Pattern Recognition 60 (2016) 681–691

Contents lists available at ScienceDirect

Pattern Recognition

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

# Sparse exponential family Principal Component Analysis

Meng Lu [a,b,*], Jianhua Z. Huang [c], Xiaoning Qian [b]

[a] Institute of Data Science, Tianjin University, Tianjin, China
[b] Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77840, USA
[c] Department of Statistics, Texas A&M University, College Station, TX 77840, USA

ARTICLE INFO

*Article history:*
Received 15 January 2016
Received in revised form
9 May 2016
Accepted 19 May 2016
Available online 21 May 2016

*Keywords:*
Dimension reduction
Sparsity
Exponential family principal component analysis

ABSTRACT

We propose a Sparse exponential family Principal Component Analysis (SePCA) method suitable for any type of data following exponential family distributions to achieve simultaneous dimension reduction and variable selection for better interpretation of the results. Because of the generality of exponential family distributions, the method can be applied to a wide range of applications, in particular when analyzing high dimensional next-generation sequencing data and genetic mutation data in genomics. The use of sparsity-inducing penalty helps produce sparse principal component loading vectors such that the principal components can focus on informative variables. By using an equivalent dual form of the formulated optimization problem for SePCA, we derive optimal solutions with efficient iterative closed-form updating rules. The results from both simulation experiments and real-world applications have demonstrated the superiority of our SePCA in reconstruction accuracy and computational efficiency over traditional exponential family PCA (ePCA), the existing Sparse PCA (SPCA) and Sparse Logistic PCA (SLPCA) algorithms.

## 1. Introduction

Dimension reduction methods are widely used for many data analytic applications such as computer vision, data mining, and bioinformatics. In addition to deriving low dimensional projections for model complexity reduction and reproducibility of learning results, people often would like to know the physical meanings of the original variables and how they contribute to these projections. For example, when analyzing images, it is of much interest to know which image regions are crucial to represent or capture the essential information contained in the given images. Identifying variables expressing the maximum data variation will also be of much interest for next-generation sequencing data analysis since it would help greatly reduce the profiling cost for biomarker discovery. To achieve these goals in diverse real-world applications, one faces two critical challenges: how to handle diverse data types arising from different applications and how to obtain meaningful interpretation of analysis results. Exponential family PCA (ePCA) methods [1–3] and Sparse PCA (SPCA) methods [4–7] are well known to address these two issues separately. However, to the best of our knowledge, it seems that no one has proposed a method to address these two issues simultaneously. In this paper, we propose a Sparse exponential family PCA (SePCA) method for dimension reduction with both the capability of addressing the interpretation issue and the generality of applications to any type of data following exponential family distributions.

For probabilistic interpretation, ePCA aims to find the low-dimensional projections of a set of canonical parameters that maximize the likelihood of the observed data. PCA is a special case of ePCA by assuming that the conditional probability of each data point given its corresponding canonical parameters—mean vectors—follows a Gaussian distribution. By extending the Gaussian distribution of the conditional probability to other members of the exponential family, ePCA naturally generalizes PCA to be more suitable to handle various data types, including binary and count data for example, other than continuous data that is often assumed to follow a Gaussian distribution. To achieve better performance for dimension reduction, an appropriate assumption of the distribution for a given certain data type is desirable. For example, Bernoulli distribution is appropriate for binary data, multinomial distribution for categorical data, and Poisson distribution for count data. This generalization of PCA is analogous to the extension of ordinary linear regression models to Generalized Linear Models (GLM) [8]. However, ePCA still suffers from the interpretation issue of PCA, which motivates us to derive SePCA by introducing sparsity regularization to the loading vectors. The sparsity could be achieved by adding a regularization term on loading vectors to the objective function of ePCA, which is similar to the way of formulating sparsity regularized GLMs.

* Corresponding author.
*E-mail address:* lvmeng0502@gmail.com (M. Lu).

http://dx.doi.org/10.1016/j.patcog.2016.05.024
0031-3203/© 2016 Elsevier Ltd. All rights reserved.

Finding an efficient algorithm to solve SePCA is challenging due to a couple of mixed difficulties from SPCA and ePCA from the involved non-convex constraints and the non-convex complex objective function of multiple variables. It has always been challenging to solve either ePCA or SPCA, not to mention solving them simultaneously in SePCA. Several approaches have been proposed to solve either ePCA or SPCA [1,3–6,9,10]. A generic algorithm directly optimizes the objective function of ePCA [1,2]; however, it is inefficient to deal with non-quadratic and complex likelihood functions for many exponential family members. An alternative effective way is to optimize an auxiliary function of the log-likelihood to achieve the solutions by approximating the objective function with its lower bound. Majorization–Minimization (MM) algorithm [11,12] has been applied to efficiently solve Sparse Logistic PCA (SLPCA) for binary data with closed-form updating rules [9]. However, it is hard to find appropriate auxiliary functions that can lead to efficient closed-form updating rules for other members of the exponential family such as multinomial distribution and Poisson distribution. Recently, Guo and Schuurmans [3] have proposed an efficient algorithm to solve ePCA by transforming the regularized primal problem to an equivalent dual problem with the optimal solution found at the stationary point. Following this idea, we transform the SePCA problem to a dual problem, in which the objective function with respect to the Principal Components (PCs) and the principal component loading vectors has a similar form as the sparse PCA problem formulated by Shen and Huang (sPCA_rSVD) [6]. We can then solve the SePCA problem by alternately updating unknown variables using efficient closed-form updating rules that lead to favorable computational efficiency.

The rest of the paper is organized as follows. Section 2 briefly reviews classical PCA in a probabilistic modeling framework, from which it could be naturally extended to ePCA. We also introduce the SPCA problem and the algorithm for solving it at the end of this section. Section 3 describes the formulation of SePCA, with an efficient alternative updating algorithm to solve it. The computational complexity analysis of the solution algorithm is also provided. Section 4 illustrates the performance of SePCA compared with Zou's SPCA [5] and a previous SLPCA method [9] via the experiments on both simulated and real-world data. Section 5 concludes the paper and discusses our future directions.

## 2. Related work

In this section, we review relevant concepts and probabilistic models that form the foundations of SePCA. We introduce PCA from a probabilistic modeling perspective and naturally extend it to the exponential family. From this point of view, PCA is formulated as a Maximum-Likelihood Estimation (MLE) problem, which estimates the low-dimensional projections of a set of canonical parameters by assuming that the conditional probability of each data point given its canonical parameters follows a Gaussian distribution [13]. Similarly, the ePCA tailored to some other types of data could also be modeled as such a MLE problem by assuming that the conditional probability follows a corresponding distribution in the exponential family other than Gaussian. To give a flavor of SePCA, we introduce SPCA as a simple case and discuss an efficient strategy to solve it at the end of this section.

### 2.1. Principal component analysis

Given a set of samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^D$, PCA projects the data into a principal-component subspace with a lower dimension $L(\leq D)$ and meanwhile attempts to preserve the maximum data variation. An alternative interpretation of PCA from a probabilistic perspective assumes that the data points are approximated by linear projections of low-dimensional latent variables plus a Gaussian noise. For each sample $\boldsymbol{x}_n$ $(1 \leq n \leq N)$, given its corresponding vector of latent variables $\boldsymbol{z}_n$ that lies in the principal-component subspace, we assume

$$\boldsymbol{x}_n = W^T\boldsymbol{z}_n + \boldsymbol{b} + \epsilon,$$

where $W$ is a principal loading matrix whose rows span the principal-component subspace; $\boldsymbol{b}$ is a bias vector and $\epsilon$ follows a Gaussian distribution $N(0, \sigma^2 I)$. Assuming a vector of canonical parameters $\boldsymbol{\theta}_n = W^T\boldsymbol{z}_n + \boldsymbol{b}$, the conditional probability of $\boldsymbol{x}_n$ given $\boldsymbol{\theta}_n$ is then represented as:

$$p(\boldsymbol{x}_n|\boldsymbol{\theta}_n) \sim N(\boldsymbol{x}_n|\boldsymbol{\theta}_n, \sigma^2 I)$$

and the conditional probability of $\boldsymbol{x}_n$ given $\boldsymbol{z}_n$ is:

$$p(\boldsymbol{x}_n|\boldsymbol{z}_n) \sim N(\boldsymbol{x}_n|W^T\boldsymbol{z}_n + \boldsymbol{b}, \sigma^2 I).$$

PCA is then formulated as an optimization problem of maximizing the log-likelihood of the data set with respect to $\boldsymbol{z}_n$, $W$, and $\boldsymbol{b}$, where the objective function is:

$$\sum_n - \| \boldsymbol{x}_n - (W^T\boldsymbol{z}_n + \boldsymbol{b})\|^2 \quad \text{s. t.} \;\; WW^T = I \tag{1}$$

up to a constant. Obviously, this problem is equivalent to minimizing the sum of Euclidean distances from data points to their projections in the principal-component subspace, which is exactly the other interpretation of PCA [14].

### 2.2. Exponential family PCA

From a probabilistic perspective, it is natural to generalize PCA to the exponential family. In the exponential family, a probabilistic latent variable model representing the conditional distribution of a data sample $\boldsymbol{x}_n$ has such a general form [1]:

$$p(\boldsymbol{x}_n|\boldsymbol{\theta}_n) = \exp(\boldsymbol{\theta}_n^T\boldsymbol{x}_n + \log q(\boldsymbol{x}_n) - A(\boldsymbol{\theta}_n)), \tag{2}$$

where $\boldsymbol{\theta}_n$ denotes the corresponding canonical parameters corresponding to the sample $\boldsymbol{x}_n$. $A(\boldsymbol{x}_n)$ is the log-normalization factor with the form based on the base measure $q(\boldsymbol{x}_n)$: $\log \int \exp(\boldsymbol{\theta}_n^T\boldsymbol{x}_n)q(\boldsymbol{x}_n)d\boldsymbol{x}_n$, which ensures that the sum of the conditional probabilities over the domain of $\boldsymbol{x}_n$ equals 1. The probability distribution functions for the members in the exponential family are mainly differentiated by the form of $A(\cdot)$ function. Consequently, the data log-likelihood with respect to the canonical parameters may be of a quadratic form (for Gaussian) or not (for others). Taking Gaussian for instance, $A(\boldsymbol{\theta}_n)$ takes a form of $\boldsymbol{\theta}_n^2/2$ to ensure a Gaussian distribution function. Then, its data log-likelihood function given $\boldsymbol{\theta}$ is equivalent to

$$\sum_n - \| \boldsymbol{x}_n - \boldsymbol{\theta}_n \|^2 \tag{3}$$

up to a constant. The canonical parameters $\boldsymbol{\theta}_n$ are further parameterized with a form of $W^T\boldsymbol{z}_n + \boldsymbol{b}$ using lower-dimensional latent variables $\boldsymbol{z}_n$, principal loading matrix $W$ and a bias vector $\boldsymbol{b}$ for dimension reduction. After substituting $\boldsymbol{\theta}_n$ into (3), we arrive at (1), which is exactly the objective function of PCA derived by MLE.

In general, ePCA can be achieved by maximizing the generalized likelihood based on a general form of the probability function shown in (2). After substituting $\boldsymbol{\theta}_n$ by $\boldsymbol{z}_n$, $W$, and $\boldsymbol{b}$, ePCA is then formulated as the following problem:

$$\min_{Z,\boldsymbol{b}} \min_{W:WW^T=I} \sum_n A(W^T\boldsymbol{z}_n + \boldsymbol{b}) - \mathrm{tr}((ZW + \boldsymbol{1}\boldsymbol{b}^T)X^T), \tag{4}$$

where $Z$ is the $N \times L$ principal component score matrix whose $n$-th row is $\boldsymbol{z}_n$. A probabilistic graphical model to illustrate ePCA is shown in Fig. 1. Note that the principal component subspace is