Local Factor Analysis with Automatic Model Selection and Data Smoothing Based Regularization

Lei Shi and Lei Xu
Department of Computer Science & Engineering
Chinese University of Hong Kong
Shatin, NT, Hong Kong
{shil,lxu}@cse.cuhk.edu.hk

Abstract

Local factor analysis (LFA) is regarded as an efficient approach that implements local feature extraction and dimensionality reduction. A further investigation is made on an automatic BYY harmony data smoothing LFA (LFA-HDS) from the Bayesian Ying-Yang (BYY) harmony learning point of view. On the level of regularization, an data smoothing based regularization technique is adapted into this automatic LFA-HDS learning for problems with small sample sizes, while on the level of model selection, the proposed automatic LFA-HDS algorithm makes parameter learning with automatic determination of both the component number and the factor number in each component. A comparative study has been conducted on simulated data sets and several real problem data sets. The algorithm has been compared with not only a recent approach called Incremental Mixture of Factor Analysers (IMoFA) but also the conventional two-stage implementation of maximum likelihood (ML) plus model selection, namely, using the EM algorithm for parameter learning on a series candidate models, and selecting one best candidate by AIC, CAIC, BIC, and cross-validation (CV). Experiments have shown that IMoFA and ML-BIC, ML-CV outperform ML-AIC or ML-CAIC. Interestingly, the data smoothing BYY harmony learning obtains comparably desired results compared to IMoFA and ML-BIC but with much less computational cost.

1 Introduction

Clustering and feature extraction are two fundamental problems in the literature of unsupervised learning. It is well known that Gaussian mixture model (GMM) with full covariance matrices requires sufficient training data to guarantee the reliability of the estimated model parameters, while GMM with diagonal covariance matrices requires a relatively large number of Gaussians to provide high recognition performance. Local factor analysis (LFA), also named mixture of factor analyzers (MFA), combines the widely-used GMM model with one well known feature extraction and dimension reduction approach, namely factor analysis (FA). Via local structure analysis, LFA is able to reduce the freedom degree of covariance matrices to achieve a good generalization. Several efforts have been made on such a topic of local feature extraction and dimensionality reduction [2, 3, 12].

In the literature of LFA research, the conventional method performs the maximum likelihood (ML) learning in help of one of typical statistical criteria to select both component number and local dimensions of local factor analysis via a two-phase procedure. However, it suffers a huge computing cost. Bayesian Ying-Yang (BYY) learning was proposed as a unified statistical learning framework firstly in 1994 and systematically developed in the past decade. Providing a general learning framework, BYY harmony learning consists of a general BYY system and a fundamental harmony learning principle as a unified guide for developing new regularization techniques, a new class of criteria for model selection, and a new family of algorithms that perform parameter learning with

automatic model selection. Readers are referred to [15, 17] for a recent systematical review. Applying the BYY harmony learning to local factor analysis, an adaptive learning algorithm has been developed to perform local factor analysis with both the local dimensions of each component and the number of components automatically determined during parameter learning [14, 16, 11].

However, when the number of observations is small, both ML learning and original BYY automatic LFA learning may obtain a poor estimation. To cover this problem, data smoothing [13] provides a useful regularization approach. The basic idea of data smoothing regularization is to learn a parametric model together with a Parzen window nonparametric model via a smoothing parameter h^2 [13]. This work focuses on automatic BYY harmony data smoothing LFA learning (LFA-HDS), which extends the previous effort in [11] by adapting the smoothing based regularization technique into the original automatic BYY harmony learning on LFA.

This paper investigates the automatic BYY harmony data smoothing LFA learning, in comparison with the ML learning via criteria of AIC, CAIC, BIC, as well as a recently proposed approach called Incremental Mixture of Factor Analyzers (IMoFA)[10] that makes an increasing model selection during learning. A comparative study is conducted via experiments on not only simulated data but also several real problem data sets, as well as a popular digit recognition database, respectively.

2 FA and LFA

Factor analysis (FA) is a classical feature extraction and dimension reduction technique aiming to find the hidden causes and sources [7]. Provided a d-dimensional vector of observable variables $\mathbf{x}$, the FA model is given by $\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{e}$, where $\mathbf{A}$ is a $d \times k$ loading matrix, $\mathbf{y}$ is an m-dimensional unobservable latent vector assumed from Gaussian $G(\mathbf{y}|\mathbf{0}, \mathbf{I_k})$ with $m < d$, $\mathbf{e}$ is a d-dimensional random noise vector assumed from Gaussian $G(\mathbf{e}|\mathbf{0}, \mathbf{\Psi})$ with $\mathbf{\Psi}$ being a diagonal matrix. Moreover, $\mathbf{y}$ and $\mathbf{e}$ are mutually independent. Therefore, $\mathbf{x}$ is distributed with zero mean and covariance $\mathbf{A}\mathbf{A}^T + \mathbf{\Psi}$. The goal of FA is to find $\theta = \{\mathbf{A}, \mathbf{\Psi}\}$ that best models the structure of $\mathbf{x}$. One widely used method to estimate θ is the maximum likelihood (ML) learning that maximizes the log-likelihood function, usually implemented by the expectation-maximization (EM) algorithm [1, 7].

Local factor analysis (LFA), also called mixture of factor analyzers (MFA), is a useful multivariate analysis tool to explore not only clusters but also local subspaces with wide applications including pattern recognition, bioinformatics, and financial engineering [14, 9]. LFA performs clustering analysis and feature extraction in each cluster simultaneously. Provided $\mathbf{x}$ as a d-dimensional random vector of observable variables, the mixture model assumes that $\mathbf{x}$ is distributed according to a mixture of k underlying probability distributions $p(\mathbf{x}) = \sum_{l=1}^{k} \alpha_l p_l(\mathbf{x})$, where $p_l(\mathbf{x})$ is the density of the lth component in the mixture, and α_l is the probability that an observation belongs to the lth component with $\alpha_l \geq 0, l = 1, \ldots, k$, and $\sum_{l=1}^{k} \alpha_l = 1$. For LFA, it is further assumed that each $p_l(\mathbf{x})$ is modelled by a single FA [14]. For a set of observations $\{\mathbf{x}_t\}_{t=1}^{n}$, supposing that the number of components k and the numbers of local factors $\{m_l\}$ are given, one widely used method to estimate the unknown parameters $\theta = \{\alpha_l, \mathbf{A}_l, \mathbf{c}_l, \mathbf{\Psi}_l\}_{l=1}^{k}$ is the maximum likelihood (ML) learning, which can be effectively implemented by expectation-maximization (EM) algorithm [1, 3].

2.1 Conventional Statistical Criteria

Two important problems for LFA are how to select the number of Gaussian components k and how to decide the numbers of sub-factors $\{m_l\}_{l=1}^{k}$. They can be addressed in a *two-phase* procedure in help of typical statistical model selection criteria such as Akaike's information criterion (AIC) [4], Bozdogan's consistent Akaike's information criterion (CAIC) [6], Schwarz's Bayesian inference criterion (BIC) [8] which coincides with Rissanen's minimum description length (MDL) criterion [5], and the cross-validation (CV) technique. These criteria are based on the maximum likelihood (ML) estimates of parameters which can be obtained by the EM algorithm [3, 7].

In the first phase, two ranges of $k \in [k_{min}, k_{max}]$ and $m_l \in [m_{min}, m_{max}]$ are selected to form a domain $\mathcal{M}$, assumed to contain the optimal $k^*, \{m_l^*\}_{l=1}^{k^*}$. At each specific choice of $k, \{m_l\}$ in $\mathcal{M}$, the parameters are estimated θ via the ML learning. In the second phase, selection is made among all candidate models obtained in the first phase according to their criterion values, that is:

$$\hat{k}, \{\hat{m_l}\} = \arg \min_{\{k, \{m_l\}\} \in \mathcal{M}} J(\hat{\theta}, k, \{m_l\}), \tag{1}$$

However, in this domain $\mathcal{M}$, we have to implement EM algorithm at least $\sum_{k=k_{min}}^{k_{max}}(m_{max} - m_{min} + 1)^k$ times, which is usually too time-consuming if we have no knowledge or assumption about the underlying model structure.

2.2 Incremental Mixture of Factor Analysers

Recently, an adaptive algorithm referred as *incremental mixture of factor analysers (IMoFA)* was proposed in [10]. Starting with a 1-factor, 1-component mixture model, in process, IMoFA either splits component or adding local factors according to the validation likelihood, which is terminated when there is no improvement on the validation likelihood. There are two variants IMoFA-L and IMoFA-A for unsupervised and supervised approaches, respectively. In this paper, we consider the unsupervised learning with IMoFA-L, shortly denoted by IMoFA. The detailed procedure and algorithm is referred to [10].

3 Automatic BYY Harmony Data Smoothing Learning for LFA

Bayesian Ying-Yang (BYY) harmony learning provides a promising tool for local factor analysis with an ability of determining the number of components as well as the number of local factors during parameters learning [14, 16, 17, 11]. Considering the idea of data smoothing regularization, a parametric model together with a Parzen window nonparametric model with a smoothing parameter h^2, i.e., $p_h(\mathbf{x}) = \frac{1}{n}\sum_{t=1}^{n} G(\mathbf{x}|\mathbf{x}_t, h^2\mathbf{I}_d)$.

Parameters $\theta_s = \{h^2, \alpha_l, \mathbf{U}_l, \mathbf{\Lambda}_l, \mathbf{c}_l, \mathbf{\Psi}_l\}_{l=1}^{k}$ can be estimated by BYY harmony learning, where the optimization problem is given as follows:

$$
\begin{aligned}
&\hat{\theta}_s = \arg\max_{\theta_s} H(\theta_s, k), \quad H(\theta_s, k) = \sum_{l=1}^{k}\sum_{t=1}^{n} P(l|\mathbf{x}_t) L_s(\theta_s, l) \\
\text{where} \quad &L_s(\theta_s, l) = \ln\alpha_l - \frac{1}{2}\ln|\mathbf{\Sigma}_l| - \frac{1}{2}tr(\mathbf{S}_{l,h}\mathbf{\Sigma}_l^{-1}) + \frac{d}{2}\ln h^2 - Z_s(h^2), \\
&\mathbf{\Sigma}_l = \mathbf{A}_l\mathbf{A}_l^T + \mathbf{\Psi}_l \\
&Z_s(h^2) = ln[\sum_{t=1}^{n}\sum_{\tau=1}^{n} exp(-\frac{||\mathbf{x}_t - \mathbf{x}_\tau||^2}{2h^2})] \\
&\mathbf{S}_{l,h} = \mathbf{S}_l + h^2\mathbf{I}_d, \text{ and } \sum_{l=1}^{k}\alpha_l = 1,\ l = 1, ..., k.
\end{aligned} \tag{2}
$$

In a B-architecture, $P(l|\mathbf{x}_t)$ is free and thus it follows from the above maximization that we have

$$
\begin{aligned}
&P(l|\mathbf{x}_t) = \begin{cases} 1, & l = l_t; \\ 0, & \text{otherwise.} \end{cases} \\
&l_t = \arg\max_{l} \ln[\alpha_l p_l(\mathbf{x}_t|\mathbf{y}_{l,t}) p_l(\mathbf{y}_{l,t})] = L_s(\theta_s, l), \\
&\mathbf{y}_{l,t} = \arg\max_{\mathbf{y}} \ln(p_l(\mathbf{x}_t|\mathbf{y}) p_l(\mathbf{y})).
\end{aligned} \tag{3}
$$

Performing (2) results in maximizing $\ln\alpha_l$, $\ln p_l(\mathbf{x}|\mathbf{y})$ and $\ln p_l(\mathbf{y})$, which will push α_l or $\mathbf{\Psi}_l$ towards zero if component l is extra. Thus we can delete component l if its corresponding α_l or $\mathbf{\Psi}_l$ is approaching to zero. Also, if the latent dimension $\mathbf{y}^{(j)}$ is extra, maximizing $\ln p_l(\mathbf{y})$ will push the variance $\mathbf{\Lambda}_l^{(j)}$ towards zero, thus factor j can be deleted. As long as k and $\{m_l\}$ are initialized at values large enough, they will be determined appropriately and automatically during parameter learning, with details referred to [14, 15]. To compare with the EM algorithm in a batch way, here we also consider a batch algorithm to implement Eq.(2), which is not shown here due to the space limitation.

After the training of LFA, for classification, we first obtain $M_{lj}, l = 1, \cdots, k_j$ by BYY-LFA for each class $j = 1, \cdots, C$. As a test data $\mathbf{y}_i$ comes, we compute the the likelihoods $p(\mathbf{y}_i|M_{lj}), l = 1, \cdots, k_j, j = 1, \cdots, C$ and find the κ largest ones. Then, we classify $\mathbf{y}_i$ to the class $j^* =$

$arg\max_j \kappa_j$, where κ_j is the account that the κ largest ones share the class label j. This decision rule actually shares the idea of the well known k-NN approach, shortly denoted by a BYY-LFA Rank-κ rule.

4 Summary

In order to make local feature extraction and dimension reduction, through the local factor analysis (LFA) model, we investigate the automatic BYY harmony data smoothing LFA learning (LFA-HDS), compared with typical model selection criteria via the conventional two-phase procedure and the IMoFA approach.

Due to the space limitation of this extended abstract, several series of experimental results and conclusion will be found in the final version.

References

References

[1] Redner, R.A. and Walker, H.F.: Mixture densities, maximum likelihood, and the EM algorithm. *SIAM Review*, vol.26, pp.195-239, 1984.

[2] Kambhatla, N. and Leen, T.K.: Fast non-linear dimension reduction. *Advances in NIPS 6*, Morgan Kaufmann, San Francisco, 1994.

[3] Hinton G.E., Revow, M. and Dayan, P.: Recognizing handwritten digits using mixtures of Linear models. *Advances in NIPS 7*, MIT Press, Cambridge, MA, 1995.

[4] Akaike, H.: A new look at statistical model identification. IEEE Trans. *Automatic Control*, vol.19, pp.716-723, 1974.

[5] Barron, A. and Rissanen, J.: The minimum description length priciple in coding and modeling. *IEEE Trans. Information Theory*, vol.44, pp.2743-2760, 1998.

[6] Bozdogan, H.: Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, vol.52(3), pp.345-370, 1987.

[7] Rubin, D. and Thayer, D.: EM algorithms for ML factor analysis. *Psychometrika*, vol.47(1), pp.69-76, 1982.

[8] Schwarz, G.: Estimating the dimension of a model. *The Annals of Statistics*, vol.6(2), pp.461-464, 1978.

[9] Ghahramani, Z. and Beal, M.: Variational inference for Bayesian mixture of factor analysers. *Advances in NIPS*, vol.12, pp.449-455, 2000.

[10] Albert Ali Salah and Ethem Alpaydin: Incremental Mixtures of Factor Analysers. *Proc. 17th Intl Conf. on Pattern Recognition*, vol.1, 276-279, 2004.

[11] Shi, L. and Xu, L.: Local Factor Analysis with Automatic Model Selection: A comparative Study and Digits Recognition Application, *Artificial Neural Networks - ICANN 2006: Lecture Notes in Computer Sciences*, Springer Berlin / Heidelberg, vol.4132, pp. 260-269.

[12] Xu, L.: Multisets Modeling Learning: An Unified Theory for Supervised and Unsupervised Learning, Invited Talk, *Proc. IEEE ICNN94*, June 26-July 2, 1994, Orlando, Florida, Vol.I, pp.315-320.

[13] Xu, L.: Data smoothing regularzation, multi-sets-learning, and problem solving strategies. *Neural Networks*, vol.16(5-6), pp.817-825, 2003.

[14] Xu, L.: Temporal BYY Encoding, Markovian State Spaces, and Space Dimension Determination. *IEEE Trans. Neural Networks*, vol.15(5), pp.1276-1295, 2004.

[15] Xu, L: Advances on BYY harmony learning: information theoretic perspective, generalized projection geometry, and independent factor auto-determination, *IEEE Trans on Neural Networks, Vol. 15, No. 4*, pp885-902, 2004.

[16] Xu, L.: A Unified Perspective and New Results on RHT Computing, Mixture Based Learning, and Multi-learner Based Problem Solving. To appear in a special issue of *Pattern Recognition*, 2006.

[17] Xu, L.: Trends on Regularization and Model Selection in Statistical Learning: A Perspective from Bayesian Ying Yang Learning. *Challenges to Computational Intelligence* (in press), Duch, W., Mandziuk, J. and Zurada, J.M. eds, the Springers series - Studies in Computational Intelligence, Springer-Verlag, 2006.