tag{9}$$

where the observation X is regarded to be generated from its inner representation $R = \{Y, \boldsymbol{\Theta}\}$ with latent variable Y and parameters $\boldsymbol{\Theta}$. As interpreted in [7], maximizing $H(p\|q)$ forces $q(X|R)q(R)$ to match $p(R|X)p(X)$. Due to a finite sample size and practical constraints on $p(R|X)$, this matching aims at but may not really reach a perfect matching $p(R|X)p(X) = q(X|R)q(R)$. Still, we get a trend at this equality which turns $H(p\|q)$ into a negative entropy that describes the complexity of system, and thus further maximizing it leads to a least complexity. Hence, this matching is not in a maximum likelihood sense but with a promising model selection nature. Readers are referred to not only a summary of nine aspects on the novelty and favorable natures of BYY harmony learning, made at the end of Section 4.1 in [12], but also the roadmap shown in Fig. A2 in [12], as well as to a systematic outline on the 13 topics about best harmony learning in Section 7 in [21].

The model selection performance of not only BYY criterion but also BYY automatic model selection on BFA has been comparatively investigated in [4], in comparison with existing typical model selection criteria, including Bayesian Information Criterion (BIC) [9] etc., which are implemented in a two-phase procedure that first trains a set of candidate models and then selects the one with the minimum criterion value. This two-stage implementation suffers from a huge computation because it requires parameter learning for each candidate model scale. Moreover, a larger model scale often implies more unknown parameters, and thus parameter estimation becomes less reliable so that the criterion evaluation reduces its accuracy, see Section 2.1 in [12] for a detailed discussion. This paper focuses on BYY based automatic model selection, incorporated with appropriate prior distributions on parameters.

Specifically, we consider the BFA model by Eq. (2) with independently and identically distributed (i.i.d.) samples in $X_N = \{\boldsymbol{x}_t\}_{t=1}^N$, from which we have

$$q(X|Y,\boldsymbol{\Theta}) = \prod_t q(\boldsymbol{x}_t|\boldsymbol{y}_t,\boldsymbol{\Theta}), q(Y|\boldsymbol{\Theta}) = \prod_t q(\boldsymbol{y}_t|\boldsymbol{\Theta}), \tag{10}$$

Download English Version:

https://daneshyari.com/en/article/410049

Download Persian Version:

https://daneshyari.com/article/410049

Daneshyari.com