# A Generalization of Principal Component Analysis to the Exponential Family

Michael Collins, Sanjoy Dasgupta, and Robert E. Shcapire (NIPS2001)

Presented by Mikko Heikelä

# Outline

- Introduction
- Exponential family
- Generalized linear models
- Bregman distances
- The generalization of PCA
- A minimization algorithm
- Examples

## Introduction—two views of PCA

- For given data $\vec{x}_i \in \mathbf{R}^d$, find a lower dimensional subspace that minimizes the sum of squared distances between $\vec{x}_i$ and their projections $\vec{\theta}_i$ to it:

$$\sum_{i=1}^{n} ||\vec{x}_i - \vec{\theta}_i||^2 \tag{1}$$

- Probabilistic alternative: each $\vec{x}_i$ is seen as drawn from a unit gaussian $P_{\vec{\theta}_i}$ with unknown mean $\vec{\theta}_i$. Maximize the likelihood of the data subject to the condition that $\vec{\theta}_i$ belong to a low dimensional subspace
- Each $\vec{x}_i$ is seen as a version of a $\vec{\theta}_i$ in a subspace, corrupted by gaussian noise.
- These are equivalent—the negative log-likelihood is (1) plus constants

## Introduction—generalizing PCA?

- For nonnegative or discrete data the gaussian noise is not natural.
- Gaussian distribution is suited for real valued data. Other distributions in the exponential family can describe other types of data, e.g. Poisson—integer, Bernoulli—binary
- A general dimensionality reduction scheme for the exponential family can be devised
- The approach permits hybrid cases where the data contains different types of dimensions
- In general a crucial difference to ordinary PCA: the natural parameter space and the space of the data are not the same. A mapping between these is needed.
- This leeds us to look at generalized linear models (GLM), exponential families and Bregman distances.

# Exponential Family

- Conditional probability can be written in form:

$$\log P(x|\theta) = \log P_o(x) + x\theta - G(\theta) \tag{2}$$

- $\theta$ is the natural parameter
- $G(\theta)$ provides normalization

$$\Longrightarrow G(\theta) = \log \sum_{x \in \chi} P_0(x) e^{x\theta} \tag{3}$$

- The derivative of $G(\theta)$, which is denoted by $g(\theta)$ gives the expectation value of $x$ given the parameter value $\theta$.

$$g(\theta) \doteq G'(\theta) = E[x|\theta] \tag{4}$$

- $g(\theta)$ is called the expectation parameter.

## Exponential Family—Examples

Normal distribution

- $\log P(x|\theta) = -\log\sqrt{2\pi} - \frac{1}{2}(x-\theta)^2$
- $\log P_0(x) = -\log\sqrt{2\pi} - x^2/2$, $\theta = \mu$, and $G(\theta) = \theta^2/2$

Bernoulli distribution

- $P(x|p) = p^x(1-p)^{(1-x)}$, where $p \in [0,1]$
- $\log P_0(x) = 1$, $\theta = \log\frac{p}{1-p}$, and $G(\theta) = \log(1+e^\theta)$

## Generalized Linear Models

The regression setup: a group of training samples $(\vec{x}_i, y_i)$ is given. The problem is to predict $y$ when given $\vec{x}$.

Linear regression:

- $y_i$ is approximated by $\vec{\beta} \cdot \vec{x}_i$
- The parameter $\vec{\beta}$ is set to $\arg\min_{\vec{\beta} \in \mathbf{R}^d} \sum_i (y_i - \vec{\beta} \cdot \vec{x}_i)^2$

Generalized linear model:

- $h(\vec{\beta} \cdot \vec{x}_i)$ is taken to approximate the expectation parameter of the exponential model
- $h$ is the inverse of the “link function”. The coice $h = g$ is called “canonical link”
- With canonical link $\vec{\beta} \cdot \vec{x}_i$ is directly an approximation for the natural parameters of the exponential model.

## Bregman Distances

Let $F : \Delta \rightarrow \mathbf{R}$ be differentiable and strictly convex in a convex set $\Delta \subset \mathbf{R}$. Bregman distance associated with $F$, defined for points $p, q \in \Delta$ is

$$B_F(p||q) \doteq F(p) - F(q) - f(q)(p - q) \tag{5}$$

where $f(x) = F'(x)$.

- For exponential family the log-likelihood $\log P(x|\theta)$ is related to a Bregman distance.
- Define a "dual" $F$ through $G$ by

$$F(g(\theta)) + G(\theta) = g(\theta)\theta \tag{6}$$

- It turns out that

$$-\log P(x|\theta) = -\log P_0(x) - F(x) + B_F(x||g(\theta)) \tag{7}$$

# From probability distribution to Bregman distance

| | normal | Bernoulli | Poisson |
|---|---|---|---|
| $\mathcal{X}$ | $\mathbb{R}$ | $\{0,1\}$ | $\{0,1,2\ldots\infty\}$ |
| $G(\theta)$ | $\theta^2/2$ | $\log(1+e^\theta)$ | $e^\theta$ |
| $g(\theta)$ | $\theta$ | $\frac{e^\theta}{(1+e^\theta)}$ | $e^\theta$ |
| $F(x)$ | $x^2/2$ | $x\log(x)+(1-x)\log(1-x)$ | $x\log(x)-x$ |
| $f(x)=g^{-1}(x)$ | $x$ | $\log\frac{x}{1-x}$ | $\log x$ |
| $B_F(p \parallel q)$ | $(p-q)^2/2$ | $p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q}$ | $p\log\frac{p}{q}+q-p$ |
| $B_F(x \parallel g(\theta))$ | $(x-\theta)^2/2$ | $\log(1+e^{-x^*\theta})$ where $x^*=2x-1$ | $e^\theta-x\theta+x\log x-x$ |

## Generalized PCA—Concepts

The idea is to find natural parameters $\vec{\theta}_i$ that are close to the data $\vec{x}_i$, and lie on a low dimensional subspace.

More formally:

- Search for a basis $\vec{v}_1$,. . . ,$\vec{v}_l$ in $\mathbf{R}^d$
- Represent each $\vec{\theta}_i$ as the linear combination of these elements $\vec{\theta}_i = \sum_k a_{ik}\vec{v}_k$ that is "closest" to $\vec{x}_i$.

Let $\mathbf{X}$ be the $n \times d$ matrix with rows $\vec{x}_i$. Let $\mathbf{V}$ be the $l \times d$ matrix with rows $\vec{v}_k$, and $\mathbf{A}$ the $n \times l$ matrix with elements $a_{ik}$. Then the natural parameters $\vec{\theta}_i$ are in the rows of the matrix $\mathbf{\Theta} = \mathbf{AV}$.

## Generalized PCA—Concepts

- The natural parameters $\mathbf{\Theta}$ define the conditional probability of the data.
- The negative log-likelihood is taken as the loss function

$$L(\mathbf{V}, \mathbf{A}) = -\log P(\mathbf{X}|\mathbf{A}, \mathbf{V}) = -\sum_i \sum_j \log P(x_{ij}|\theta_{ij}) \qquad (8)$$

- Equation (7) leads to the following form for the loss function

$$L(\mathbf{V}, \mathbf{A}) = \sum_i \sum_j B_F(x_{ij}||\ g(\theta_{ij})) = \sum_i B_F(\vec{x}_i||\ g(\vec{\theta}_i)) \qquad (9)$$

The generalized PCA can be seen as a search for low dimensional surface $Q(\mathbf{V})$, that passes near all the points $\vec{x}_i$ (in terms of the Bregman distance $B_F$), given by by $Q(\mathbf{V}) = \{g(\vec{a}\mathbf{V})|\vec{a} \in \mathbf{R}^l\}$.

## Generalized PCA—Summary

- The loss function is the negative log likelihood
- The matrix $\mathbf{\Theta} = \mathbf{AV}$ is the matrix of natural parameter values
- The derivative $g(\theta)$ of $G(\theta)$ maps the natural parameters to a matrix of expectation parameters, $g(\mathbf{AV})$
- The function $F$ is derived in terms of $G$, and from it further the Bregman distance $B_F$.
- Now the loss can be written in terms of the Bregman distances $B_F$ alone.

## Generalized PCA—a Minimization Algorithm

The simplest case: search for a one dimensional subspace ($l = 1$)

For $i = 1 \ldots n :\ a_i^{(t)} = \arg\min_{a \in \mathbf{R}} \sum_j B_F(x_{ij} || g(a v_j^{(t-1)}))$

For $j = 1 \ldots d :\ v_j^{(t)} = \arg\min_{v \in \mathbf{R}} \sum_i B_F(x_{ij} || g(a_i^t v)$

- $n + d$ problems, each of which is essentially a very simple GLM regression problem.

## Generalized PCA—a Minimization Algorithm

- One possibility to multiple component optimization is to cycle through the $l$ components, keeping all but one fixed at any given time.

**//Initialization**

Set $\mathbf{A} = \mathbf{0}$, $\mathbf{V} = \mathbf{0}$

**//Cycle through $\ell$ components $N$ times**

For $n = 1, \ldots, N$, $c = 1, \ldots, \ell$:

**//Now optimize the $c$'th component with other components fixed**

Initialize $\mathbf{v}_c^{(0)}$ randomly, and set $s_{ij} = \sum_{k \neq c} a_{ik} v_{kj}$

For $t = 1, \ldots,$ convergence

For $i = 1, \ldots, n$, $\quad a_{ic}^{(t)} = \arg\min_{a \in \mathbb{R}} \sum_j B_F\left(x_{ij} \parallel g(a v_{cj}^{(t-1)} + s_{ij})\right)$

For $j = 1 \ldots d$, $\quad v_{cj}^{(t)} = \arg\min_{v \in \mathbb{R}} \sum_i B_F\left(x_{ij} \parallel g(a_{ic}^{(t)} v + s_{ij})\right)$

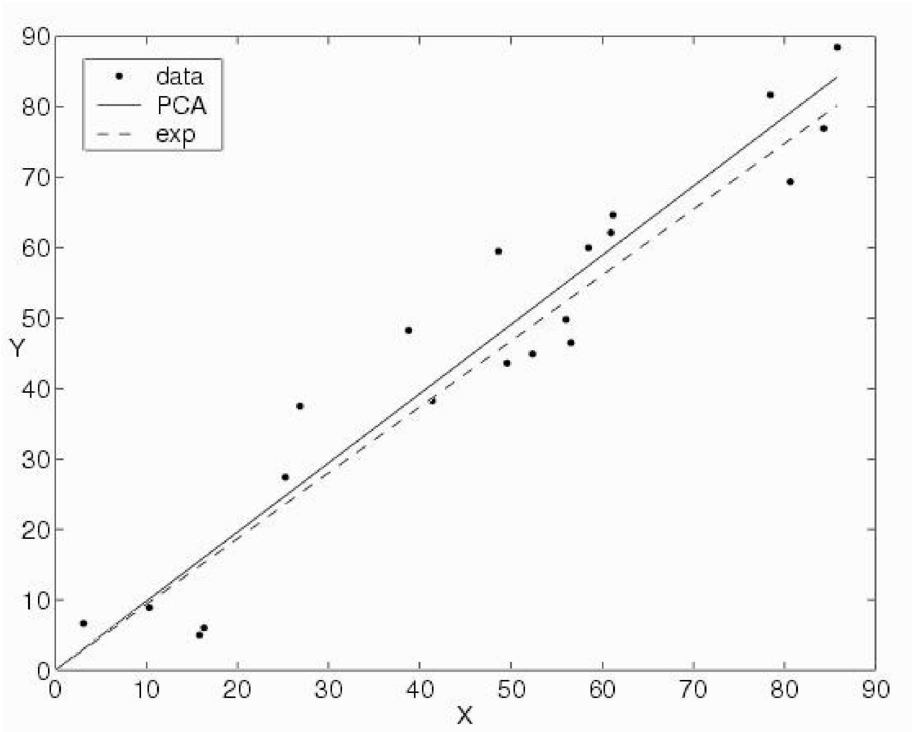
## Examples—Exponential distribution

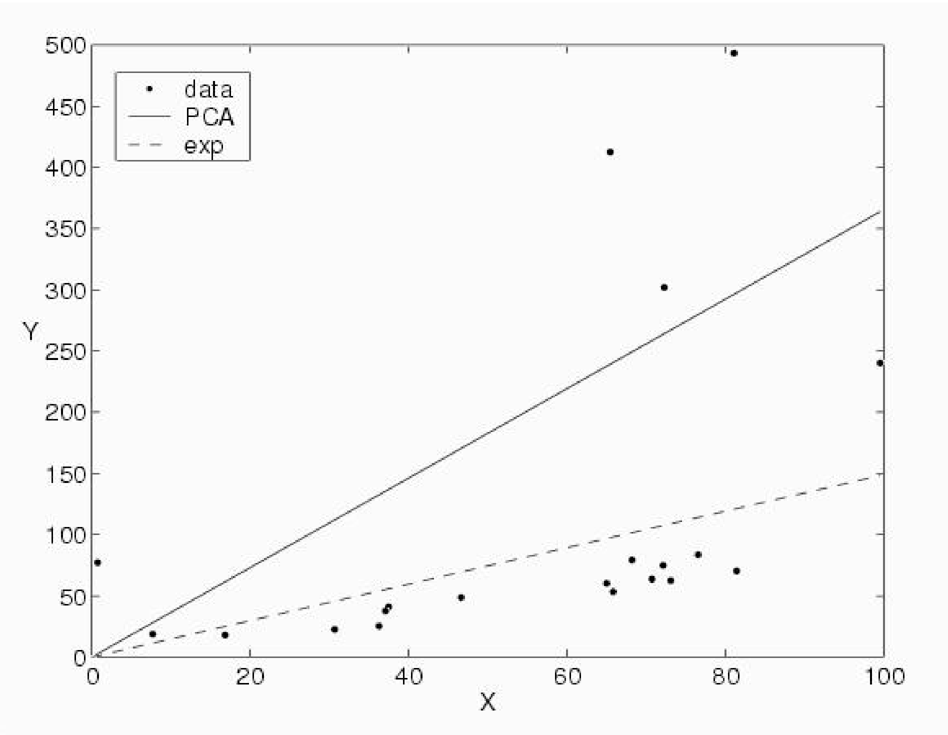
- Given nonnegative data $\mathbf{X} \in \mathbf{R}^{n\times d}$ we want the best one dimensional approximation
- Find a vector $\vec{v}$ and coefficients $\vec{a}$ such that the approximation $\vec{x}_i \approx g(a_i\vec{v})$ has minimum loss
- Closed form update rule turns out to be

$$\frac{1}{\vec{v}} \longleftarrow \frac{n}{d}\mathbf{X}^T \cdot \frac{1}{\mathbf{X}\vec{v}}, \tag{10}$$

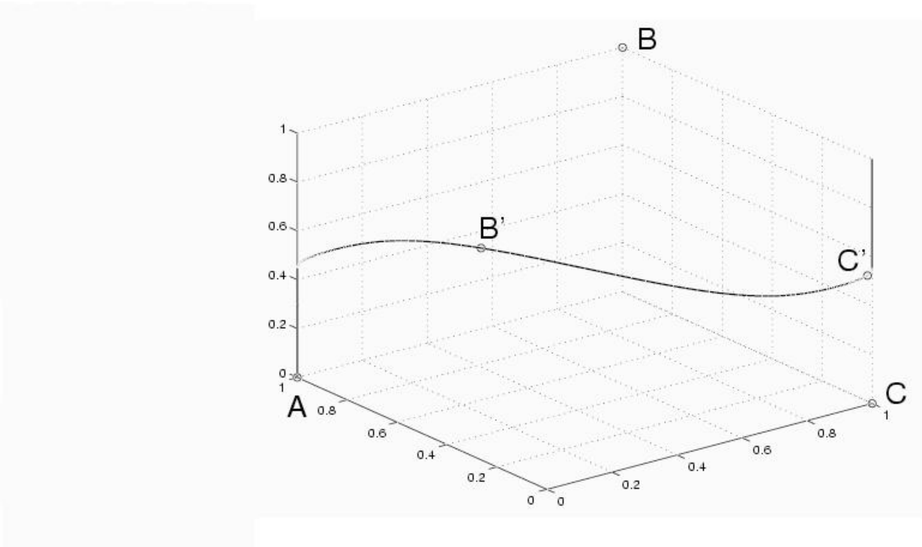
  where $\frac{1}{\vec{v}}$ means componentwise reciprocal
- The link function in this case is $g(\theta) = -\frac{1}{\theta}$ (naturally the mean of the distribution).
- Thus points of the form $g(a_i\vec{v})$ lie on a straight line and comparison to ordinary PCA becomes meaningful

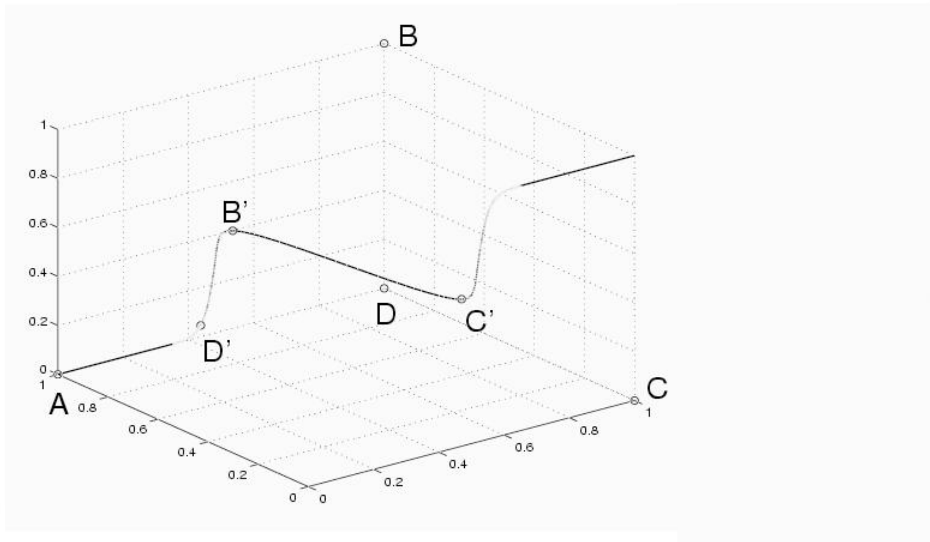
# Examples—Exponential distribution



## Examples—Bernoulli distribution




- A mapping of $\{0,1\}^3$ cube to one dimension via the generalized PCA
- Here the linear subspace of the natural parameter space is mapped by $g(\theta)$ to a nonlinear curve in the cube. Note the symmetry around $(1/2, 1/2, 1/2)$