Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

Maximum entropy discrimination factor analyzers

Sotirios P. Chatzis

Department of Electrical Engineering, Computer Engineering, and Informatics, Cyprus University of Technology, 33 Saripolou Str., Limassol 3036, Cyprus

ARTICLE INFO

Article history:
Received 11 December 2015
Received in revised form
29 February 2016
Accepted 3 August 2016
Communicated by Deng Cai

Keywords:
Large-margin modeling
Maximum-entropy discrimination
Mean-field inference
Latent variable representation
Factor analyzers

ABSTRACT

Devising generative models that allow for inferring low dimensional latent feature representations of high-dimensional observations is a significant problem in statistical machine learning. Factor analysis (FA) is a well-established linear latent variable scheme addressing this problem by modeling the covariances between the elements of multivariate observations under a set of linear assumptions. FA is closely related to principal components analysis (PCA), and might be considered as a generalization of both PCA and its probabilistic version, PPCA. Recently, the invention of Gaussian process latent variable models (GP-LVMs) has given rise to a whole new family of latent variable modeling schemes that generalize FA under a nonparametric Bayesian inference framework. In this work, we examine generalization of FA models under a different Bayesian inference perspective. Specifically, we propose a *large-margin* formulation of FA under the maximum entropy discrimination (MED) framework. The MED framework integrates the large-margin principle with Bayesian posterior inference in an elegant and computationally efficient fashion, allowing to leverage existing high-performance solvers for convex optimization problems. We devise efficient mean-field inference algorithms for our model, and exhibit its advantages by evaluating it in a number of diverse application scenarios, dealing with high-dimensional data classification and reconstruction.

1. Introduction

Factor analysis (FA) is a well-established linear latent variable scheme, modeling the covariances between the elements of multivariate observations by dividing them into two parts: an unobserved systematic part, taken as a linear combination of a relatively small number of unobserved latent variables called *factors*, and an unobserved *error* part, whose elements are considered as uncorrelated [1,2]. Factor analysis is closely related to principal components analysis (PCA) [3], and might be considered as a generalization of both PCA and its probabilistic version, PPCA [3], overcoming their drawbacks which namely are (a) PCA does not correspond to an underlying density function for the data and (b) both PCA and PPCA assume a uniform variation for the components of the feature vectors outside the principal subspace, which in general is a strong and restrictive assumption.

Recently, a considerable amount of work has been devoted to the generalization of FA under a nonparametric Bayesian perspective. This line of research has been motivated by the seminal work on Gaussian process latent variable models (GP-LVMs), first presented in [4]. GP-LVMs can be considered as multiple-output Gaussian process (GP) regression models where only the output data are given. The inputs are unobserved and are treated as latent variables which are optimized in the context of model training. The adoption of this optimization strategy (rather than the straightforward solution of performing inference over these latent variables) is a trick that renders the model tractable; the theoretical grounding for this approach is based on the fact that GP-LVM can be seen as a nonlinear extension of FA and PPCA [4]. Following the ideas presented in [4], several researchers have since proposed a variety of extensions to GP-LVM, including adaptations of the model to allow for modeling data with complex underlying dynamics, e.g. sequential data [5,6], and multimodal extensions of GP-LVMs (e.g., [7]), to name just a few.

In this work, for the first time in the literature, we consider generalization of FA models under a different Bayesian inference perspective. Specifically, we propose a generative latent feature model that leverages the large-margin principle to learn the function mapping the obtained latent representations to the original (observed) data presented to our model. Introduction of the large-margin learning principle allows for obtaining a technique with higher discriminative power, that makes more effective use of our training data during estimation of the postulated latent data representations. To introduce the large-margin principle in the context of our hierarchical Bayesian model, we build upon the maximum entropy discrimination (MED) framework [8,9].

The MED framework integrates the large-margin principle with Bayesian posterior inference in an elegant and computationally efficient fashion, allowing to leverage existing high-performance

E-mail address: sotirios.chatzis@eecei.cut.ac.cy

http://dx.doi.org/10.1016/j.neucom.2016.08.007
0925-2312/

techniques for both hierarchical Bayesian models and convex optimization problems. Adoption of the MED framework in the context of our model yields a mean-field inference algorithm, which obtains a regularized posterior distribution in a feasible space defined by a set of *expected* margin constraints generalized from the familiar support vector regression (SVR)-style margin constraints. We dub our approach maximum entropy discrimination factor analysis (MED-FA).

To examine the effectiveness of our approach, and how it compares to the existing approaches, we perform a number of experimental evaluations; we consider application scenarios from diverse domains, dealing with high-dimensional data classification and reconstruction. The remainder of this paper is organized as follows: in the next section, we introduce our proposed approach, and derive its inference algorithm and its predictive density expression. Further, we perform our experimental evaluations, and examine the advantages of our approach over existing alternatives. Finally, in the concluding section of this paper, we summarize our results and discuss directions for future research.

2. Proposed approach

Let us consider a dataset $Y = \{\boldsymbol{y}_d\}_{d=1}^D$ comprising N-dimensional i.i.d. observations $\boldsymbol{y}_d = [y_{dn}]_{n=1}^N \in \mathbb{R}^N$. Let us also assume that these observations are generated under an FA model of the form

$$p(\boldsymbol{y}_d|\boldsymbol{z}_d; \boldsymbol{H}, \boldsymbol{\Psi}) = \prod_{n=1}^{N} p(y_{dn}|\boldsymbol{z}_d; \boldsymbol{\eta}_n, \Psi_n) \tag{1}$$

where $\boldsymbol{H} = \{\boldsymbol{\eta}_n\}_{n=1}^N$, $\boldsymbol{\eta}_n \in \mathbb{R}^D$, and

$$p(y_{dn}|\boldsymbol{z}_d; \boldsymbol{\eta}_n, \Psi_n) = \mathcal{N}(y_{dn}|\boldsymbol{\eta}_n^T \boldsymbol{z}_d, \Psi_n) \tag{2}$$

In the likelihood function (1) of the postulated model, $\boldsymbol{H}$ is the factor loadings matrix of the postulated model, while $\boldsymbol{\Psi} = \mathrm{diag}([\Psi_n]_{n=1}^N)$ is its *diagonal* noise covariance matrix. On the other hand, the $\boldsymbol{z}_d = [z_{dm}]_{m=1}^M \in \mathbb{R}^M$ are M-dimensional latent representations (factors) of the observed data inferred by our model, with $M < N$. As usual in FA, we elect to impose a spherical prior over the latent factors $\boldsymbol{z}_d$, yielding

$$p(\boldsymbol{z}_d) = \mathcal{N}(\boldsymbol{z}_d|\boldsymbol{0}, \boldsymbol{I}_M) \tag{3}$$

From this starting point, we further stipulate that the postulated linear scheme (1) that connects the obtained latent representations $Z = \{\boldsymbol{z}_d\}_{d=1}^D$ and the observed data $Y = \{\boldsymbol{y}_d\}_{d=1}^D$ be subject to the following large-margin constraints:

$$\begin{cases} y_{dn} - \mathbb{E}[\boldsymbol{\eta}_n^T \boldsymbol{z}_d] \leq \varepsilon + \xi_{dn} \\ -y_{dn} + \mathbb{E}[\boldsymbol{\eta}_n^T \boldsymbol{z}_d] \leq \varepsilon + \xi_{dn}^* \\ \xi_{dn}, \xi_{dn}^* \geq 0 \end{cases} \quad \forall\, d, n \tag{4}$$

Our imposed constraints are inspired from large-margin approaches, and especially, the literature on MED regression models [8, Chapter 4], as they are based on maximization of an *expected*margin, that takes into account the Bayesian inferential formulation of our model.

Note that, in (4), ε is a precision parameter, functioning similar to the precision parameter in SVR, and $\{\xi_{dn}, \xi_{dn}^*\}_{d,n}$ are some slack-variables, used in a way similar to SVR [10]. Note also that our model postulates *soft constraints* (by introducing the slack variables $\{\xi_{dn}, \xi_{dn}^*\}_{d,n}$), so as to allow for better handling *outliers* in the modeled datasets (which are quite common in real-world application scenarios).

Finally, in order to facilitate data-driven selection of the most appropriate dimensionality M of the latent factor vectors $\boldsymbol{z}_d$, we resort to a technique widely referred to as *automatic relevance determination* (*ARD*) [11]. The ARD mechanism is implemented by imposing a hierarchical prior over the factor loadings matrix $\boldsymbol{H}$ to discourage large values, with the width along each latent dimension controlled by a Gamma-distributed precision hyperparameter ϕ_m, $m \in \{1, \ldots, M\}$. If one of these precisions ϕ_m tends to infinity, then the outgoing weights $\eta_{hm}\ \forall\, n$ will have to be very close to zero in order to maintain a high likelihood under this prior. As such, the model ignores the corresponding (e.g., the mth) direction in the latent subspace, which is effectively "switched off" of the model.

On this basis, we impose a *hierarchical* conjugate prior distribution over the factor loadings matrix $\boldsymbol{H}$, as follows: Each row $\boldsymbol{\eta}_n$ of the factor loadings matrix $\boldsymbol{H}$ imposed a zero-mean Gaussian prior with axis-aligned elliptical covariance

$$p(\boldsymbol{\eta}_n|\boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{\eta}_n|\boldsymbol{0}, \mathrm{diag}(\boldsymbol{\phi})^{-1}) \tag{5}$$

where $\boldsymbol{\phi}$ is the precision hyperparameter vector of the prior. In addition, we postulate

$$p(\boldsymbol{\phi}|\bar{\omega}_0, \hat{\omega}_0) = \prod_{m=1}^{M} \mathcal{G}(\phi_m|\bar{\omega}_0, \hat{\omega}_0) \tag{6}$$

where $\bar{\omega}_0$ and $\hat{\omega}_0$ are the shape and inverse-scale hyper-hyperparameters of the Gamma prior imposed on the precision vector $\boldsymbol{\phi}$.

This concludes the formulation of our MED-FA model.

2.1. Inference algorithm

To perform inference for our model, we adopt the MED inference framework. Specifically, conventional MED inference in the context of our model comprises solution of the following minimization problem:

$$\min_{q(Z,\boldsymbol{H},\boldsymbol{\phi}),\boldsymbol{\xi},\boldsymbol{\xi}^*,\boldsymbol{\Psi}} \mathrm{KL}\big(q(Z, \boldsymbol{H}, \boldsymbol{\phi}) \| p(Z, \boldsymbol{H}, \boldsymbol{\phi})\big) + \gamma \sum_{d=1}^{D}\sum_{n=1}^{N} (\xi_{dn} + \xi_{dn}^*) \tag{7}$$

under the constraints (4), where $\boldsymbol{\xi} = [\xi_{dn}]_{d,n}$, $\boldsymbol{\xi}^* = [\xi_{dn}^*]_{d,n}$, and γ is a regularization constant. However, in our work, we elect to optimize a composite objective function that also takes into consideration the *expected* (negative) log-likelihood of our hierarchical Bayesian model, which measures the goodness of fit to the training data, similar to [12]. This way, inference for our model eventually reduces to solution of the following problem:

$$\begin{aligned} &\min_{q(Z,\boldsymbol{H},\boldsymbol{\phi}),\boldsymbol{\xi},\boldsymbol{\xi}^*,\boldsymbol{\Psi}} \mathrm{KL}\big(q(Z, \boldsymbol{H}, \boldsymbol{\phi}) \| p(Z, \boldsymbol{H}, \boldsymbol{\phi})\big) \\ &\quad + \gamma \sum_{d=1}^{D}\sum_{n=1}^{N} (\xi_{dn} + \xi_{dn}^*) - \sum_{d=1}^{D} \mathbb{E}[\log p(\boldsymbol{y}_d|\boldsymbol{z}_d)] \\ &\forall\, d, n,\ \text{s. t.}: \begin{cases} y_{dn} - \mathbb{E}[\boldsymbol{\eta}_n^T \boldsymbol{z}_d] \leq \varepsilon + \xi_{dn} \\ -y_{dn} + \mathbb{E}[\boldsymbol{\eta}_n^T \boldsymbol{z}_d] \leq \varepsilon + \xi_{dn}^* \\ \xi_{dn}, \xi_{dn}^* \geq 0 \end{cases} \end{aligned} \tag{8}$$

Note that, in the above expressions, all the expectations $\mathbb{E}[\cdot]$ are computed w.r.t. the posterior $q(Z, \boldsymbol{H}, \boldsymbol{\phi})$.

Our inference algorithm proceeds in an iterative fashion, under the mean-field principle [13]: on each iteration, we consecutively minimize (8) over each one of the factors of the sought posterior $q(Z, \boldsymbol{H}, \boldsymbol{\phi})$, as well as the noise covariance $\boldsymbol{\Psi}$, and the slack-variables $\boldsymbol{\xi}, \boldsymbol{\xi}^*$, one at a time, holding the others fixed. It has been shown that such an iterative consecutive updating procedure is guaranteed to monotonically optimize the objective function of our problem [9]. Under this procedure, the posterior over the (rows of the) factor loadings matrix yields

Please cite this article as: S.P. Chatzis, Maximum entropy discrimination factor analyzers, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.08.007