Laplace Exponential Family PCA

Fangqi Li, Xudie Ren

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University,
800 Dongchuan Road, Min Hang District, Shanghai 200240,P.R. China
{solour_lfq}@sjtu.edu.cn

Abstract. Considering numerous types of data, this paper discusses application of PCA to exponential family distributions. Reviewing the probabilistic basis of PCA, we propose a model using Laplace approximation, which was widely used in classification context, Laplace exponential family PCA (LePCA). The proposed approach provides a more probabilistic solution compared with numerous models before. Standard EM algorithm can be applied to this model, while only a degraded form of EM is applicable on previous exponential PCA models. LePCA absorbs probabilistic PCA, as well as the traditional PCA as its specialization by taking the Gaussian assumption for granted.

Keywords: machine learning, Laplace approximation, principal component analysis, EM algorithm

1 Introduction

Principal component analysis (PCA) [1] is a widely used algorithm for dimension-reduction. Intuitively, it looks for a low-dimensional subspace and represents the original data by their projection to a set of orthonormal basis in it. The projected data have a large variance, i.e. a relatively large amount of information is saved [2]. From a probabilistic perspective, we assume that there is a low-dimensional latent variable $\mathbf{z}$ subject to identity Gaussian. We further assume that one data term $\mathbf{x}$ is sampled from a linear Gaussian relationship. When the coefficient for covariance matrix is set to zero, the model degrades into a deterministic algorithm equal to the traditional PCA, the general latent variable approach is known as probabilistic principal component analysis (PPCA) [3]. PPCA is a special form of factor analysis, with traditional PCA as its specialization. It is the Gaussian linear hypothesis that gives rise to measurement of information by variance. But this hypothesis that data items are subject to Gaussian distribution is sometimes inappropriate. One generali-zation is replacing the Gaussian distribution with the more general exponential family distributions [4], which is known as exponential PCA (EPCA) [5].

Like all probabilistic models, PPCA needs to cope with the over-fitting caused by maximum likelihood estimation. It is naturally to introduce a-posteriori estimation of parameters on PPCA. Similarly, EPCA can also perform MAP as well. However, the prior distribution of the general exponential distribution is not normally distributed, which deprives the posterior estimation of a closed form. Hence the optimization of

adfa, p. 1, 2011.

parameters for EPCA is usually done through an iterative approach. It is common to relate this procedure of optimization to expectation-maximum (EM) algorithm as long as they both aim at maximizing the likelihood for a model with latent variables.

Another problem with traditional PCA as well as PPCA is the dimensionality of latent variables. The dimensionality to which data are reduced to generally needs to be given beforehand, however, it is more reasonable to learn this parameter from data. Works have been done to solve the problem by applying Automatic Relevance Determination (ARD) [6] procedure on PPCA as well as EPCA, but in the second case, the conjugate nature of parameter subspace is lost.

In this paper, we concentrate on giving a closed form for PCA extended to exponential family distributions by making use of Laplace approximation as in [7], chapter 8.4.1. Laplace exponential family PCA (LePCA). We will show how parameters can be estimated by using straightforward EM. The relationship between our method and other ones for parameter estimation in exponential PCA before is showed as well. We further illustrate that it is natural and foresightful to assume a Gaussian hypothesis for PPCA instead of other emission distributions. It is also showed that using Laplace approximation enables this model to generalize factor analysis (FA) [8] by introducing a variable Hessian.

The rest of paper is organized as follows: Section two reviews the generalization process of PCA. Section three introduces the proposed Laplace exponential family PCA. Experiment and illustration example are given in Section four. Section five concludes the paper.

2 Related Work

2.1 Probabilistic PCA

The probabilistic perspective of principal component analysis was introduced in [3]. In PPCA, there is a latent variable $\mathbf{z}$ with fewer dimensionality $q < d$ .Where d denotes the dimensionality of data $\mathbf{x}$. After transformation, it forms the mean for the observed variables. The probabilistic graphical model for PPCA, together with its Bayesian version [9] is given in **Fig.1**, N denotes the size of training set.

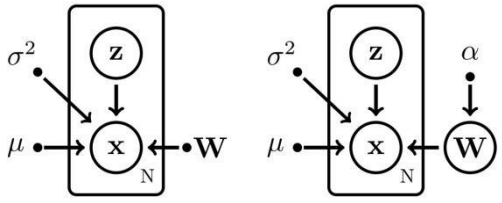

Fig. 1. PGM for PPCA and Bayesian PPCA

For ordinary PPCA, we use point estimation on $\boldsymbol{\mu}$, σ^2 and $\mathbf{W}$, where we have assumed a linear Gaussian relationship:

$$\mathrm{p}(\mathbf{z}) = \mathrm{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I}_{\mathrm{q}}),\ \mathrm{p}(\mathbf{x}|\mathbf{z}) = \mathrm{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2 \mathbf{I}_d)$$

$$\mathrm{p}(\mathbf{x}) = \int \mathrm{p}(\mathbf{x} \mid \mathbf{z})\mathrm{p}(\mathbf{z})\mathrm{d}\mathbf{z} = \mathrm{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2 \mathbf{I}_d)$$

The ARD prior takes the form:

$$\mathrm{p}(\mathbf{W} \mid \boldsymbol{\alpha}) = \prod_{i=1}^{q} (\frac{\alpha_i}{2\pi})^{\frac{d}{2}} \exp\{-\frac{\alpha_i}{2} \mathbf{w}_i^{\mathrm{T}} \mathbf{w}_{\mathrm{i}}\}$$

Thanks to the linear Gaussian relationship, the parameter's update has a closed form given in [3].

PPCA provides a probabilistic approach of dimension reduction from d to q by identifying the most appropriate linear transformation introduced by loading matrix $\mathbf{W}$. Original data $\mathbf{x}$ is encoded by its correlated latent variable, which subjects to $\mathrm{p}(\mathbf{z} \mid \mathbf{x})$. This posterior distribution has a closed form as long as the correlation are Gaussian.

One problem of PPCA is that q needs to be determined beforehand. In [9], an algorithm that can learn an appropriate q given data was proposed by introducing an ARD prior on loading matrix $\mathbf{W}$. A sparse solution can be obtained by learning hyper-parameter $\boldsymbol{\alpha}$.

2.2 Generalization of PCA to Exponential Family

The distribution of $\mathbf{x}$ introduced by PPCA is a Gaussian. However, there are numerous data with a non-Gaussian distribution. In order to applying the technique of PCA onto other data, a generalization of PCA to exponential family distributions was proposed [5]. It uses the same idea of looking for a subspace that reduces the dimension while preserves information to its best.

Under a Gaussian context, fitting the likelihood is equal to minimizing a loss function in a quadratic form, hence a least-square target. For general exponential family, the loss function is given by the Bregman distance. Thanks to the property of exponential family, this optimization task is convex with respect to two independent parameters. To estimate parameters, [5] gave an iterative solution during which $\mathbf{W}$ and $\mathbf{z}$ are optimized alternately. EPCA' s posterior form was studied by [10], where a MAP estimation for parameters are drawn through hybrid Monte Carlo given prior distribution for $\mathbf{W}$ and $\mathbf{z}$ .The graphical model for EPCA is shown in **Fig.2**.

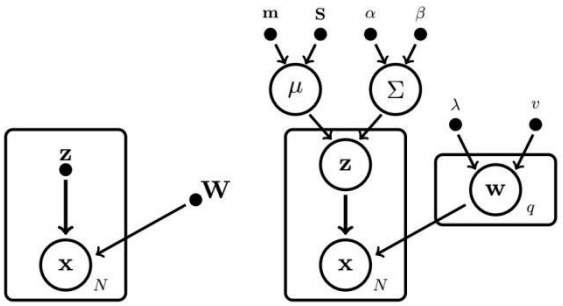

Fig. 2. PGM for EPCA and its Bayesian version, BXPCA

And we have:

$$p(\mathbf{x} \mid \mathbf{z}, \mathbf{W}) = p_e(\mathbf{x} \mid \mathbf{W}\mathbf{z})$$

EPCA is different from PPCA in their ways of identifying conditional probability $p(\mathbf{x} \mid \mathbf{z}, \mathbf{W})$. PPCA uses a Gaussian distribution while EPCA uses a context-related exponential family distribution denoted by p_e .

2.3 Simple Exponential Family PCA (SePCA)

As PPCA, EPCA faces the problem of finding the best dimensionality for latent variable, namely the number of principal components. One of the solutions has been proposed in [11], with a graphical model as **Fig.3**.

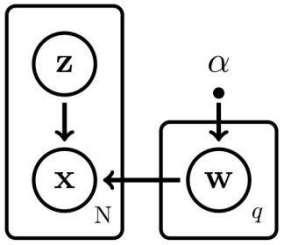

Fig. 3. PGM for SePCA

By using an ARD prior on **W** as in Bayesian PPCA, dimensionality is driven to its optimum by learning hyper-parameter $\boldsymbol{\alpha}$. It is noticeable that the PGM for SePCA is the same to that of Bayesian PPCA as shown in Fig.1, after removing $\boldsymbol{\mu}$ and σ^2 .

On parameter estimation, SePCA uses an iterative method similar to that used in EPCA. It was interpreted in [11] that this iterative optimization on two set of parameters alternatively can be seen as a degenerated EM.

3 Proposed Model

In this section, we begin with reviewing some basic properties of exponential family distributions, followed by introducing the proposed model. By sticking to a PGM language, we list the theoretical improvements and some comparisons. On elaborating the method for parameter estimation, we elucidate the feature of proposed model further.

3.1 Exponential Family

An exponential family distribution takes the form:

$$\mathrm{p}(\mathbf{x} \mid \boldsymbol{\theta}) = \exp\{\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta} + h(\mathbf{x}) - G(\boldsymbol{\theta})\}$$

Where we work on the sufficient statistics. The distribution's cumulant function $G(\boldsymbol{\theta})$ is always a convex one. And $h(\mathbf{x})$ is a scaling factor which often set to one. It is straightforward to recognize that a lot of widely-used distributions belongs to exponential family. Typical examples are Gaussian, Bernoulli, multinoulli, etc.

One significance for exponential family is that it is the only family with conjugate prior distributions, which can simplify the estimation as well as increase the interpretability.

The conjugate prior for an exponential family distribution takes the form:

$$\mathrm{p}(\boldsymbol{\theta} \mid \nu, \boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{\lambda}, \nu)} \exp\{\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\lambda} - \nu \cdot G(\boldsymbol{\theta})\}$$

From the perspective of observation, this prior is tantamount to taking ν prior data terms with their sufficient statistics sum up to $\boldsymbol{\lambda}$.

3.2 Graphical Model

The structure of Laplace exponential family PCA is defined by **Fig.4**.

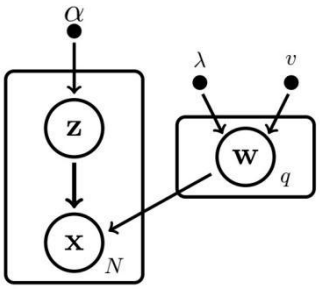

Fig. 4. PGM for LePCA

The conjugation on basis for the subspace is preserved by introducing conjugate prior on $\mathbf{W}$. However, as long as the dimensionality needs to be learned, an ARD

prior is applied on $\mathbf{Z}$, denoting the collection of all latent variables. Unlike SePCA, the conjugate relationship saves more space for tuning parameters in a more intuitive and interpretive way. As it is pointed out in [12], the difference in status between $\mathbf{W}$ and $\mathbf{Z}$ is imposed. Traditional BPPCA obtains a sparse result by treating $\mathbf{Z}$ as latent variables and integrating out $\mathbf{W}$. It is also applicable to integrate out $\mathbf{Z}$, and yield a closed form of solution at length. In exponential family context, integrating out a ARD prior $\mathbf{Z}$ preserves sparsity and conjugation at the same time.

Formally, the prior for basis takes the form:

$$\mathrm{p}(\mathbf{w} \mid \nu, \boldsymbol{\lambda}) \propto \exp\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\lambda} - \nu \cdot G(\mathbf{w})\}$$

Where $\mathbf{w}$ is one column of loading matrix $\mathbf{W}$. As for latent variables:

$$\mathrm{p}(\mathbf{z}) = \mathrm{N}((\mathbf{0},1)^{T}, \boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1} = \mathrm{diag}\{\alpha_1^{-1}, \alpha_2^{-1}, ..., \alpha_q^{-1}\})$$

By setting $\alpha_q^{-1} \approx 0$, the biased term is taken off, so the emission distribution is:

$$\mathrm{p}(\mathbf{x} \mid \mathbf{W}, \mathbf{z}) = \exp\{\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{z} - G(\mathbf{W}\mathbf{z})\}$$

Since this is a probabilistic model with latent variables, we naturally resort to EM on parameter estimation. In the E-step, estimate the conditional distribution on $\mathbf{z}$:

$$\mathrm{p}(\mathbf{z} \mid \mathbf{W}, \mathbf{x}) = \frac{\mathrm{p}(\mathbf{x} \mid \mathbf{W}, \mathbf{z})\mathrm{p}(\mathbf{z})}{\int \mathrm{p}(\mathbf{x}, \mathbf{z}' \mid \mathbf{W})\mathrm{d}\mathbf{z}'} = \frac{\exp\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{z} + \mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{z} - G(\mathbf{W}\mathbf{z})\}}{c \cdot \int \exp\{-\frac{1}{2}\mathbf{z}'^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{z}' + \mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{z}' - G(\mathbf{W}\mathbf{z}')\}\mathrm{d}\mathbf{z}'}$$

The difficulty within this approach is that on the denominator, latent variable $\mathbf{z}'$ can not be integrated out conveniently due to the variety in form of cumulant function G and the fraction takes an irregular form. However, it is still possible to estimate parameters with proper approximation. In previous works as EPCA or SePCA, an iterative and alternative approach is used. We are now to give another solution using approximation and indicate that the previous approaches are variants of the one we proposed.

3.3 On Parameter Estimation: Laplace Approximation

Interpretation between iterative optimization and EM within exponential PCA is given in [11]. It is also straightforward to elaborate the similarity by using point estimation to optimize joint likelihood instead of calculating the complete posterior distribution. Thus E-step is reduced to:

$$\mathbf{z}^{(\mathrm{t})} = \arg\min_{\mathbf{z}}\{-\log \mathrm{p}(\mathbf{x} \mid \mathbf{W}^{(\mathrm{t}-1)}, \mathbf{z}) - \log \mathrm{p}(\mathbf{z})\}$$

Where it is assumed that the posterior distribution is peaked at its maximum. Using the literature of Bregman distance[13]:

$$\mathrm{B}_F(\mathbf{p} \,\|\, \mathbf{q}) = F(\mathbf{p}) - F(\mathbf{q}) - \nabla F(\mathbf{q})^{\mathrm{T}}(\mathbf{p} - \mathbf{q})$$

The optimized target can be rewrite into:

$$\mathrm{B}_F(\mathbf{x} \,\|\, g(\mathbf{W}^{(\mathrm{t}-1)}\mathbf{z})) + \frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{z}$$

Where we have $g = G'$ and $F = \int g^{-1}$. Once the context-related distribution is determined, these functions can be calculated analytically. This optimization of $\mathbf{z}$ is reduced into a standard convex optimization.

To estimate $\mathbf{W}$, applying the M-step:

$$\begin{aligned}
\mathbf{W}^{(\mathrm{t})} &= \arg\max_{\mathbf{W}} \{ \int \mathrm{p}(\mathbf{Z} \mid \mathbf{X}, \mathbf{W}^{(\mathrm{t}-1)}) \log \mathrm{p}(\mathbf{X}, \mathbf{Z} \mid \mathbf{W}) \mathrm{d}\mathbf{Z} \} \\
&= \arg\max_{\mathbf{W}} \{ \log \mathrm{p}(\mathbf{X}, \mathbf{Z}^{(\mathrm{t})} \mid \mathbf{W}) \} \\
&= \arg\max_{\mathbf{W}} \{ \log \mathrm{p}(\mathbf{X} \mid \mathbf{Z}^{(\mathrm{t})}, \mathbf{W}) \}
\end{aligned}$$

The cancellation of integral is due to the fact that the posterior distribution has been approximated as a Dirac function. The degenerated M-step is again a convex optimization.

To use the full EM algorithm, it is feasible to use Laplace approximation, which was firstly used in classification context to handle non-Gaussian likelihood, hence fit our issue. This equals using the second-order Taylor expansion of $f(\mathbf{z}) = G(\mathbf{W}\mathbf{z})$ to approximate the exact culumant function:

$$\begin{aligned}
G(\mathbf{W}\mathbf{z}) &= f(\mathbf{z}) \\
&= f(\mathbf{z}_0 + (\mathbf{z} - \mathbf{z}_0)) \\
&\approx f(\mathbf{z}_0) + \nabla f(\mathbf{z}_0)^{\mathrm{T}}(\mathbf{z} - \mathbf{z}_0) + \frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^{\mathrm{T}} \nabla^2 f(\mathbf{z}_0)(\mathbf{z} - \mathbf{z}_0) \\
&= G(\mathbf{W}\mathbf{z}_0) + \nabla G(\mathbf{W}\mathbf{z}_0)^{\mathrm{T}} \mathbf{W}(\mathbf{z} - \mathbf{z}_0) + \frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \nabla^2 G(\mathbf{W}\mathbf{z}_0) \mathbf{W}(\mathbf{z} - \mathbf{z}_0) \\
&= \frac{1}{2}\mathbf{z}^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \nabla^2 G(\mathbf{W}\mathbf{z}_0) \mathbf{W}\mathbf{z} + (\nabla G(\mathbf{W}\mathbf{z}_0)^{\mathrm{T}} \mathbf{W} - \mathbf{z}_0^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \nabla^2 G(\mathbf{W}\mathbf{z}_0) \mathbf{W}) \mathbf{z} + c \\
&= \frac{1}{2}\mathbf{z}^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \mathbf{H} \mathbf{W}\mathbf{z} + \alpha^{\mathrm{T}} \mathbf{z} + c
\end{aligned}$$

Where we denote the Hessian matrix for G at $\mathbf{z}_0$ by $\mathbf{H}$ and absorb constant term into C and coefficient of linear term into $\boldsymbol{\alpha}$.

Laplace approximation addresses the marginal likelihood which is independence of $\mathbf{z}$, and we can now approximate the full posterior of $\mathbf{z}^{(t)}$ as $\mathrm{N}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$, where the Hessian is computed as $\mathbf{z}_0 = \mathbf{z}^{(t-1)}$ for better approximation:

$$\boldsymbol{\Sigma}^{(t)} = (\boldsymbol{\Lambda} + \mathbf{W}^{(t-1)^{\mathrm{T}}} \mathbf{H}^{(t-1)} \mathbf{W}^{(t-1)})^{-1} \tag{1}$$

$$\boldsymbol{\mu}^{(t)} = \boldsymbol{\Sigma}^{(t)} (\mathbf{W}^{(t-1)^{\mathrm{T}}} \mathbf{x} + \boldsymbol{\alpha}^{(t-1)}) \tag{2}$$

In M-step, we optimize $\mathbf{W}$ with respect to the expectation of log-likelihood under posterior distribution obtained before can be processed using Laplace approximation again:

$$\begin{aligned} &\int \mathrm{p}(\mathbf{z} \mid \mathbf{W}^{(t-1)}, \mathbf{x}) \{-\frac{1}{2} \mathbf{z}^{\mathrm{T}} \boldsymbol{\Lambda} \mathbf{z} + \mathbf{x}^{\mathrm{T}} \mathbf{W} \mathbf{z} - G(\mathbf{W}\mathbf{z})\} \mathrm{d}\mathbf{z} \\ \approx &\int \mathrm{p}(\mathbf{z} \mid \mathbf{W}^{(t-1)}, \mathbf{x}) \{-\frac{1}{2} \mathbf{z}^{\mathrm{T}} (\boldsymbol{\Lambda} + \mathbf{W}^{\mathrm{T}} \mathbf{H}^{(t-1)} \mathbf{W}) \mathbf{z} + (\mathbf{x}^{\mathrm{T}} \mathbf{W} + \boldsymbol{\alpha}^{(t-1)^{\mathrm{T}}}) \mathbf{z} - C\} \mathrm{d}\mathbf{z} \end{aligned}$$

Which can be solved analytically since it only composed of the expectation of first and second order moment of a Gaussian variable.

If the Hessian is a constant, calculation can be reduced sharply. A trivial case is Gaussian distribution with cumulant function $G(\theta) = \theta^2 / 2$, hence yields a constant Hessian. For general exponential family, the Hessian needs to be computed for times during iterations. Addressing the similarity between our model and FA, FA assumes $\mathrm{p}(\mathbf{x} \mid \mathbf{W}, \mathbf{z})$ to be normally distributed $\mathrm{N}(\mathbf{W}\mathbf{z}, \boldsymbol{\Psi})$. Applying Laplace approximation on the emission distribution results in a variable $\boldsymbol{\Psi}^{(t)}$ (notice the similarity between equation(1) and the E-step for FA, and sufficient statistics we used through does not equal the one for Gaussian in factor analysis). The approximation should only be used during estimation so the overall distribution of data remains general.

LePCA' s parameter estimation can be concluded as follow:

Table 1. EM algorithm for LePCA

- **Initialization:** give $\mathbf{z}^{(0)}$ and $\mathbf{W}^{(0)}$ according to prior distribution; $t = 0$;
- **Iteration:** while not converge:
  - Compute $\mathbf{H}^{(t)} = \nabla^2 G(\mathbf{W}\mathbf{z})|_{\mathbf{z}^{(t)}}$;
  - E-step: Compute $\mathrm{p}(\mathbf{z}^{(t+1)})$ w.r.t $\mathbf{H}^{(t)}$, $\mathbf{W}^{(t)}$ using equation (1) and (2), then calculate expectation for $\mathbf{z}$ and $\mathbf{z}\mathbf{z}^{\mathrm{T}}$;
  - M-step: Optimize $\mathbf{W}^{(t+1)}$ w.r.t the expectations calculated before; similar to M-step for FA;
  - $t = t + 1$;

3.4 On Parameter Estimation: Evidence Framework

To learn the dimension of principal components q , we take $\mathbf{W}$ as granted and integrate out $\mathbf{z}$ to learn the hyper-parameter $\boldsymbol{\alpha}$. In PPCA, $\mathbf{z}$ can be integrated out easily using Gaussian properties. The difficulty within exponential family model is the same as the one arises in E-step. Hence it can be handled by using Laplace approximation as well. For latent variable $\mathbf{z}$, we collect all i th component in it from N data into $\mathbf{Z}_i$, using the evidence framework as linear regression, we have the re-estimation formula as in [9] or [14], Chapter 3.5:

$$\alpha_i := \frac{d - \alpha_i \mathrm{tr}(\mathbf{H}^{-1})}{\mathbf{Z}_i^{\mathrm{T}} \mathbf{Z}_i}$$

Where $\mathbf{H}$ is the Hessian for $\log \mathrm{p}(\mathbf{Z}_i \mid \mathbf{X}, \mathbf{W})$ with respect to $\mathbf{Z}_i$, which consists of constant terms thanks to the Laplace approximation. In practice, it is possible to ignore the second term in numerator to simplify the re-estimation further. By evidence framework, the dimension with less support from the data will increase related component in $\boldsymbol{\alpha}$ quickly. This ends in an decrease in the prior variance for that dimension. Since we have the prior to be a zero-mean Gaussian, this is tantamount to drive all components in that dimension to zero.

4 Experiment and Illustration

On synthetic data set used in [10] we test SePCA. The data set consists of $N = 120$ data that subject to Bernoulli distribution with $d = 16$. The "genuine" data is divided into three groups with consistent content. Each component in each data item is further flipped with probability 0.1 to form the training set. The data set is illustrated in **Fig.5**., where vertical and horizontal axes represent dimension and data respectively.

(a) **(b)**

Fig. 5. Synthetic data set, the prototype **(a)** and the one added with noise**(b)**

Bernoulli distribution belongs to the exponential family. As for this data set, its cumlant function takes a concise form:

$$G(\boldsymbol{\theta}) = -\sum_{i=1}^{d} \log(1-\sigma(\theta_i))$$

Where σ denotes sigmoid function. With this fact the gradient and Hessian(which is diagonal) can be easily computed.

Three typical phases in iteration are selected and demonstrated in **Fig.6**:

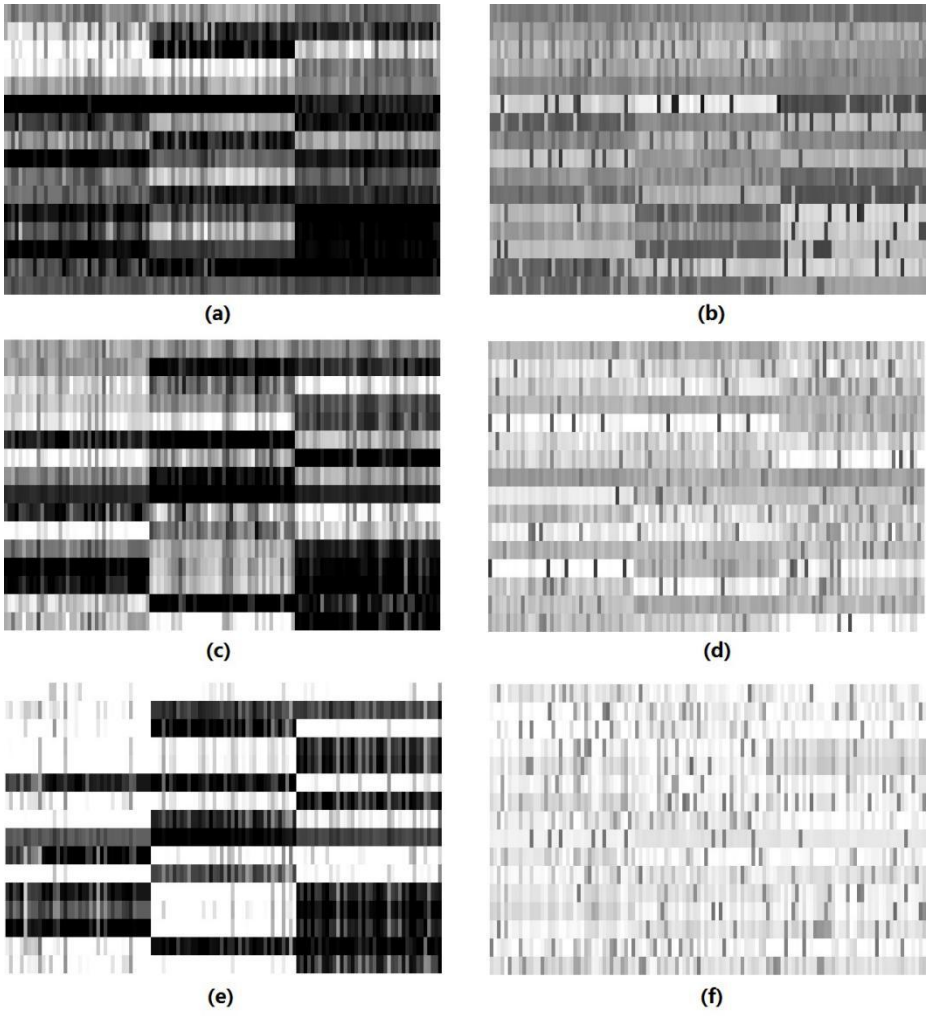

Fig. 6. The chronological mean for natural parameter during iterations **(a)**, **(c)**, **(e)**, and the corresponding difference between the mean and the actual data set **(b)**, **(d)**, **(f)**.

To quantify the effect, we calculate the log-likelihood $\log p(\mathbf{X} \mid \mathbf{W}, \mathbf{Z})$ at the final stage, and reach a log-likelihood on training set and genuine data on -302.37 and -266.95 respectively. Together with features in model structure, we compare LePCA with other models thoroughly:

Table 2. Comparison between models(figures for some previous models are tested in [11])

	BPPCA	EPCA	BXPCA	SePCA	LePCA
Prior of $\mathbf{z}$	$N(\mathbf{0}, \mathbf{I})$	flat	$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$N(\mathbf{0}, \mathbf{I})$	ARD
Prior of $\mathbf{w}$	ARD	flat	Conjugate	ARD	Conjugate
Canceled	$\mathbf{w}$	no	no	$\mathbf{w}$	$\mathbf{z}$
Full EM	Applicable	MAP	MAP	MAP	Applicable
Train LL		-13.3	-202.7	-338.3	-302.4
Prototype LL		-808.1	-517.4	-231.0	-267.0

It is noticeable that both SePCA and LePCA yield to relatively better effect on both training set and genuine data. EPCA suffers from over-fitting by using MLE. BXPCA handles with Bayesian approach but the number for free parameters are still too large(due to the unidentifiability of models). SePCA results in a better performance. However, its probabilistic characteristic is reduced by using MAP instead of a full EM. Allowing conjugate prior, LePCA increase the interpretability significantly. And it is also possible to apply a sequential learning process based on conjugation.

5 Conclusion

In this paper we propose LePCA as a solution to handle the difficulties of generalization of PCA to exponential family. From a probabilistic perspective, the difficulty arises from the fact that latent variable can hardly be integrated out for general exponential family distributions. Instead of using MAP as an approximation during EM as previous models, we use Laplace approximation to cancel latent variables, hence provide a more probabilistic solution. By addressing the similarity between LePCA and the most general model with continuous latent variables as factor analysis, our model can be taken as a more dynamic approach with the Hessian of the emission distribution re-estimated at each iteration. It is easy to specify LePCA by setting context-related parameters to give rise to other basic models as PPCA.

Acknowledgement This research work is funded by the National Key Research and Development Project of China (2016YFB0801003)

References

1. Wold, Svante, K.Esbensen, and P.Geladi. "Principal component analysis." *Chemometrics & Intelligent Laboratory Systems* 2.1–3(1987):37-52.
2. Hotelling, H. "Analysis of a complex of statistical variables into principal components. " *British Journal of Educational Psychology* 24.6(1933):417-520.
3. Tipping, Michael E., and C. M. Bishop. "Probabilistic Principal Component Analysis." *Journal of the Royal Statistical Society*61.3(1999):611-622.
4. Guo, Yuhong. "Supervised exponential family principal component analysis via convex optimization." *International Conference on Neural Information Processing Systems* Curran Associates Inc. 2008:569-576.
5. Collins, Michael, S. Dasgupta, and R. E. Schapire. "A generalization of principal component analysis to the exponential family." *International Conference on Neural Information Processing Systems: Natural and Synthetic* MIT Press, 2001:617-624.
6. Mackay, David J C. "Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks." *Network Computation in Neural Systems* 6.3(1995):469-505.
7. Robert, Christian. *Machine Learning, a Probabilistic Perspective*. MIT Press, 2012.
8. Akaike, Hirotugu. "Factor analysis and AIC." *Psychometrika*52.3(1987):317-332.
9. Bishop, Christopher M. "Bayesian PCA." *Advances in Neural Information Processing Systems* DBLP, 1999:382-388.
10. Mohamed, Shakir, K. A. Heller, and Z. Ghahramani. "Bayesian Exponential Family PCA." *Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December* DBLP, 2008:1089-1096.
11. Li, J., and D. Tao. "Simple exponential family PCA." *IEEE Trans Neural Netw Learn Syst* 24.3(2013):485-497.
12. Nakajima, Shinichi, M. Sugiyama, and D. Babacan. "On Bayesian PCA: automatic dimensionality selection and analytic solution." *International Conference on International Conference on Machine Learning*Omnipress, 2011:497-504.
13. Miller, Frederic P., et al. "Bregman Divergence." *Alphascript Publishing*(2010).
14. Bishop, Christopher M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc. 2006.