T-122.102 Co-occurrence methods in analysis of discrete data
A Generalization of Principal Component Analysis to the Exponential Family [1]

Original article: Michael Collins,
Sanjoy Dasgupta, and Robert E. Shcapire
Summary: Mikko Heikelä

March 8, 2004

1 Introduction

Collins, Dasgupta, and Shcapire present a way to generalize the popuar dimensionality reduction method principal component analysis (PCA) to exponential families of distributions [1]. First they present a probabilistic interpretation of PCA that will lead to the generalization. Exponential families, generalized linear models (GLM), and Bregman distances are preliminaries that are briefly studied before formulating the generalized PCA. A minimization algorithm is presented and two simple examples are discussed.

Perhaps the most videly used formulation of the PCA problem is as a search for a linear subspace that passes near all the data points. For given data $\mathbf{x}_i \in \mathbb{R}^d$ the lower dimensional subspace that minimizes the sum of squared distances between $\mathbf{x}_i$ and their projections $\boldsymbol{\theta}_i$ to it is found. The cost function is the sum of euclidean distances:

$$\sum_{i=1}^{n} \|\mathbf{x_i} - \boldsymbol{\theta}_i\|^2. \tag{1}$$

In a probabilistic alternative each $\mathbf{x}_i$ is seen as drawn from a unit gaussian $P_{\boldsymbol{\theta}_i}$ with unknown mean $\boldsymbol{\theta}_i$. Maximizing the likelihood of the data subject to the condition that $\boldsymbol{\theta}_i$ belong to a low dimensional subspace is equivalent to equation (1). In this view the data points $\mathbf{x}_i$ are noise-corrupted versions of actual points $\boldsymbol{\theta}_i$ in a linear subspace, where the noise is understood to be gaussian noise with unit variance.

This interpretation that contains gaussian noise is not natural e.g. when the data is discrete valued or nonnegative. Gaussian distribution is just one of the distributions that make up the exponential family, and it is the particular one that suits real valued data. Other distributions in the exponential family can describe other types of data, e.g. Poisson—integer, Bernoulli—binary.

Collins et. al. show that a general dimensionality reduction scheme for the exponential family can be devised. In many parts of the generalization the dimensions of the data are treated separately, and it turns out that hybrid cases where the data contains different types of dimensions are permitted. In general there is a crucial difference to ordinary PCA: the natural parameter space of the distribution family from which the data is drawn and the space of the data are not the same. A mapping between these is needed. This difference can be bridged after looking at generalized linear models (GLM), exponential families and Bregman distances.

2 Some Theoretical Background

In the exponential family[1] the conditional probability of a data point x is given by

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

where θ is the natural parameter, and $G(\theta)$ provides normalization. Therefore $G(\theta)$ must be given by

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

where χ is the domain of x, and the sum is naturally replaced by an integral if x is continuous valued.

An important funtion is the derivative of $G(\theta)$, which is denoted by $g(\theta)$. $g(\theta)$ gives the expectation value of x given the parameter value θ

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

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

[1] The term exponential family is used both to describe all distributions that can be written in the form of equation (2), and for particular families of distributions parameterized by θ, where all the functions in the this equation are fixed. There should be no room for confusion.

As an example it is shown how normal and Bernoulli distributions fit this picture. The normal distribution can be written as $\log P(x|\mu) = -\log\sqrt{2\pi} - \frac{1}{2}(x-\mu)^2$. From this we can read $\log P_0(x) = -\log\sqrt{2\pi} - x^2/2$, $\theta = \mu$, and $G(\theta) = \theta^2/2$. In the Bernoulli distribution the probability is usually written as $P(x|p) = p^x(1-p)^{(1-x)}$, where $p \in [0,1]$ and $x \in \{0,1\}$. This is translated to the exponential form as follows: $\log P_0(x) = 1$, $\theta = \log\frac{p}{1-p}$, and $G(\theta) = \log(1+e^{\theta})$.

In the previous discussion the parameter θ and the data point x have both been taken as one-dimensional. Although we are deriving a dimensionality reduction method we do not need to consider the multidimensional case here because each data dimension has its own unique natural parameter dimension. The Bregman distance that is derived later and used as a loss function can be computed as a sum of the componentwise distances. The vector version of much of the following discussion can be found in [2].

2.1 Generalized Linear Models

The generalization of PCA by Collins et. al. is analogous to the way in which generalized linear models (GLM's) generalize regression. In a regression problem a group of training samples $(\mathbf{x}_i, y_i)$ is given. The problem is to predict y when given $\mathbf{x}$. In linear regression y_i is approximated by $\boldsymbol{\beta}\cdot\mathbf{x}_i$. The parameter $\boldsymbol{\beta}$ is set to $\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d}\sum_i(y_i - \boldsymbol{\beta}\cdot\mathbf{x}_i)^2$.

In a GLM $h(\boldsymbol{\beta}\cdot\mathbf{x}_i)$ is taken to approximate the expectation parameter of the exponential model. h is the inverse of the "link function". The coice h = g (as defined before) is called "canonical link". With canonical link $\boldsymbol{\beta}\cdot\mathbf{x}_i$ is directly an approximation for the natural parameters of the exponential model.

2.2 Bregman Distances

In the generalized PCA the euclidean distances of PCA are replaced with Bregman distances related to the underlying exponential probability distribution. Bregman distances are defined as follows: Let $F : \Delta \to \mathbb{R}$ be a differentiable and strictly convex function in a convex set $\Delta \subset \mathbb{R}$. The 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 directly related to a Bregman distance. Following the discussion in [2], a "dual" function F can be defined through G by

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

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

Table 1: The derivation of a Bregman distance related to the log-likelihood of various exponential family distributions.

It turns out that

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

where F is the dual of G as defined above, and B_F the Bregman distance derived from it. This form is very useful for likelihood maximization because on the right hand side there is θ dependence only in the Bregman distance term, and the other terms can therefore be neglected in optimization. We have here arrived at a systematic procedure to derive an expression for log-likelihood in exponential family in terms of a Bregman distance. The results of applying this procedure to some distributions is shown in table 1.

3 Generalized PCA

In an analogy with the probabilistic interpretation of ordinary PCA, the idea of the generalized algorithm is to find natural parameters $\boldsymbol{\theta}_i$ that lie in a low dimensional subspace, and are close to the data $\mathbf{x}_i$ in the sense that the likelihood is maximized.

More formally the problem is to search for a basis $\mathbf{v}_1,\ldots,\mathbf{v}_l$ in $\mathbb{R}^d$, and a representation of each $\boldsymbol{\theta}_i$ as a linear combination of these elements $\boldsymbol{\theta}_i = \sum_k a_{ik}\mathbf{v}_k$ in such a way that the likelihood of the data is mazimized. As discussed previously this amounts to minimizing the sum of Bregman distances from the data to the expectation images $g(\theta)$ of the natural parameters θ.

In describing the method and a minimization algorithm the following notations are used. Let $\mathbf{X}$ be the $n \times d$ matrix with the data points $\mathbf{x}_i$ as its rows. Let $\mathbf{V}$ be the $l \times d$ matrix with rows $\mathbf{v}_k$, and $\mathbf{A}$ the $n \times l$ matrix with elements a_{ik}. Then the natural parameters $\boldsymbol{\theta}_i$ are in the rows of the matrix $\boldsymbol{\Theta} = \mathbf{AV}$.

In the generalized PCA, the loss function is taken to be the negative log-likelihood of the data, which depends of the matrix of the natural parameters $\boldsymbol{\Theta}$, and therefore on the matrices $\mathbf{V}$, and $\mathbf{A}$:

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

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

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

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

As a summary:

- The loss function is the negative log likelihood
- The matrix $\boldsymbol{\Theta} = \mathbf{AV}$ 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.

4 Algorithm

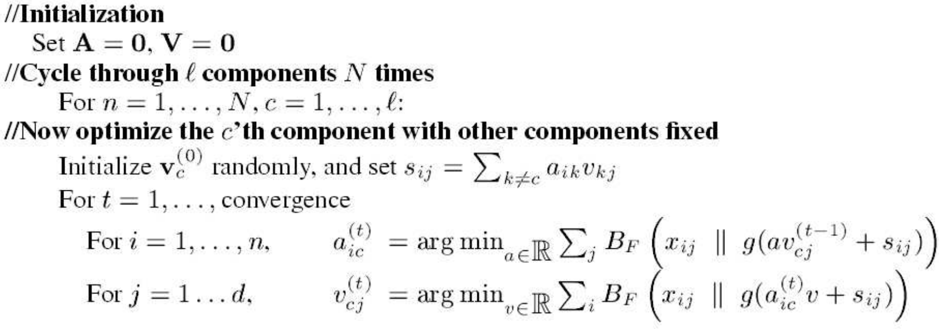
The minimization algorithm the writers propose is discussed in this section. The simplest case is a search for a one dimensional subspace ($l = 1$). In this case the minimum of the loss function is searched for by starting with randomized $\mathbf{V}$ and iterating the following two steps until convergence is observed.

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

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

Each loop here consists of $n+d$ problems, each of which is essentially a very simple GLM regression problem (simple because there is only one parameter to be optimized over). The loss is convex in eihter of its argumets alone, but not in general when they are considered together. Thus, convergence is not easy to prove. The gaussian case is known better than others, and there it is known that the hessian of the loss is only positive semi-definite for the global minimum [3]. Whether this is true in the general case is an open problem as well as the convergence to the global optimum in a general case. The algorithm has behaved well in this respect in preliminary numerical studies.

One possibility to multiple component optimization is to cycle through the l components, keeping all but one fixed at any given time. This approach leads to the algorithm below:



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

Multiple rounds over the components are needed because unlike in odinary PCA the later components affect the most important components.

5 Examples

The authors give two very simple illustrative examples which are repeated here. First they consider finding a one-dimensional subspace of two-dimensional data, using the exponential distribution suited for nonnegative data. In this case the minimization procedure can be written in closed form and it turns out that the mapping $g(\theta)$ from natural parameters to the expectation parameters is such that the image of a linear subspace of the natural parameters is a straight line. This is similar as in ordinary PCA, and it is interesting to compare these cases.

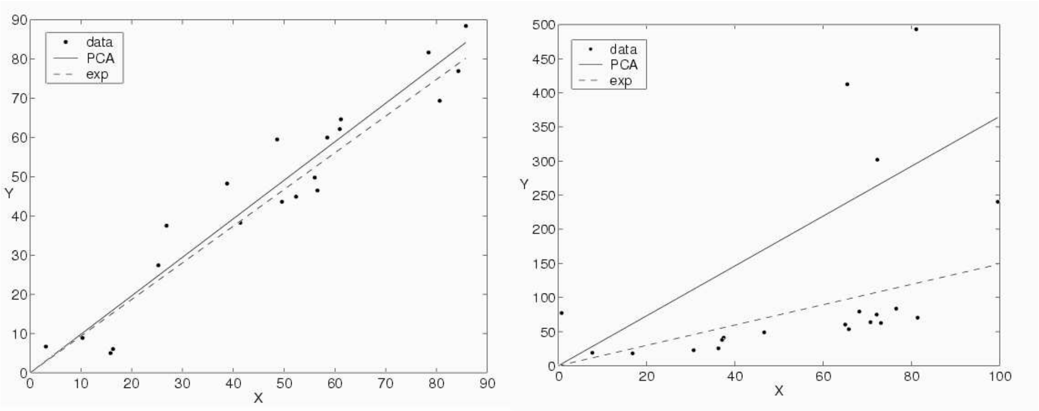


Figure 1: Comparison of regular PCA and the PCA for exponential distribution.

Figure 1 shows the comparison. On the left the data points fit well on a straight line, and both methods give similar results. On the right there are some outliers, and it turns out that ordinary PCA is much more sensitive to them than the exponential variant.

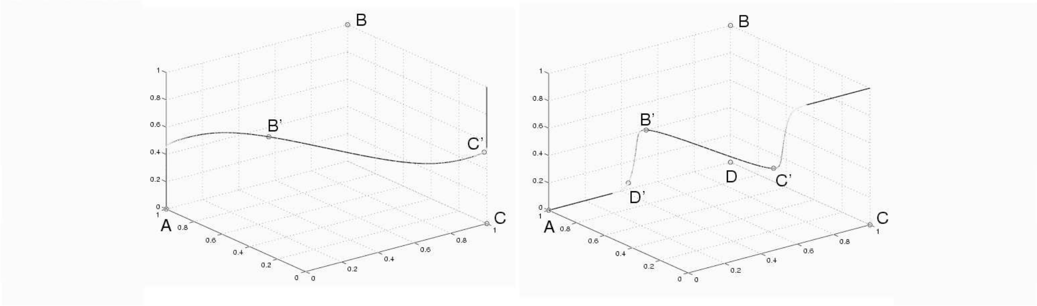


Figure 2: Projections from 3- to 1-dimensional space with Bernoulli PCA. The data points are marked by the capital letters, and their projections to the curve (as given by the minimum Bregman distance) are marked with the primed capitals.

The other example is finding a one dimensional subspace of the parameters of three-dimensional Bernoulli distribution. In this case the linear subspace of the natural parameters, is mapped by $g(\theta)$ to a nonlinear curve on the space of the data. Results for this toy experiment are shown in figure 5. In this case the choice of data points is perhaps a bit unfortunate as they are about as far away from a common subspace as possible. The example thus fails to illustrate the form of the solution in a case where dimensionality reduction would be possible without severe distortion.

References

[1] M. Collins, S. Dasgupta, R. Schapire. A generalization of principal component analysis to the exponential family. In: NIPS*13 (2001)

[2] Katy S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. Machine Learning, 43:211-246 (2001)

[3] I. T. Jolliffe. Principal Component Analysis. Springer-Verlag (1986)