Bayesian Ying-Yang Learning on Orthogonal Binary Factor Analysis

Ke Sun and Lei Xu

Department of Computer Science and Engineering
The Chinese University of Hong Kong, Hong Kong, P.R. China
{ksun,lxu}@cse.cuhk.edu.hk

Abstract. Binary Factor Analysis (BFA) aims to discover latent binary structures in high dimensional data. Parameter learning in BFA suffers from exponential complexity and a large number of local optima. Model selection in BFA is therefore difficult. The traditional approach for model selection is implemented in a two phase procedure. On a prefixed range of model scales, maximum likelihood (ML) learning is performed for each candidate scale. After this enumeration, the optimum scale is selected according to some criterion. In contrast, the Bayesian Ying-Yang (BYY) learning starts from a high dimensional model and automatically deduces the dimension during parameter learning. The enumeration overhead in the two phase approach is saved. This paper investigates a subclass of BFA called Orthogonal Binary Factor Analysis (OBFA). A BYY machine for OBFA is constructed. The harmony measure, which serves as the objective function in the BYY harmony learning, is more accurately estimated by recovering a term that was missing in the previous studies on BYY learning based BFA. Comparison with traditional two phase implementations shows good performance of the proposed approach.

1 Introduction

Latent structures often exist in high dimensional observations. Depending on the characteristics of data, such a structure can be a manifold of much a lower dimension or just a discrete set. Discovering the underlying structure with an appropriate scale of parametric model is critical in parameter learning. In the classical Factor Analysis (FA) [1], the underlying structure is assumed to be a low dimensional Gaussian distribution. A variant model called Binary Factor Analysis (BFA) adopts a vector of independent Bernoulli distribution as the latent model. As in FA, BFA also faces the difficulty of model selection, that is, to determine an appropriate number of binary factors so that the resulting model represents the regularity well but does not overfit the training data. The problem is even more difficult in BFA because of the combinatorial complexity in computation and a huge number of local optima. Research on BFA under the names of latent trait model, item response theory, latent class model or multiple cause model [2,3,4] is widely used in data reduction, psychological measurement, political science, etc.
This paper focuses on a subclass of BFA called Orthogonal Binary Factor Analysis (OBFA) which further restricts the loading matrix in BFA to be orthogonal. OBFA provides a general method to construct a (orthogonal) coordinate system so that the clusters in the training data are separated by certain coordinate planes. This coordinate system is valuable in many practical applications. In a 2-level Orthogonal Experiment Design (OED) \[5\] with a set of high dimensional structured experiment inputs available, the task is to extract \(2^m\) well separated representative inputs corresponding to \(m\) independent factors. Here OBFA can be applied to learn a coordinate system with the representation inputs lying at \(\{-1, 1\}^m\). A parallel binary channel with additive noise and a rotation transformation at the output end results naturally in the OBFA model. In an artificial neural network with one hidden layer, the first layer can be regarded as encoding of space regions as binary vectors and thus can be initialized with the equations of such separation planes. Psychological measurement with questionnaires also falls in the category of OBFA if the distribution of answers is linear separable by several orthogonal planes. Moreover, taking advantage of the formal analysis results, OBFA can be utilized to initialize BFA.

The conventional two phase approach for model selection has to enumerate a candidate set of model scales, denoted as \(\mathcal{K}\). ML learning is performed for each \(k \in \mathcal{K}\) to estimate the parameters \(\Theta\). The optimal scale \(k^*\) is selected with

\[
\hat{k}^* = \arg\min_{k \in \mathcal{K}} J(\hat{\Theta}, k),
\]

where \(J(\hat{\Theta}, k)\) is a model selection criterion, such as Akaike’s Information Criterion (AIC) \[6\], Bozdogan’s Consistent Akaike’s Information Criterion (CAIC) \[7\], Schwarz’s Bayesian Information Criterion (BIC) \[8\], Rissanen’s Minimum Description Length (MDL) criterion \[9\]. Huge computation is evolved in this enumeration, especially for problems with a high computational complexity as BFA.

Proposed firstly in 1995 \[10\] and systematically developed in the past decade \[11,12,13\], the BYY harmony learning provides a general framework for parameter learning and model selection. Under this framework, model selection can be performed automatically during parameter learning. In this way the computational overhead in the two phase enumeration is saved. The BYY harmony learning is mathematically implemented by maximizing the following harmony measure

\[
H(p \mid \mid q) = \int p(R \mid X)p(X) \log [q(X \mid R)q(R)] dXdR,
\]

where \(X\) is the external observations, \(R = \{Y, \Theta\}\) consists of the inner states \(Y\) and all unknown parameters \(\Theta\). The joint distribution \(q(X, R) = q(X \mid R)q(R)\) represents the underlying model; the complementary joint distribution \(p(X, R) = p(R \mid X)p(X)\) is the posterior model constructed from the observations.

Much work has been dedicated to FA and BFA using the BYY harmony learning \[11,12,14\]. When the other structures are fixed, several typical choices