Neural Networks 23 (2010) 226–238

Contents lists available at ScienceDirect

Neural Networks

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

Robust extraction of local structures by the minimum β-divergence method

Md. Nurul Haque Mollah [a,b,*], Nayeema Sultana [c], Mihoko Minami [e], Shinto Eguchi [a,d]

[a] The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
[b] The Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh
[c] Department of Statistics, Islamia College, Binodpur, Rajshahi-6206, Bangladesh
[d] The Graduate University for Advanced Studies, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
[e] Department of Mathematics, Keio University, Yokohama, Kanagawa 223-8522, Japan

ARTICLE INFO

Article history:
Received 12 April 2007
Accepted 12 November 2009

Keywords:
Local PCA
β-divergence
Initialization of the parameters
Adaptive selection for the tuning parameter
Cross validation
Sequential estimation

ABSTRACT

This paper discusses a new highly robust learning algorithm for exploring local principal component analysis (PCA) structures in which an observed data follow one of several heterogeneous PCA models. The proposed method is formulated by minimizing β-divergence. It searches a local PCA structure based on an initial location of the shifting parameter and a value for the tuning parameter β. If the initial choice of the shifting parameter belongs to a data cluster, then the proposed method detects the local PCA structure of that data cluster, ignoring data in other clusters as outliers. We discuss the selection procedures for the tuning parameter β and the initial value of the shifting parameter $\boldsymbol{\mu}$ in this article. We demonstrate the performance of the proposed method by simulation. Finally, we compare the proposed method with a method based on a finite mixture model.

1. Introduction

Principal component analysis (PCA) is one of the most popular technique for processing, compressing and visualizing multivariate data. It is widely used for reducing dimensionality of multivariate data (Jolliffe, 2002). In general, PCA aims to extract the most informative q-dimensional output vector $\boldsymbol{y}(t)$ from input vector $\boldsymbol{x}(t)$ of dimension m, which is achieved by obtaining the $m \times q$ orthogonal matrix Γ (i.e. $\Gamma^{\mathrm{T}}\Gamma = I_q$, identity matrix). Thus Γ linealy relates $\boldsymbol{x}(t)$ to $\boldsymbol{y}(t)$ by

$$\boldsymbol{y}(t) = \Gamma^{\mathrm{T}}\Big(\boldsymbol{x}(t) - \boldsymbol{\mu}\Big), \quad t = 1, 2, \ldots, n \tag{1}$$

such that components of $\boldsymbol{y}(t)$ are mutually uncorrelated, satisfying the order of the variances according to the component number of $\boldsymbol{y}(t)$. In the context of off-line learning Γ and $\boldsymbol{\mu}$ are directly obtained as the q dominant eigenvectors of the sample covariance matrix and the sample mean vector. The classical PCA is characterized by minimizing the empirical loss function

$$\frac{1}{n}\sum_{t=1}^{n} z\Big(\boldsymbol{x}(t), \boldsymbol{\mu}, \Gamma\Big) \tag{2}$$

with respect to $\boldsymbol{\mu}$ and Γ, where

$$z(\boldsymbol{x}, \boldsymbol{\mu}, \Gamma) = \frac{1}{2}\Big\{\|\boldsymbol{x} - \boldsymbol{\mu}\|^2 - \|\Gamma^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{\mu})\|^2\Big\} \tag{3}$$

or half the squared residual distance of $\boldsymbol{x} - \boldsymbol{\mu}$ projected onto the subspace spanned by the columns of Γ (Hotelling, 1933). Higuchi and Eguchi (2004) proposed a variant of this classical procedure for robust PCA by minimizing the empirical loss function

$$L_{\Psi}(\boldsymbol{\mu}, \Gamma) = \frac{1}{n}\sum_{t=1}^{n} \Psi\Big(z\Big(\boldsymbol{x}(t), \boldsymbol{\mu}, \Gamma\Big)\Big) \tag{4}$$

where $\Psi(z)$ is assumed to be a monotonically increasing. Various choices of Ψs yield various procedures for PCA including the identity function $\Psi_0(z) = z$ as the classical PCA and the sigmoid function as the self-organizing rule, cf. (Xu & Yuille, 1995). In general, Ψ is interpreted as a generic function to give the loss function L_{Ψ}. The minimization of L_{Ψ} in Eq. (4) is referred as *minimum psi principle generated by* Ψ. Based on an argument similar to that of the classical PCA, Higuchi and Eguchi (2004) showed that the minimizer of $L_{\Psi}(\boldsymbol{\mu}, \Gamma)$ satisfies the stationary equation system for $\boldsymbol{\mu}$ and Γ. In neural networks, Γ is interpreted as the coefficient matrix connecting m neurons to q neurons, where a learning process works by off-line renewal of Γ based on a batch of input vectors or on-line renewal of Γ based on sequential input vectors (Amari, 1977; Haykin, 1999; Oja, 1989). See also Croux and Haesbroeck (2000) and Campbell (1980) for robust PCA methods. All PCA algorithms mentioned above are well discussed and

* Corresponding author at: The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan. Tel.: +81 3 5421 8728.
E-mail addresses: nhmollah@ism.ac.jp (Md. Nurul Haque Mollah), nasultana@yahoo.com (N. Sultana), mminami@ism.ac.jp (M. Minami), eguchi@ism.ac.jp (S. Eguchi).

0893-6080/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.11.011

established in a context in which the data distribution is uni-modal, that is, there is only one data center in the entire data space.

In the case of multi-modal distribution, the performance of the PCA algorithms as early discussed are not so good. In this aspects, several interesting algorithms for local dimensionality reduction have been proposed. As for example, mixtures of PCA and mixtures of factor analysers proposed by Hinton, Dayan and Revow (1997), VQPCA (Vector-Quantization PCA) algorithm of Kambhatla and Leen (1997), mixtures of PPCA algorithm of Tipping and Bishop (1999), a nonlinear neural network model of mixture of local PCA proposed by Zhang, Fu and Yan (2001), resolution-based complexity control for Gaussian mixture models of Meinicke and Ritter (2001), automated hierarchical mixtures of PPCA algorithm of Su and Dy (2004) and an extension of neural gas to local PCA proposed by Möller and Hoffmann (2004). However, when applying any one of these algorithms, one may encounter a difficult problem that the number of data clusters in the entire data space should be known in advance. To overcome such problems for local dimensionality reduction, there exist some alternative ideas which includes variational inference for Bayesian mixtures of factor analysers proposed by Ghahramani and Beal (2000), unsupervised learning of finite mixture models suggested by Figueiredo and Jain (2002) and accelerated variational Dirichlet mixture models of Kurihara, Welling and Vlassis (2006). Anyway, these type of algorithms may gives misleading results in presence of outliers (Hampel, Ronchetti, Rousseeuw & Stahel, 1986). In this aspect, Ridder and France (2003) offered robust algorithm based on mixture of PPCA using t-distributions. However, one major problem in this algorithm is that it needs number of data clusters in advance. Therefore some researchers or users may expect a highly robust algorithm against outliers which does not require number of data clusters in advance.

In this paper we propose a new highly robust algorithm for exploring local PCA structures by minimizing β-divergence in a situation where we do not know whether the data distribution is uni-modal or multi-modal. The key idea is the use of a super robust PCA algorithm with a volume adjustment based on β-divergence, in which it properly detects one data cluster by ignoring all data in other clusters as outliers. See Higuchi and Eguchi (1998, 2004), Kamiya and Eguchi (2001) and Mollah, Minami and Eguchi (2007) for the robust procedures. The proposed method has a close link with the mixture ICA method proposed by Mollah, Minami and Eguchi (2006). See also Lee, Lewicki and Sejnowski (2000) for model based mixtures of ICA models. We introduce the β-divergence satisfying a condition of volume matching which naturally defines the learning algorithm for both uni-modal and multi-modal distributions. Also the behavior of the expected loss function based on the β-divergence in a context of multi-modal distributions is investigated. Thus the performance of the proposed learning algorithm beyond robustness is viewed. The proposed learning algorithm is based on an empirical loss function

$$L_\beta(\boldsymbol{\mu}, V) = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{\beta}\Big[1 - \det(V)^{-\frac{1}{2}\frac{\beta}{\beta+1}}\exp\{-\beta w(\boldsymbol{x}(t), \boldsymbol{\mu}, V)\}\Big] \tag{5}$$

where V is a variance matrix and

$$w(\boldsymbol{x}, \boldsymbol{\mu}, V) = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}V^{-1}(\boldsymbol{x}-\boldsymbol{\mu}). \tag{6}$$

Thus the shifting parameter $\boldsymbol{\mu}$ is defined by the minimizer of the loss function (5) with respect to $\boldsymbol{\mu}$; the connection matrix Γ is defined by eigen-decomposition of the minimizer of the loss function (5) with respect to V. The loss function $L_\beta(\boldsymbol{\mu}, V)$ is closely connected with minimum psi-principle if we choose $\Psi_\beta(z) = \{1 - \exp(-\beta z)\}/\beta$. The main difference from the minimum Ψ-principle is that the loss function is defined by a function of V and w in place of Γ and z in Eq. (4). It may suggest one of robust procedures for PCA by direct application of the discussion in Higuchi and Eguchi (2004). Furthermore, we will show that the loss function $L_\beta(\boldsymbol{\mu}, V)$ satisfies a remarkable property beyond robustness as follows. Let us consider a probabilistic situation in which the data distribution is J-modal. Then the dataset $\mathscr{D} = \{\boldsymbol{x}(t) : t = 1, \ldots, n\}$ in $\mathbf{R}^m$ can be decomposed into J mutually disjoint subsets $\{\mathscr{D}_j : j = 1, \ldots, J\}$ such that jth local mode along with mean vector $\boldsymbol{\mu}_j$ and variance matrix V_j occurs in

$$\mathscr{D}_j = \{\boldsymbol{x}(t) : t \in T_j\} \tag{7}$$

with partitioned index sets $\{T_j : j = 1, \ldots, J\}$. Then we will show that the proposed learning algorithm by minimizing $L_\beta(\boldsymbol{\mu}, V)$ can be extract $\boldsymbol{\mu}_j$ and V_j if it start from an initial point $\boldsymbol{\mu} \in \mathscr{D}_j$ and appropriately chosen V. At this time, under a Gaussian mixture density $p(\boldsymbol{x}) = \sum_{j=1}^{J}\pi_j\varphi(\boldsymbol{x}, \boldsymbol{\mu}_j, V_j)$, we will observe that

$$\begin{aligned}(\boldsymbol{\mu}_j, V_j) &= \operatorname*{argmin}_{(\boldsymbol{\mu},V)\in\mathscr{D}_j\times\mathscr{S}_m} L_\beta(\boldsymbol{\mu}, V; p)\\ &= \operatorname*{argmin}_{(\boldsymbol{\mu},V)\in\mathscr{D}_j\times\mathscr{S}_m} L_\beta(\boldsymbol{\mu}, V), \quad (j = 1, 2, \ldots, J),\end{aligned} \tag{8}$$

where $\mathscr{S}_m$ denotes the space of all the symmetric, positive-definite matrices of order m. Thus minimization of $L_\beta(\boldsymbol{\mu}, V)$ with respect to $(\boldsymbol{\mu}, V) \in (\mathscr{D}_j \times \mathscr{S}_m)$ offers J local minima $\{(\boldsymbol{\mu}_j, V_j); j = 1, 2, \ldots, J\}$ for J-modal data distribution.

Section 2 describes the new proposal for local PCA by minimizing β-divergence. In Section 3, we discuss the consistency of the proposed method for local PCA. Section 4 discuss the proposed learning algorithm. In Section 5, we discuss the adaptive selection procedure for the tuning parameter. Simulation and discussion is given in Section 6. Finally, Section 7 presents the conclusions of this study.

2. Local principal component analysis

Let $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ be probability density functions on a data space in $\mathbf{R}^m$. The β-divergence of $p(\boldsymbol{x})$ with $q(\boldsymbol{x})$ is defined as

$$\begin{aligned}D_\beta(p, q) = \int\Bigg[&\frac{1}{\beta}\left\{p^\beta(\boldsymbol{x}) - q^\beta(\boldsymbol{x})\right\}p(\boldsymbol{x})\\ &- \frac{1}{\beta+1}\left\{p^{\beta+1}(\boldsymbol{x}) - q^{\beta+1}(\boldsymbol{x})\right\}\Bigg]\mathrm{d}\boldsymbol{x}, \quad \text{for } \beta > 0\end{aligned}$$

which is non-negative, that is $D_\beta(p, q) \geq 0$, equality holds if and only if $p(\boldsymbol{x}) = q(\boldsymbol{x})$ for almost all $\boldsymbol{x}$ in $\mathbf{R}^m$, see Basu, Harris, Hjort and Jones (1998) and Minami and Eguchi (2002). We note that β-divergence reduces to Kullback–Leibler (KL) divergence when we take a limit of the tuning parameter β to 0 as

$$\begin{aligned}\lim_{\beta\downarrow 0} D_\beta(p, q) &= \int p(\boldsymbol{x})\log\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}\\ &= D_{\mathrm{KL}}(p, q).\end{aligned}$$

Let $p(\boldsymbol{x})$ be the density function of data distribution of $\boldsymbol{x}$. Then the minimum β-divergence method is defined by

$$\min_{q\in M} D_\beta(p, q),$$

where M denotes a statistical model. Let us consider a kind of volume match by

$$\begin{aligned}D^*_\beta(p, q) &= \min_\kappa D_\beta(p, \kappa q)\\ &= \frac{1}{\beta(\beta+1)}\left[\int p^{\beta+1}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \frac{\left\{\int p(\boldsymbol{x})q^\beta(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right\}^{\beta+1}}{\left\{\int q^{\beta+1}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right\}^\beta}\right].\end{aligned} \tag{9}$$

We note that for a fixed data density p the functional $D^*_\beta(p, \cdot)$ is defined on the space of nonnegative functions with a finite mass and that $D^*_\beta(p, \kappa q) = D^*_\beta(p, q)$ for any positive scalar κ. If we neglect terms depending only on p, then we get the term