Efficient Global Optimization for Exponential Family PCA and Low-Rank Matrix Factorization

Yuhong Guo
Computer Sciences Laboratory
Australian National University
Canberra, ACT 0200, Australia
yuhongguo.cs@gmail.com

Dale Schuurmans
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8, Canada
dale@cs.ualberta.ca

***Abstract*—We present an efficient global optimization algorithm for exponential family principal component analysis (PCA) and associated low-rank matrix factorization problems. Exponential family PCA has been shown to improve the results of standard PCA on non-Gaussian data. Unfortunately, the widespread use of exponential family PCA has been hampered by the existence of only local optimization procedures. The prevailing assumption has been that the non-convexity of the problem prevents an efficient global optimization approach from being developed. Fortunately, this pessimism is unfounded. We present a reformulation of the underlying optimization problem that preserves the identity of the global solution while admitting an efficient optimization procedure. The algorithm we develop involves only a sub-gradient optimization of a convex objective plus associated eigenvector computations. (No general purpose semidefinite programming solver is required.) The low-rank constraint is exactly preserved, while the method can be kernelized through a consistent approximation to admit a fixed non-linearity. We demonstrate improved solution quality with the global solver, and also add to the evidence that exponential family PCA produces superior results to standard PCA on non-Gaussian data.**

I. Introduction

Few methods are more commonly used in machine learning and statistical data analysis than principal component analysis (PCA). PCA provides a convenient form of dimensionality reduction that is useful for visualization, compression, feature discovery, embedding, and data cleaning. Three key features of PCA make it particularly well suited for these purposes; namely, that it optimally approximates the original data under a well understood reconstruction loss, it re-codes data using uncorrelated features, and it admits an efficient procedure for computing optimal embeddings. PCA is not without its shortcomings however, as it suffers from well known weaknesses; particularly the tacit assumptions of linearity and Gaussian data.

In this paper we focus on the implicit Gaussian assumption and consider generalized forms of PCA that are better suited to non-Gaussian data. Many extensions to PCA have been proposed that relax the assumption of a Gaussian data distribution. Exponential family PCA [1], [2], [3], [4] is the most prominent example, where the underlying statistical inference principle for dimensionality reduction is extended to the general exponential family. Closely related approaches use non-least squares cost functions for generalized linear modeling [5], [6], [7] and generalized matrix factorization [8], [9], [10]. All of this previous work has shown that improved forms of dimensionality reduction can be obtained by using exponential family models appropriate for the data at hand, such as multinomial models for discrete data [2], [3], [4], Poisson models for integer data [1], and exponential models for nonnegative data [6], [7]. (Although not focused on dimensionality reduction per se, independent component analysis [11] and its extensions [12] also deploy non-Gaussian assumptions.) Given data from a non-Gaussian distribution these techniques are better able than PCA to capture the intrinsic low dimensional structure or independent feature structure.

The main drawback with existing non-Gaussian dimensionality reduction methods, however, is that they rely on iterative local optimization procedures, which are not guaranteed to produce an optimal embedding under their respective cost functions. In fact, it has been assumed [5], [8] that efficient global optimization is not possible in these cases, since the cost function being minimized is non-convex and an additional non-convex constraint is used to enforce feature orthogonality. The apparent difficulty of solving the underlying optimization problem has dissuaded researchers from investigating global solution methods for these problems. Unfortunately, the lack of global solution methods has kept exponential family PCA and its relatives from being widely deployed, even when Non-Gaussian data is ubiquitous.

In this paper, we show that, despite the non-convex nature of the underlying problem, an efficient global optimization strategy can be developed for exponential family PCA and related low-rank matrix factorization problems. In particular, we demonstrate through a series of problem reformulations, exploiting convex duality of sub-problems and eigenvector properties, that an efficiently solvable formulation of the problem can be derived that preserves solution equivalence to the original. Although our focus in this paper is not on non-linearity per se, we also show how a general empirical approximation can be kernelized, thus enabling a fixed non-linearity to be incorporated in the model in a manner analogous to kernel PCA [13], [14]. (We do not specifically address the question of how to *estimate* the non-linearity itself, as is done in [15], [16], [17]—this remains an issue for future work.)

The remainder of this paper is organized as follows. First, we provide a brief recap of PCA in Section II, showing its connection to approximate matrix factorization and maximum likelihood estimation of Gaussian means. Then using PCA as a foundation, we briefly review the extension to exponential family PCA, re-express it in our framework, and explain the difficulty that exists in the resulting optimization. Then in Section III we present the main result of the paper: we show how the apparently hard optimization problem can be reformulated in an equivalent but efficiently solvable form. Finally, we discuss an extension to a flexible empirical approximation approach in Section IV that enables the general introduction of a kernel, and thus allows us to incorporate non-linearities in the model. Section VI concludes the paper with a discussion of future directions.

II. Preliminaries

Assume we are given a $t \times n$ data matrix, X, consisting of t observations of n-dimensional feature vectors, $X_{i:}$, from which we would like to recover a k-dimensional re-representation of the data. That is, one would like to assign a low dimensional vector, $Z_{i:}$, as a reduced representation of each high dimensional observation, $X_{i:}$. This is often thought of as discovering a hidden low dimensional manifold in the high dimensional feature space. A key restriction that we would also like to enforce is that the features used for coding, $Z_{:j}$, should be statistically uncorrelated; that is, we would like to enforce the constraint $Z^\top Z = I$, which ensures that the codes are expressed by orthogonal features in the low dimensional representation.

A. PCA

Given the above setup, PCA adds the assumptions of linearity and Gaussian distributed data, albeit implicitly. In fact, there are many mathematically equivalent ways to derive PCA. In multivariate statistics, principal components are originally defined to be the (orthogonal) directions of maximum variance in the feature space [18]. For our purposes, we will approach PCA from the standpoint of self-supervised regression: we seek a low dimensional encoding that allows the original data to be optimally reconstructed by a linear map. That is, we would like to simultaneously assign a $t \times k$ matrix of low dimensional codes, Z, and a $k \times n$ matrix of reconstruction weights, W, such that $X \approx ZW$. If least squares reconstruction error is used as cost function to minimize, then one obtains a form of self-supervised regression that is equivalent to PCA

$$\min_{Z:Z^\top Z=I} \min_{W} \frac{1}{2}\text{tr}\left((X-ZW)(X-ZW)^\top\right), \quad (1)$$

where tr denotes matrix trace. This formulation also suggests that PCA can be interpreted as a form of approximate matrix factorization under a constraint of bounded rank k [8], [9], [10].

Note that, superficially, the optimization problem (1) does not appear to be easy. The objective is not jointly convex in Z and W (although it is marginally convex in W for fixed Z, and vice versa) nor is the constraint on Z convex. Despite the ostensive difficulty, the task has sufficient structure that a globally optimal solution can be easily recovered, as is well known. Since we will need to understand and adapt they key aspects of this argument, below we briefly recap the major steps in the derivation.

First, since (1) is convex in W for fixed Z, the inner minimization can be solved by determining a critical point, given by a W that satisfies $d/dW = Z^\top(ZW - X) = W - Z^\top X = 0$, immediately implying that $W = Z^\top X$. Substituting this in the original problem yields

$$\min_{Z:Z^\top Z=I} \frac{1}{2}\text{tr}\left((I-ZZ^\top)(I-ZZ^\top)XX^\top\right). \quad (2)$$

Although the objective remains non-convex in Z, given the prevailing constraint on Z the objective simplifies to $(I - ZZ^\top)(I-ZZ^\top) = I - 2ZZ^\top + ZZ^\top ZZ^\top = I - ZZ^\top$, thus yielding a simpler problem with a convex objective

$$\begin{aligned} &\arg\min_{Z:Z^\top Z=I} \frac{1}{2}\text{tr}\left((I-ZZ^\top)XX^\top\right) \\ = \ &\arg\max_{Z:Z^\top Z=I} \text{tr}\left(ZZ^\top XX^\top\right). \quad (3)\end{aligned}$$

Famously, despite the non-convex constraints, (3) has an efficiently computable solution given by $Z = Q^{(k)}_{max}(XX^\top)$—the top k eigenvectors of $XX^\top$ [19]. That is, the low dimensional basis for re-coding the data is given by the top k eigenvectors of the empirical covariance matrix.[1] Unfortunately, each step of this derivation faces a non-convex problem, and it appears that it is only by very special structure that a globally optimal solution is obtained. Generalizing this derivation to other conditions has been considered difficult. Nevertheless, we demonstrate concrete progress below.

B. Gaussian Model

The generalized PCA models we consider below are based on a probabilistic interpretation of classical PCA. Once again, there are many formulations of PCA as the solution of maximum likelihood or maximum a posteriori problems [20], [21], [22], all inevitably based on a multivariate Gaussian model. The specific form of probabilistic PCA we consider will lead directly to exponential family PCA below, and provide a bridge between the classical least squares formulation and more recent formulations based on alternative loss functions.

Consider a conditional Gaussian model where an n-dimensional vector $\mathbf{x}$ is generated from a k-dimensional code $\mathbf{z}$ and a set of parameters W, according to $\mathbf{x} = W^\top \mathbf{z} + \epsilon$ where $\epsilon \sim N(0, \alpha I)$ for some fixed variance $\alpha > 0$. Then the conditional density of $\mathbf{x}$ given $\mathbf{z}$ and W is given by

$$p(\mathbf{x}|\mathbf{z}, W) \quad = \quad \exp\left(\frac{1}{\alpha}\mathbf{z}^\top W\mathbf{x} - \frac{1}{2\alpha}\mathbf{x}^\top\mathbf{x} - A(\mathbf{z}^\top W)\right), (4)$$

[1] Here we are assuming that the data has already had its mean subtracted so that X is centered; that is $X^\top \mathbf{1} = \mathbf{0}$, meaning that $\hat{\text{cov}}(X) = XX^\top/t - (X^\top\mathbf{1})(\mathbf{1}^\top X)/t^2 = XX^\top/t$. We suppress such details and assume the data is centered as necessary.

such that

$$A(\mathbf{z}^\top W) \;=\; \frac{1}{2\alpha}\mathbf{z}^\top W W^\top \mathbf{z} + \frac{n}{2}\log(2\pi\alpha). \tag{5}$$

Once again, given a matrix of data X, we would like to assign a low dimensional code, $Z_{i:}$, to each high dimensional observation, $X_{i:}$, and also solve for a matrix of shared reconstruction parameters, W, but now following the goal of maximizing the conditional likelihood of the data, $p(X|ZW)$. As before, we maintain the constraint that the code features remain uncorrelated (i.e. $Z^\top Z = I$). Combining these notions yields the optimization problem

$$\begin{aligned} &\min_{Z:Z^\top Z=I}\ \min_{W}\ \frac{1}{2\alpha}\mathrm{tr}\left(ZWW^\top Z^\top\right) + \frac{nt}{2}\log(2\pi\alpha) \\ &\qquad -\frac{1}{\alpha}\mathrm{tr}\left(ZWX^\top\right) + \frac{1}{2\alpha}\mathrm{tr}\left(XX^\top\right) \qquad (6) \\ = &\min_{Z:Z^\top Z=I}\ \min_{W}\ \frac{1}{2\alpha}\mathrm{tr}\left((X-ZW)(X-ZW)^\top\right) \\ &\qquad +\frac{nt}{2}\log(2\pi\alpha). \qquad (7) \end{aligned}$$

It is obvious that (7) has the same minimizer as (1) since the objectives are identical up to a multiplicative and additive constant. Note that the result does not depend on the assumed Gaussian variance α.

C. Exponential Family PCA

The form (4) suggests an immediate generalization of the Gaussian conditional model to a general class exponential family models. The resulting extension of Gaussian to non-Gaussian PCA is precisely the original proposal of [1]. A general exponential family representation of the conditional distribution of an observation vector $\mathbf{x}$ given a low dimensional code vector $\mathbf{z}$ and parameter matrix W can be written as

$$p(\mathbf{x}|\mathbf{z}, W) \;=\; \exp\left(\mathbf{z}^\top W\mathbf{x} + \log q(\mathbf{x}) - A(\mathbf{z}^\top W)\right), \tag{8}$$

where

$$A(\mathbf{z}^\top W) \;=\; \log\int \exp\left(\mathbf{z}^\top W\mathbf{x}\right)\, q(\mathbf{x})\, d\mathbf{x}\,. \tag{9}$$

By altering the choice of base measure $q(\mathbf{x})$, features of $\mathbf{x}$, domain of $\mathbf{x}$, and domain of W, one can change the form of the conditional model on $\mathbf{x}$ given $\mathbf{z}$. For example, exponential, Poisson, Bernoulli, and multinomial distributions on $\mathbf{x}$ can be obtained through suitable choices of these components [1], [23]. General Markov random field models on $\mathbf{x}$ can also be obtained in this way [23]. A key practical concern is whether one can compute the log normalization factor, $A(\mathbf{z}^\top W)$. It turns out that this quantity has a closed form solution for the standard distributional forms mentioned above, and can be efficiently computed for Markov random fields with sufficient structure [23]. We revisit this issue in Section IV below.

An exponential family model can be used for dimensionality reduction in the same manner as the Gaussian model. Specifically, dimensionality reduction can be cast as a maximum conditional likelihood problem, where the codes, Z, and parameters, W, are simultaneously optimized, subject to the independence constraint on Z. That is, we minimize

$$\min_{Z:Z^\top Z=I}\ \min_{W}\ \Big(\sum_i A(Z_{i:}W) - \chi_i\Big) - \mathrm{tr}\left(ZWX^\top\right), \tag{10}$$

where $\chi_i = \log q(X_{i:})$. This form is equivalent to the proposed objective in [1], with the exception of the orthogonality constraint $Z^\top Z = I$ (which was not originally imposed). Despite simplifying the problem by omitting the orthogonality constraint, [1] did not provide a global solution procedure. Instead, an alternating minimization procedure was suggested that is not guaranteed to avoid local minima. The original formulation of [1] has since been modified by [2], [3] and [5], and encapsulated in an independently developed model [4]. All of these formulations differ slightly in detail, but uniformly rely on iterative solution procedures that do not bring any guarantee of global optimality. [8] notes that using a non-least squares reconstruction loss in (1) tends to create local minima. However, we now show that even with the non-convex orthogonality constraint, a regularized version of (10) can be reformulated that bypasses local minima, and allows a global solution to be obtained. Our derivation exploits both results on eigenvector analysis [19] and convex duality [23].

III. Efficient Global Optimization

The optimization objective we consider for dimensionality reduction augments the original exponential family PCA formulation (10) with a quadratic regularizer on W. The regularization term can be interpreted as a zero-mean Gaussian prior on W with diagonal covariance. The addition of a regularization term allows a maximum a posteriori formulation of the problem and also simplifies subsequent mathematical details. The original formulation (10) can be approximated arbitrarily closely by setting the regularization parameter, β, close to zero. We do not impose any prior on Z, but rather maintain the orthogonality constraint $Z^\top Z = I$. Thus we are left with the main optimization problem

$$\begin{aligned} &\min_{Z:Z^\top Z=I}\ \min_{W}\ \Big(\sum_i A(Z_{i:}W) - \chi_i\Big) - \mathrm{tr}\left(ZWX^\top\right) \\ &\qquad +\frac{\beta}{2}\mathrm{tr}\left(W^\top W\right). \qquad (11) \end{aligned}$$

As before, the problem is not jointly convex in W and Z and has a non-convex constraint. The hope for an efficient solution rests on the fact that the problem is convex in W for fixed Z. In particular, it can be verified that $A(\mathbf{z}^\top W)$, as defined in (9) is convex in W for given $\mathbf{z}$ [23], [24]. This fact, combined with a key result from the semidefinite optimization literature allows us to derive our main result.

Theorem 1: The optimization problem (11) is equivalent to

$$\begin{aligned} &\max_{U}\ \min_{Z:Z^\top Z=I} -\Big(\sum_i A^*(U_{i:}) + \chi_i\Big) \\ &\qquad -\frac{1}{2\beta}\mathrm{tr}\left((X-U)(X-U)^\top ZZ^\top\right), \qquad (12) \end{aligned}$$

where U is a $t \times n$ matrix, $A^*(U_{i:})$ is the Fenchel conjugate of $A(Z_{i:}W)$, and the parameters W can be recovered from U and Z by

$$W \quad = \quad \frac{1}{\beta} Z^\top (X - U). \tag{13}$$

The proof will proceed by a series of reformulations stated in lemmas.

Lemma 1:

$$\begin{aligned} \min_W \quad & \Big(\sum_i A(Z_{i:}W) - \chi_i\Big) - \operatorname{tr}\left(ZWX^\top\right) \\ & + \frac{\beta}{2} \operatorname{tr}\left(W^\top W\right) \qquad (14) \\ = \quad \max_U \quad & -\Big(\sum_i A^*(U_{i:}) + \chi_i\Big) \\ & - \frac{1}{2\beta} \operatorname{tr}\left((X-U)(X-U)^\top ZZ^\top\right). \qquad (15) \end{aligned}$$

Proof: First consider $A(Z_{i:}W)$ as defined in (9). This function is convex in W and can be re-expressed as $A(Z_{i:}W) = \max_{U_{i:}} \operatorname{tr}\left(Z_{i:}WU_{i:}^\top\right) - A^*(U_{i:})$ where A^* is the Fenchel conjugate of A. Importantly, A^* is a closed convex function [24], [23]. Thus, we can rewrite the inner minimization of (11) as

$$\begin{aligned} \min_W \max_U \quad & -\Big(\sum_i A^*(U_{i:}) + \chi_i\Big) + \operatorname{tr}\left(ZW(U - X)^\top\right) \\ & + \frac{\beta}{2} \operatorname{tr}\left(W^\top W\right). \end{aligned} \tag{16}$$

Let $F(W, U)$ denote the objective in (16). Crucially, one can verify that F satisfies the conditions of the strong minmax theorem (stated in Theorem 2 below), which allows the order of the minimization and maximization to be reversed. That is, based on the strong minmax theorem we conclude that (16) is equivalent to

$$\begin{aligned} \max_U \min_W \quad & -\Big(\sum_i A^*(U_{i:}) + \chi_i\Big) + \operatorname{tr}\left(ZW(U - X)^\top\right) \\ & + \frac{\beta}{2} \operatorname{tr}\left(W^\top W\right). \end{aligned} \tag{17}$$

Now, the inner minimization on W can be easily solved by determining a critical point, since the function is convex in W for fixed U. Thus, $d/dW = Z^\top(U - X) + \beta W = 0$ which implies $W = \frac{1}{\beta} Z^\top (X - U)$. Substituting this into (17) yields the result. ■

Theorem 2: (Strong Minmax Property) Consider a joint function $f(x, y)$ defined over $x \in X$ and $y \in Y$. Assume (1) $f(\cdot, y)$ is a closed and convex for all $y \in Y$; (2) $f(x, \cdot)$ is closed and concave for all $x \in X$; (3) $\sup_{y \in Y} f(x, y) < \infty$ for all $x \in X$; and (4) $f(\cdot, y)$ has bounded level sets $\{x : f(x, y) \leq t\}$ for all $y \in Y$. Then $\inf_{x \in X} \sup_{y \in Y} \; f(x, y) = \sup_{y \in Y} \inf_{x \in X} f(x, y)$ and the solution value is finite.

Proof: This theorem is just a specialization of a standard result in convex analysis; specifically [25, Theorem 37.3] and [26, Page 95]. We will make use of it one more time below. ■

Next we turn to the outer minimization on Z from the original problem (11), which involves a non-convex constraint. Note that in the optimization problem (15) Z only appears as the outer product $ZZ^\top$. This allows us to rewrite the minimization in terms of a square matrix M. Although initially this will appear like a relaxation of the original optimization problem, later we will verify that the optimal solution is preserved.

Lemma 2:

$$\begin{aligned} \min_{Z: Z^\top Z = I} \max_U \quad & -\Big(\sum_i A^*(U_{i:}) + \chi_i\Big) \\ & - \frac{1}{2\beta} \operatorname{tr}\left((X-U)(X-U)^\top ZZ^\top\right) \qquad (18) \\ \geq \quad \min_{M: I \succeq M \succeq 0, \operatorname{tr}(M) = k} \max_U \quad & -\Big(\sum_i A^*(U_{i:}) + \chi_i\Big) \\ & - \frac{1}{2\beta} \operatorname{tr}\left((X-U)(X-U)^\top M\right). \qquad (19) \end{aligned}$$

Proof: Consider the relationships that hold between the following sets of constraints on M

$$\begin{aligned} & \{M : M = ZZ^\top \text{ for some } Z \text{ such that } Z^\top Z = I\} && (20) \\ & = \quad \{M : I \succeq M \succeq 0, \operatorname{tr}(M) = k, M^2 = M\} && (21) \\ & \subseteq \quad \{M : I \succeq M \succeq 0, \operatorname{tr}(M) = k\}. && (22) \end{aligned}$$

The first equality holds because both sets of constraints bound the eigenvalues of the matrices to be either 0 or 1, with exactly k being 1 [27]. Unfortunately, neither of the first two sets of constraints is convex on M. Therefore, we relax the second set of constraints merely by dropping the non-convex constraint $M^2 = M$, which then means the eigenvalues of M are only constrained to be between 0 and 1 with their sum totaling to k. Obviously since the problem is a minimization, a lower bound is obtained by relaxing the constraint. ■

Thus, we have achieved a formulation where the problem is convex, albeit with a relaxation. That is, the outer minimization in (19) is convex in M, since a maximum of linear functions is convex and the constraints are convex [24]. Let $G(M, U)$ denote the objective in (19). Once again, one can verify that this function satisfies the conditions of the strong minmax theorem, which allows another reversal of a minimization and maximization.

Lemma 3: An equivalent optimization problem to (19) is

$$\begin{aligned} \max_U \quad \min_{M: I \succeq M \succeq 0, \operatorname{tr}(M) = k} \quad & -\Big(\sum_i A^*(U_{i:}) + \chi_i\Big) \\ & - \frac{1}{2\beta} \operatorname{tr}\left((X-U)(X-U)^\top M\right). \end{aligned} \tag{23}$$

Proof: The proof is immediate upon verifying that $G(M, U)$ satisfies the conditions of Theorem 2. ■

Note that the inner minimization is now in the form of a standard semidefinite program. We invoke a fundamental theorem about semidefinite programs of this form to achieve the following result.

Lemma 4: The optimization problem (23) is equivalent to

$$\max_{U} \min_{Z: Z^\top Z = I} -\Big(\sum_i A^*(U_{i:}) + \chi_i\Big) - \frac{1}{2\beta}\text{tr}\left((X-U)(X-U)^\top ZZ^\top\right). \quad (24)$$

Proof: The result comes from [19], which shows that the semidefinite program $\max_{M: I \succeq M \succeq 0, \text{tr}(M)=k} \text{tr}(MA)$ has the solution $Z = Q^{(k)}_{max}(A)$, which is equivalent to solving $\min_{Z: Z^\top Z = I} \text{tr}(ZZ^\top A)$. ■

The final step of the proof of Theorem 1 is to establish that nothing has been lost by the lower bound relaxation in Lemma 2.

Lemma 5: The optimization problem (24) is equivalent to (18).

Proof: Let (U^*, Z^*) be a solution to (24). Note that (24) is an equivalent optimization problem to (19) by the previous two lemmas. Hence for $M^* = Z^* Z^{*\top}$ we have that (U^*, M^*) is a solution of (19). Recall that (19) was a lower bound on (18) only insofar as the constraint $M^2 = M$ was dropped. However, the solution M^* to (19) automatically satisfies $M^{*2} = M^*$. Hence it also is a solution of (18). ■

A. Algorithm

Theorem 1 suggests an efficient algorithmic approach for solving the exponential family PCA problem (11), consisting of an outer maximization loop on U and an inner minimization on Z. By Lemma 4 the inner loop can be solved by setting $Z = Q^{(k)}_{max}((X-U)(X-U)^\top)$, hence relying on an efficient eigenvector computation. Crucially, the objective (12) is concave in U, since $-A^*(U_{i:})$ is concave, and a minimum of concave functions (i.e. the concave quadratics in the trace term) is also concave.

To solve the outer maximization we deploy a bundle method [28], which only requires that a supergradient direction (i.e. a locally improving direction) can be found at every step. At a given U, once the inner solution Z is computed, it can be verified that a supergradient direction is given by the gradient of the objective at Z [29], which in this case is given by $-\nabla_U(\sum_i A^*(U_{i:}) - \frac{1}{\beta}ZZ^\top(U-X)$.

IV. Empirical Approximation

An interesting and practical extension of the previous approach is possible. In practice, the log normalization function A might not be easily computable, which means that its Fenchel conjugate A^* might not be easily computable either, which would make implementation of the resulting exponential family PCA model virtually impossible.

In general, one can always use a sample approximation to the integral (9) and achieve an empirical approximation to the true underlying exponential family model as follows. If one replaces the integral definition (9) with an empirical definition of A,

$$A(\mathbf{z}^\top W) \quad = \quad \log \sum_i \exp\left(\mathbf{z}^\top W X_{i:}\right)/t, \qquad (25)$$

then the conjugate function can be given by

$$A^*(\Theta_{i:}) \quad = \quad \text{tr}\left(\Theta_{i:} \log \Theta_{i:}^\top\right) \qquad (26)$$

for a $1 \times t$ vector $\Theta_{i:}$ such that $\Theta_{i:} \geq 0$ and $\Theta_{i:}\mathbf{1} = 1$. With this model the exponential family PCA problem can be expressed as

$$\max_{\Theta: \Theta \geq 0, \Theta\mathbf{1}=\mathbf{1}} \min_{Z: Z^\top Z = I} -\text{tr}\left(\Theta \log \Theta^\top\right) - \frac{1}{2\beta}\text{tr}\left((I-\Theta)XX^\top(I-\Theta)^\top ZZ^\top\right). \quad (27)$$

The same algorithm as above can be employed to optimize this objective.

One of the main benefits of working with the empirical approximation is that it is automatically kernelized. That is, the data matrix X only appears in the training objective through the kernel matrix $XX^\top$. This property allows one to work with a fixed non-linearity defined through a kernel.

Interestingly, the empirical model would give equivalent results to classical PCA (or kernel PCA) if Θ were uniform, in which case $I - \Theta$ would correspond to the centering matrix $H = I - \mathbf{1}\mathbf{1}^\top$, and (27) would be equivalent to $\arg\max_{Z: Z^\top Z = I} \text{tr}(HXX^\top HZZ^\top)$. The solution is simply $Z = Q^{(k)}_{max}(HXX^\top H)$, the top k eigenvectors of the centered covariance matrix [13], [14].

V. Experimental Results

To investigate the performance of the proposed approach, we conducted experiments on both synthetic and real world data, comparing the proposed global optimization algorithm for exponential family PCA to the standard alternating minimization approach [1] and to standard PCA.

We first conducted two sets of synthetic experiments on data generated by Bernoulli and exponential distributions respectively. For the Bernoulli model, we generated a random set of latent codes Z forming four clusters in a 100×2 matrix, thus creating two dimensional codes for 100 data points, as shown in Figure 1. We then randomly generated a 2×20 parameter matrix W to be used to expand the two dimensional codes into 20 dimensional parameter vectors. Finally we sampled a 100×20 matrix X from Bernoulli distributions defined by the natural parameters ZW.

The dimensionality reduction methods were then applied to the data to produce a two dimensional embedding of X. In particular, the two exponential family PCA methods, using global and local optimization respectively, were both applied using the Bernoulli model, and these were compared to standard PCA. The results produced by the three approaches are shown in Figure 2. Interestingly, the global optimization method and standard PCA both achieve similar results on this data, and both successfully separate the four clusters in the low dimensional embedding. However, the local optimization approach does not perform nearly as well on this data set, demonstrating that local minimization might not be the most effective training approach in every circumstance.

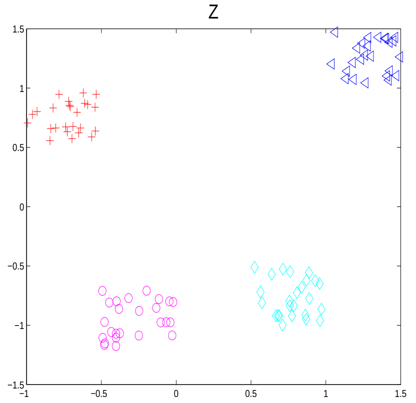

Fig. 1. Visualization of the two dimensional Z used for data generation in the synthetic Bernoulli experiment.

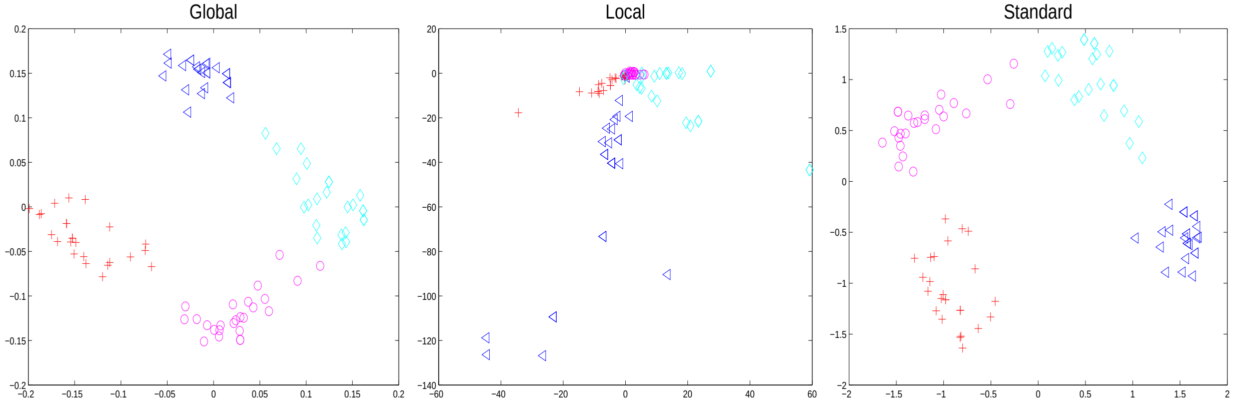

Fig. 2. Two dimensional embeddings produced by the three approaches on two dimensional space for the synthetic experiment with the Bernoulli model.

The second synthetic experiment involved data generated from a multivariate exponential distribution. We generated latent codes Z and the parameter matrix W in a slightly different manner than above. In particular, we generated a 100×2 matrix Z and the 2×3 parameter matrix W to satisfy the constraint $ZW < 0$. The 100×3 matrix of data X was then sampled componentwise from exponential distributions with mean parameters given by $\Theta = -1/(ZW)$. Figure 3 shows the data matrix X generated in three dimensional space.

As before, we applied each of the dimensionality reduction techniques to embed the data into two dimensions, but here we applied the exponential family PCA techniques (global and local) with an exponential model. The results are shown in Figure 4. In this case, one can observe that both the global and local exponential family PCA methods are more robust to outliers than standard PCA, given highly spread data generated from an exponential distribution.

We also conducted experiments using a real world data taken from the 20 newsgroups data set, which contains 200 documents sampled from four newsgroups: comp.*, rec.*, sci.* and talk.*. This document data is represented using 100 binary word indicator features. Thus we conducted experiments using Bernoulli models.

Each of the three methods were used to embed the data into 2, 5 and 10 dimensional representations. To evaluate the quality of the embeddings, we measured the integrity with which the clusterings were preserved in the dimensionality reduction process. This was done by running k-means clustering on the reduced representations and comparing the clustering accuracy to the different cluster labels given by the newsgroup identities. Table I shows the clustering accuracies obtained by the different dimensionality algorithms using 2, 5 and 10 dimensional embeddings. These results are averaged over 10 runs where an independent sample of documents is sampled each time. In this table one can see that the global optimization approach for exponential family PCA with a Bernoulli model achieves a slight advantage over local optimization for exponential family PCA and over standard PCA.

Finally, we conducted experiments using the Yale face image data set, which consists of 165 grayscale images of 15 individuals. There are 11 images per subject, each with a different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light,

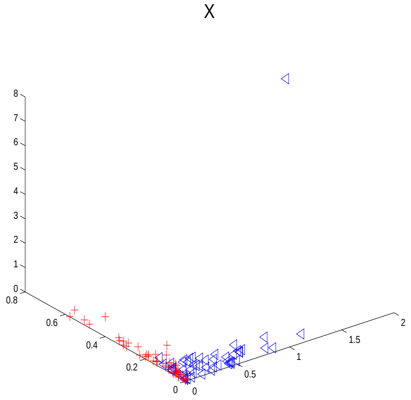

Fig. 3. Visualization of the data X generated in the synthetic experiment for the exponential distribution model.

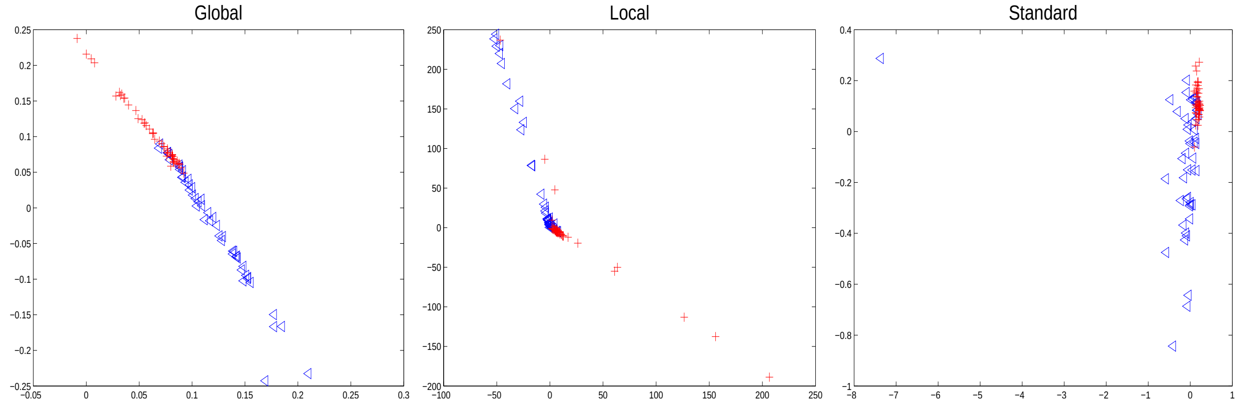

Fig. 4. Two dimensional embeddings produced by the three approaches on two dimensional space for the synthetic experiment with the exponential model.

TABLE I

EVALUATING EMBEDDING QUALITY VIA K-MEANS CLUSTERING ACCURACY OBTAINED ON NEWSGROUP DATA. COMPARING BERNOULLI EXPONENTIAL FAMILY PCA TO STANDARD PCA. GLOBAL DENOTES GLOBAL OPTIMIZATION METHOD, LOCAL DENOTES LOCAL OPTIMIZATION METHOD AND STANDARD DENOTES STANDARD PCA.

#Dim	Full Data	Global	Local	Standard
2	0.37	0.42	0.40	0.40
5	0.37	0.40	0.36	0.39
10	0.37	0.39	0.38	0.38

TABLE II

EVALUATING EMBEDDING QUALITY VIA K-MEANS CLUSTERING ACCURACY OBTAINED ON YALE IMAGE DATA. COMPARING BERNOULLI EXPONENTIAL FAMILY PCA TO STANDARD PCA. GLOBAL DENOTES GLOBAL OPTIMIZATION METHOD, LOCAL DENOTES LOCAL OPTIMIZATION METHOD AND STANDARD DENOTES STANDARD PCA.

Dataset	Full Data	Global	Local	Standard
Yale 1	0.5	0.68	0.64	0.61
Yale 2	0.59	0.52	0.43	0.45
Yale 3	0.51	0.44	0.38	0.38

sad, sleepy, surprised, and wink. Each image is represented with 1024 pixels. From this data we formed three subsets—Yale 1, Yale 2 and Yale 3—and used the three algorithms to produce a two dimensional embedding of the data. Once again we evaluated embedding quality by measuring the integrity with which the clusters (images from different people) were preserved by the dimensionality reduction process. We evaluated clustering performance using k-means in the reduced two-dimensional space and compared the classes to the original person identities. Yale 1 has 4 classes, containing Subjects 1, 3, 5 and 7, totaling 44 images; Yale 2 has 4 classes, containing Subjects 2, 4, 6 and 8, totaling 44 images; and Yale 3 has 5 classes, containing Subjects 2, 4, 6, 8, and 10, totaling 55 images. Since we cannot assume this data comes from any specific exponential family distribution, we used the empirical approximation presented in Section IV for both the global and local optimization approaches and compared these to standard PCA. Table II shows the clustering accuracies achieved by the respective two dimensional embeddings. in the two dimensional space. The convex approach presents an obvious advantage over the other two approaches.

VI. Conclusion

We have introduced a global optimization algorithm for exponential family PCA. The derivation exploits a number of facts from convex duality and eigenvector analysis to achieve a global solution technique for this class of problems. Experimental results demonstrate the benefits of the global solution technique over existing local optimization methods. We are also investigating the benefits of kernelization in the context of non-Gaussian PCA models on real data.

Acknowledgment

Some of this work was performed while the first author was at the University of Alberta. We gratefully acknowledge the support of NSERC, MITACS, Alberta Ingenuity, CFI, and the Canada Research Chairs program.

References

[1] M. Collins, S. Dasgupta, and R. Schapire, "A generalization of principal component analysis to the exponential family," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2001.

[2] Sajama and A. Orlitsky, "Semi-parametric exponential family pca," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, vol. 17, 2004.

[3] ——, "Semi-parametric exponential family PCA: Reducing dimensions via non-parametric latent distribution estimation," UCSD, Tech. Rep. CS2004-0790, 2004.

[4] A. Kaban and M. Girolami, "A combined latent class and trait model fo the analysis and visualization of discrete data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 859–872, 2001.

[5] G. Gordon, "Generalized^2 linear^2 models," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, vol. 15, 2002.

[6] N. Roy and G. Gordon, "Exponential family PCA for belief compression in POMDPs," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, vol. 15, 2002.

[7] N. Roy, G. Gordon, and S. Thrun, "Finding approximate POMDP solutions through belief compression," *Journal of Artificial Intelligence Research*, vol. 23, pp. 1–40, 2005.

[8] N. Srebro, J. Rennie, and T. Jaakkola, "Maximum-margin matrix factorization," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, vol. 17, 2004.

[9] N. Srebro and T. Jaakkola, "Linear dependent dimensionality reduction," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, vol. 16, 2003.

[10] ——, "Weighted low-rank approximations," in *Proceedings of International Conference on Machine Learning (ICML)*, 2003.

[11] A. Hyvarinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[12] G. Blanchard, M. Sugiyama, M. Kawanabe, V. Spokoiny, and K.-R. Mueller, "Non-gaussian component analysis: a semi-parametric framework for linear dimension reduction," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, vol. 18, 2005.

[13] B. Schoelkopf, A. Smola, and K.-R. Mueller, "Kernel principal component analysis," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, vol. 12, 1999.

[14] B. Schoelkopf and A. Smola, *Learning with Kernels*. MIT Press, 2001.

[15] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[16] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[17] K. Weinberger, F. Sha, and L. Saul, "Learning a kernel matrix for nonlinear dimensionality reduction," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.

[18] H. Hotelling, "Analysis of complex statistical variables into components," *Journal of Educational Psychology*, vol. 24, pp. 417–441, 1933.

[19] M. Overton and R. Womersley, "Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices," *Mathematical Programming*, vol. 62, pp. 321–357, 1993.

[20] S. Roweis, "EM algorithms for PCA and SPCA," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, vol. 10, 1997.

[21] M. Tipping and C. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society, B*, vol. 6, no. 3, pp. 611–622, 1999.

[22] N. Lawrence, "Probabilistic non-linear principle component analysis with gaussian process latent variable models," *Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.

[23] M. Wainwright and M. Jordan, "Graphical models, exponential families, and variational inference," UC Berkeley, Dept. Statistics, Tech. Rep. TR-649, 2003.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge U. Press, 2004.

[25] R. Rockafellar, *Convex Analysis*. Princeton Univ. Press, 1970.

[26] J. Borwein and A. Lewis, *Convex Analysis and Nonlinear Optimization*. Springer, 2000.

[27] J. Peng and Y. Wei, "Approximating k-means-type clustering via semidefinite programming," *SIAM Journal on Optimization*, vol. 18, no. 1, pp. 186–205, 2007.

[28] A. Belloni, "Introduction to bundle methods," MIT, Tech. Rep., 2005.

[29] R. Freund, "Subgradient optimization, generalized. programming, and nonconvex duality," MIT, Tech. Rep., 2004.