A Unifying Viewpoint of some Clustering Techniques Using Bregman Divergences and Extensions to Mixed Data Sets

Cécile Levasseur, Brandon Burdge, Ken Kreutz-Delgado
University of California, San Diego
Jacobs School of Engineering
La Jolla, California, USA
{clevasseur, bburdge, kreutz}@ucsd.edu

Uwe F. Mayer
University of Utah
Department of Mathematics
Salt Lake City, Utah, USA
mayer@math.utah.edu

Abstract

We present a general viewpoint using Bregman divergences and exponential family properties that contains as special cases the three following algorithms: 1) exponential family Principal Component Analysis (exponential PCA), 2) Semi-Parametric exponential family Principal Component Analysis (SP-PCA) and 3) Bregman soft clustering. This framework is equivalent to a mixed data-type hierarchical Bayes graphical model assumption with latent variables constrained to a low-dimensional parameter subspace. We show that within this framework exponential PCA and SP-PCA are similar to the Bregman soft clustering technique with the addition of a linear constraint in the parameter space. We implement the resulting modifications to SP-PCA and Bregman soft clustering for mixed (continuous and/or discrete) data sets, and add a nonparametric estimation of the point-mass probabilities to exponential PCA. Finally, we compare the relative performances of the three algorithms in a clustering setting for mixed data sets.

1. Introduction

We present a general point of view that relates the exponential family Principal Component Analysis (exponential PCA) technique of [7] to the Semi-Parametric Principal Component Analysis (SP-PCA) technique of [20] and to the Bregman soft clustering method presented in [4]. The proposed standpoint is then illustrated with a clustering problem in mixed data sets.

The three techniques considered here all utilize Bregman divergences and can all be explained within a single hierarchical Bayes graphical model framework shown in Figure 1. They are not separate unrelated algorithms but different manifestations of parameter choices taken within a common framework. The proposed model is mathematically equivalent to equation (6) and we will demonstrate that various parametric choices symbolically shown in Figure 2 determine the three algorithms. Because of this insight, we will readily extend the algorithms to deal with the important mixed data type case.

There are two ways in which Bregman divergences are important. First, they generalize the squared Euclidean distance to a class of distances that all share similar properties. Second, there exists a bijection between Bregman divergences and exponential family distributions [3, 4]. Recently, researchers have shown that many important algorithms can be generalized from Euclidean metrics to distances defined by a Bregman divergence [7, 20, 4, 6], i.e., from Gaussian distributed data components to data components distributed according to an exponential family, such as binary- or integer-valued.

Data mining techniques seek potentially useful information contained in complex data sets. Complexity of these data generally comes from a high number of components and also from the fact that the components usually are of mixed data types (categorical, count, continuous, etc.). In order to address the latter, techniques should allow for the components to have different parametric forms by using the large range of exponential family distributions and their associated Bregman divergences. While a modification for mixed data sets was presented in [12] for exponential PCA, we implement here a modification for SP-PCA and Bregman soft clustering. The issue of a high number of components of the original data is addressed by exponential PCA and SP-PCA through dimensionality reduction by projecting the data to a low-dimensional parameter subspace.

Finally, we consider synthetic data examples of mixed types. We demonstrate that exponential PCA, with the addition of a nonparametric estimation tool, rivals SP-PCA and Bregman soft clustering in terms of clustering performance.

2. Theoretical background

To motivate theoretical developments, the hierarchical Bayes graphical model for hidden or latent variables shown in Figure 1 is considered.

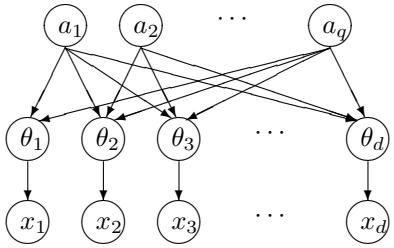

Figure 1. Graphical model for our framework.

The row vector $\mathbf{x} = [x_1, \ldots, x_d] \in \mathbb{R}^d$ consists of observed features in a d-dimensional space. It is assumed that training points can be drawn from populations having class-conditional probability density functions

$$p(\mathbf{x}|\boldsymbol{\theta}) = p_1(x_1|\theta_1) \cdot \ldots \cdot p_d(x_d|\theta_d), \tag{1}$$

where, when conditioned on the random parameter vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_d] \in \mathbb{R}^d$, the components of $\mathbf{x}$ are independent. It is further assumed that $\boldsymbol{\theta}$ can be written as

$$\boldsymbol{\theta} = \mathbf{a}\mathbf{V} + \mathbf{b}$$

with the hidden or latent variable $\mathbf{a} = [a_1, \ldots, a_q] \in \mathbb{R}^q$ random and unknown with $q < d$ (and ideally $q \ll d$), $\mathbf{V} \in \mathbb{R}^{q \times d}$ and $\mathbf{b} \in \mathbb{R}^d$ deterministic and unknown. The latent variable $\mathbf{a}$ in some way explains part (or all) of the random behavior of the observed variables. The subscript i on $p_i(\cdot|\cdot)$ serves to indicate that the marginal densities can all be different, allowing for the possibility of $\mathbf{x}$ containing categorical, discrete, and continuous valued components. Also, the marginal densities are each assumed to be one-parameter exponential family densities, and θ_i is taken to be the natural parameter (or some simple bijective function of it) of the exponential family density p_i. Hence, each component density $p_i(x_i|\theta_i)$ in (1) for $x_i \in \mathcal{X}_i, i = 1, \ldots, d$, is of the form

$$p(x_i|\theta_i) = \exp\big(\theta_i x_i - G(\theta_i)\big), \tag{2}$$

where $G(\cdot)$ is the cumulant generating function defined as

$$G(\theta_i) = \log \int_{\mathcal{X}_i} \exp\big(\theta_i x_i\big)\nu(\mathrm{d}x_i), \tag{3}$$

with $\nu(\cdot)$ a σ-finite measure that generates the exponential family. It can be shown, using Fubini's theorem [11], that $G(\boldsymbol{\theta}) = \sum_{i=1}^{d} G(\theta_i)$.

The maximum likelihood identification of the blind random effect model

$$p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \int \prod_{i=1}^{d} p_i(x_i|\theta_i)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{4}$$

where $\pi(\boldsymbol{\theta})$ is the probability density function of $\boldsymbol{\theta}$, is quite a difficult problem. It corresponds to identifying $\pi(\boldsymbol{\theta})$, which, under the condition $\boldsymbol{\theta} = \mathbf{a}\mathbf{V} + \mathbf{b}$, corresponds to identifying the matrix $\mathbf{V}$, the vector $\mathbf{b}$, and a density function on the random effect $\mathbf{a}$ via a maximization of the likelihood function $p(\mathbf{X})$ with respect to $\mathbf{V}$, $\mathbf{b}$, and the random effect density function, where

$$\begin{aligned} p(\mathbf{X}) &= \prod_{k=1}^{n} p\big(\mathbf{x}[k]\big) = \prod_{k=1}^{n} \int p\big(\mathbf{x}[k]|\boldsymbol{\theta}\big)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &= \prod_{k=1}^{n} \int \prod_{i=1}^{d} p_i\big(x_i[k]|\theta_i\big)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{5}$$

and $\mathbf{X}$ is the $(n \times d)$ observation matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}[1] \\ \mathbf{x}[2] \\ \vdots \\ \mathbf{x}[n] \end{pmatrix} = \begin{pmatrix} x_1[1] & \ldots & x_d[1] \\ x_1[2] & \ldots & x_d[2] \\ \vdots & \ddots & \vdots \\ x_1[n] & \ldots & x_d[n] \end{pmatrix}.$$

This difficulty can be avoided by NonParametric Maximum Likelihood (NPML) estimation of the random effect distribution, concurrently with the structural model parameters. The NPML estimate is known to be a discrete distribution on a finite number of support points or "atoms" [9, 10, 13]. Finding the NPML estimate is widely regarded as computationally intensive, the particular difficulty being the location of the atoms [1].

As shown in [12], with $\boldsymbol{\theta} = \mathbf{a}\mathbf{V}+\mathbf{b}$, with $\mathbf{V}$, $\mathbf{b}$ fixed and $\mathbf{a}$ random, the single-sample likelihood (4) is equal to

$$p(\mathbf{x}) = \sum_{l=1}^{m} p\big(\mathbf{x}|\underline{\boldsymbol{\theta}}[l]\big)\pi_l = \sum_{l=1}^{m} p\big(\mathbf{x}|\underline{\mathbf{a}}[l]\mathbf{V} + \mathbf{b}\big)\pi_l$$

and the data likelihood (5) is equal to

$$p(\mathbf{X}) = \prod_{k=1}^{n}\sum_{l=1}^{m} p\big(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[l]\big)\pi_l = \prod_{k=1}^{n}\sum_{l=1}^{m} p\big(\mathbf{x}[k]|\underline{\mathbf{a}}[l]\mathbf{V}+\mathbf{b}\big)\pi_l, \tag{6}$$

with point-mass probability estimates π_l, unknown point-mass support points $\underline{\mathbf{a}}[l]$, and the linear predictor $\underline{\boldsymbol{\theta}}[l] = \underline{\mathbf{a}}[l]\mathbf{V} + \mathbf{b}$ in the lth mixture component, $l = 1, \ldots, m$.

The data likelihood is thus approximately the likelihood of a finite mixture of exponential family densities with unknown mixture proportions or point-mass probability estimates π_l and unknown point-mass support points $\underline{\mathbf{a}}[l]$,

with the linear predictor $\underline{\boldsymbol{\theta}}[l] = \underline{\mathbf{a}}[l]\mathbf{V} + \mathbf{b}$ in the lth mixture component [2]. The combined problem of Maximum Likelihood Estimation (MLE) of the parameters $\mathbf{V}$, $\mathbf{b}$, the point-mass support points (atoms) $\underline{\mathbf{a}}[l]$ and the point-mass probability estimates $\pi_l, l = 1, \ldots, m$, is known as the Semiparametric Maximum Likelihood mixture density Estimation (SMLE) problem [14, 16]. It can be attacked by using the Expectation-Maximization (EM) algorithm [8, 10, 13, 19, 15, 2, 5, 16, 18], as done in particular in the Semi-Parametric Principal Component Analysis (SP-PCA) technique proposed in [20] and discussed below. Note that, historically, [10] appears to be generally acknowledged as the first paper that proposed the EM algorithm for NPML estimation in the mixture density context; then, [13] improved upon the theoretical foundations of the NPML estimation approach and later [15] further explored some of the fundamental issues raised in [13].

However, the SMLE problem can also be attacked by simply considering the special case of uniform point-mass probabilities, i.e., $\pi_l = 1/m$ for $l = 1, \ldots, m$, for which the number of support points equals the number of data samples, i.e., $m = n$. It was demonstrated in [12] that this special uniform case corresponds to exponential PCA.

3. A unifying framework

Within the proposed hierarchical Bayes graphical model framework, exponential PCA, SP-PCA and Bregman soft clustering are not separate uncorrelated algorithms but different manifestations of parameter choices.

Figure 2 considers the number of atoms as a common characteristic for comparison purposes. It symbolizes how various parametric choices determine the three algorithms. Whereas the exponential PCA approach requires a number of atoms equal to the number of data points and hence can be seen as an extreme case of the NPML technique, SP-PCA deals with a smaller number of atoms. Finally, the Bregman soft clustering approach considers an even smaller number of atoms, viewed as cluster centers, since its primary goal is clustering. Furthermore, both exponential PCA and SP-PCA impose a low-dimensional (unknown) latent variable subspace in their structure. However, Bregman soft clustering does not impose this lower dimensional constraint and hence can be seen as a degenerate case.

It becomes clear while looking at Table 1, Table 2 and Table 3 shown below that both SP-PCA and Bregman soft clustering utilize the EM algorithm for estimation purposes whereas exponential PCA does not. This is because exponential PCA assumes uniform point-mass probabilities and does not need to estimate them. Both exponential PCA and SP-PCA impose the low-dimensional parameter subspace constraint, hence the need for the Newton-Raphson iterative algorithm to find it as discussed below.

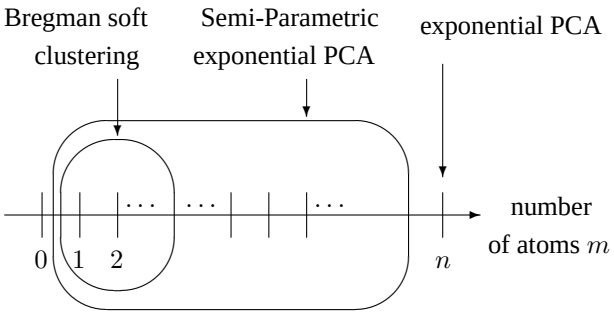

Figure 2. General point of view on SP-PCA, exponential PCA and Bregman soft clustering based on the number of NPML atoms.

4. Extensions to mixed data sets

Now that the relationship between exponential PCA, SP-PCA and Bregman soft clustering has been demonstrated within a unifying mixed data type framework, we can extend SP-PCA and Bregman soft clustering to deal with the important mixed data case. Recalling the modifications on exponential PCA presented in [12], we also introduce a penalty function that improves exponential PCA convergence efficiency. Towards these ends, following the notation in [5], we use the *mixing distribution* $\mathcal{Q}=\left\{\underline{\boldsymbol{\theta}}[l], \pi_l\right\}_{l=1}^{m}$ which contains the parameters $\underline{\boldsymbol{\theta}}[l]$ and their associated point-mass probabilities π_l. The mixing distribution needs to be estimated in all three algorithms.

4.1. Semi-parametric exponential PCA

The Semi-Parametric exponential family Principal Component Analysis (SP-PCA) approach presented in [20] attacks the Semiparametric Maximum Likelihood mixture density Estimation (SMLE) problem by using the Expectation-Maximization (EM) algorithm [8]. As done previously in [12] for exponential PCA, we present an SP-PCA modified approach for mixed data types. For simplicity of presentation, we consider that the f first attributes are distributed according to the exponential family distribution $p^{(\mathbf{1})}$ and the $(d-f)$ last attributes are distributed according to the exponential family distribution $p^{(\mathbf{2})}$. The following notation is used:

$$\begin{aligned}\mathbf{x}[k] &= \left[x_1[k], \ldots, x_f[k], x_{f+1}[k], \ldots, x_d[k]\right] \\ &= \left[\mathbf{x}^{(\mathbf{1})}[k] \middle| \mathbf{x}^{(\mathbf{2})}[k]\right],\end{aligned}$$

for $k = 1, \ldots, n$, and similarly for the observation matrix,

$$\mathbf{X} = \left(\mathbf{X}^{(\mathbf{1})} \;\middle|\; \mathbf{X}^{(\mathbf{2})}\right).$$

The EM approach introduces a *missing* (unobserved) variable $\mathbf{z}_k = [z_{k1}, \ldots, z_{km}]$, for $k = 1, \ldots, n$. This variable is an m-dimensional binary vector whose lth component equals 1 if the observed variable $\mathbf{x}[k]$ was drawn from the lth mixture component and 0 otherwise; its value is estimated during the E-step. Using this information, a *complete* log-likelihood function is defined as follows:

$$L^{(c)}\big(\mathcal{Q}, \{\mathbf{z}_k\}_{k=1}^n\big) = \log \prod_{k=1}^{n} \prod_{l=1}^{m} p\big(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[l]\big)^{z_{kl}} \pi_l^{z_{kl}},$$

with $\underline{\boldsymbol{\theta}}[l] = \underline{\mathbf{a}}[l]\mathbf{V} + \mathbf{b}$. Its maximization during the M-step yields parameters $\underline{\mathbf{A}}$, $\mathbf{V}$ and $\mathbf{b}$ estimates. Then,

$$p\big(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[l]\big) = p^{(\mathbf{1})}\big(\mathbf{x}^{(\mathbf{1})}[k]|\underline{\boldsymbol{\theta}}^{(\mathbf{1})}[l]\big) \cdot p^{(\mathbf{2})}\big(\mathbf{x}^{(\mathbf{2})}[k]|\underline{\boldsymbol{\theta}}^{(\mathbf{2})}[l]\big) \tag{7}$$

using the assumption from (1), where, for $l = 1, \ldots, m$,

$$\underline{\boldsymbol{\theta}}[l] = \big[\underline{\theta}_1[l], \ldots, \underline{\theta}_f[l], \underline{\theta}_{f+1}[l], \ldots, \underline{\theta}_d[l]\big] = \big[\underline{\boldsymbol{\theta}}^{(\mathbf{1})}[l] \big| \underline{\boldsymbol{\theta}}^{(\mathbf{2})}[l]\big],$$

$$\underline{\boldsymbol{\Theta}} = \begin{pmatrix} \underline{\boldsymbol{\theta}}[1] \\ \underline{\boldsymbol{\theta}}[2] \\ \vdots \\ \underline{\boldsymbol{\theta}}[n] \end{pmatrix} = \big(\underline{\boldsymbol{\Theta}}^{(\mathbf{1})} \;\big|\; \underline{\boldsymbol{\Theta}}^{(\mathbf{2})}\big).$$

The following decompositions arise:

$$\mathbf{V} = \big(\mathbf{V}^{(\mathbf{1})} \;\big|\; \mathbf{V}^{(\mathbf{2})}\big),$$

$$\mathbf{B} = \big(\mathbf{B}^{(\mathbf{1})} \;\big|\; \mathbf{B}^{(\mathbf{2})}\big),$$

where $\mathbf{B}^{(\mathbf{1})} = [\mathbf{b}^{(\mathbf{1})}, \ldots, \mathbf{b}^{(\mathbf{1})}]^T$ and $\mathbf{b}^{(\mathbf{1})} = [b_1, \ldots, b_f]$, $\mathbf{B}^{(\mathbf{2})} = [\mathbf{b}^{(\mathbf{2})}, \ldots, \mathbf{b}^{(\mathbf{2})}]^T$ and $\mathbf{b}^{(\mathbf{2})} = [b_{f+1}, \ldots, b_d]$. Hence,

$$\underline{\boldsymbol{\Theta}} = \big(\underline{\mathbf{A}}\mathbf{V}^{(\mathbf{1})} + \mathbf{B}^{(\mathbf{1})} \;\big|\; \underline{\mathbf{A}}\mathbf{V}^{(\mathbf{2})} + \mathbf{B}^{(\mathbf{2})}\big).$$

Note that there is no such split for $\underline{\mathbf{A}}$. The complete log-likelihood function becomes:

$$\begin{aligned} L^{(c)}\big(\mathcal{Q}, \{\mathbf{z}_k\}_{k=1}^n\big) &= \sum_{k=1}^{n}\sum_{l=1}^{m} z_{kl} \log \pi_l \\ &+ \sum_{k=1}^{n}\sum_{l=1}^{m} z_{kl} \log p^{(\mathbf{1})}\big(\mathbf{x}^{(\mathbf{1})}[k]|\underline{\boldsymbol{\theta}}^{(\mathbf{1})}[l]\big) \cdot p^{(\mathbf{2})}\big(\mathbf{x}^{(\mathbf{2})}[k]|\underline{\boldsymbol{\theta}}^{(\mathbf{2})}[l]\big). \end{aligned} \tag{8}$$

The E-step yields for $k = 1, \ldots, n$ and $l = 1, \ldots, m$:

$$\begin{aligned} \widehat{z}_{kl} &= \mathrm{E}\left\{z_{kl}|\mathbf{x}[k], \pi_1, \ldots, \pi_m\right\} \\ &= \frac{p^{(\mathbf{1})}\big(\mathbf{x}^{(\mathbf{1})}[k]|\underline{\boldsymbol{\theta}}^{(\mathbf{1})}[l]\big) \cdot p^{(\mathbf{2})}\big(\mathbf{x}^{(\mathbf{2})}[k]|\underline{\boldsymbol{\theta}}^{(\mathbf{2})}[l]\big)\big)\pi_l}{\sum_{r=1}^{m} p^{(\mathbf{1})}\big(\mathbf{x}^{(\mathbf{1})}[k]|\underline{\boldsymbol{\theta}}^{(\mathbf{1})}[r]\big) \cdot p^{(\mathbf{2})}\big(\mathbf{x}^{(\mathbf{2})}[k]|\underline{\boldsymbol{\theta}}^{(\mathbf{2})}[r]\big)\pi_r}. \end{aligned}$$

For all l and all k, each data point $\mathbf{x}[k]$ has an estimated probability $\widehat{z}_{kl}$ of belonging to the lth mixture component.

The M-step first yields the estimates for the point-mass probabilities:

$$\pi_l^{(new)} = \frac{\sum_{k=1}^{n} \widehat{z}_{kl}}{n}.$$

The second part of the M-step, i.e., the estimation of the parameters $\mathbf{V}$, $\mathbf{b}$, and the point-mass support points $\underline{\mathbf{A}} = \big[\underline{\mathbf{a}}[1]^T, \ldots, \underline{\mathbf{a}}[m]^T\big]^T \in \mathbb{R}^{m,q}$, is affected by the mixed data type assumption. It consists of maximizing the complete log-likelihood function (8) with respect to these parameters:

$$\underset{\underline{\mathbf{A}}, \mathbf{V}, \mathbf{b}}{\arg\max}\, \mathrm{E}\left\{L^{(c)}\big(\big\{\underline{\boldsymbol{\theta}}[l], \pi_l^{(new)}\big\}_{l=1}^m, \{\widehat{\mathbf{z}}_k\}_{k=1}^n\big)\right\}.$$

Following the notation in [20], we set, for $l = 1, \ldots, m$:

$$\widetilde{\mathbf{x}}[l] = \frac{\sum_{k=1}^{n} \widehat{z}_{kl}\mathbf{x}[k]}{\sum_{k=1}^{n} \widehat{z}_{kl}},$$

the lth mixture component center. It can be shown, using exponential family properties, that the loss function is:

$$\begin{aligned} &L(\underline{\mathbf{A}}, \mathbf{V}, \mathbf{b}) \\ &= \sum_{l=1}^{m}\Big\{G^{(\mathbf{1})}\big(\underline{\mathbf{a}}[l]\mathbf{V}^{(\mathbf{1})} + \mathbf{b}^{(\mathbf{1})}\big) - \big(\underline{\mathbf{a}}[l]\mathbf{V}^{(\mathbf{1})} + \mathbf{b}^{(\mathbf{1})}\big)\widetilde{\mathbf{x}}[l]^T\Big\} \\ &+ \sum_{l=1}^{m}\Big\{G^{(\mathbf{2})}\big(\underline{\mathbf{a}}[l]\mathbf{V}^{(\mathbf{2})} + \mathbf{b}^{(\mathbf{2})}\big) - \big(\underline{\mathbf{a}}[l]\mathbf{V}^{(\mathbf{2})} + \mathbf{b}^{(\mathbf{2})}\big)\widetilde{\mathbf{x}}[l]^T\Big\}, \end{aligned} \tag{9}$$

where $G^{(\mathbf{1})}(\cdot)$, $G^{(\mathbf{2})}(\cdot)$ respectively, is the cumulant generating function associated with the exponential family distribution $p^{(\mathbf{1})}(\cdot)$, $p^{(\mathbf{2})}(\cdot)$ respectively, as seen in equation (3).

Following the derivations in [12], the Newton-Raphson method is used for the iterative minimization of the loss function (9) and the resulting update equations are as follows. First at iteration t, for $l = 1, \ldots, m$,

$$\begin{aligned} \underline{\mathbf{a}}^{(t+1)}[l]^T &= \underline{\mathbf{a}}^{(t)}[l]^T - \alpha_{\underline{\mathbf{a}}}^{(t+1)} \\ &\cdot \Big\{\mathbf{V}^{(\mathbf{1})(t)} G^{(\mathbf{1})''}\big(\underline{\mathbf{a}}^{(t)}[k]\mathbf{V}^{(\mathbf{1})(t)} + \mathbf{b}^{(\mathbf{1})(t)}\big)\mathbf{V}^{(\mathbf{1})(t),T} \\ &+ \mathbf{V}^{(\mathbf{2})(t)} G^{(\mathbf{2})''}\big(\underline{\mathbf{a}}^{(t)}[k]\mathbf{V}^{(\mathbf{2})(t)} + \mathbf{b}^{(\mathbf{2})(t)}\big)\mathbf{V}^{(\mathbf{2})(t),T}\Big\}^{-1} \\ &\cdot \Big\{\mathbf{V}^{(\mathbf{1})(t)}\big(G^{(\mathbf{1})'}(\underline{\mathbf{a}}^{(t)}[k]\mathbf{V}^{(\mathbf{1})(t)} + \mathbf{b}^{(\mathbf{1})(t)}) - \widetilde{\mathbf{x}}[k]^T\big) \\ &+ \mathbf{V}^{(\mathbf{2})(t)}\big(G^{(\mathbf{2})'}(\underline{\mathbf{a}}^{(t)}[k]\mathbf{V}^{(\mathbf{2})(t)} + \mathbf{b}^{(\mathbf{2})(t)}) - \widetilde{\mathbf{x}}[k]^T\big)\Big\}. \end{aligned} \tag{10}$$

For the second step, the two sets of row vectors $\big\{\mathbf{v}_j^{(\mathbf{1})}\big\}_{j=1}^q$ and $\big\{\mathbf{v}_j^{(\mathbf{2})}\big\}_{j=1}^q$ are updated separately. The update equations can then be used for $\big\{\mathbf{v}_j^{(\mathbf{1})}\big\}_{j=1}^q$ and $\big\{\mathbf{v}_j^{(\mathbf{2})}\big\}_{j=1}^q$ by changing $\mathbf{v}_j$ to $\mathbf{v}_j^{(\mathbf{1})}$, respectively to $\mathbf{v}_j^{(\mathbf{2})}$, $\mathbf{b}$ to $\mathbf{b}^{(\mathbf{1})}$, respectively to $\mathbf{b}^{(\mathbf{2})}$, $G(\cdot), G'(\cdot)$, and

$G''(\cdot)$ to $G^{(\mathbf{1})}(\cdot), G^{(\mathbf{1})'}(\cdot)$, and $G^{(\mathbf{1})''}(\cdot)$, respectively to $G^{(\mathbf{2})}(\cdot), G^{(\mathbf{2})'}(\cdot)$, and $G^{(\mathbf{2})''}(\cdot)$. For $j = 1, \ldots, q$:

$$\begin{aligned}\mathbf{v}_j^{(t+1),T} &= \mathbf{v}_j^{(t),T} \\ &- \alpha_{\mathbf{v}}^{(t+1)} \left(\sum_{l=1}^{m} \underline{a}_j^{(t+1)}[l]^2 G''\big(\underline{\mathbf{a}}^{(t+1)}[l]\mathbf{V}^{(t)} + \mathbf{b}^{(t)}\big) \right)^{-1} \\ &\cdot \left(\sum_{k=1}^{n} \underline{a}_j^{(t+1)}[l] \big\{ G'\big(\underline{\mathbf{a}}^{(t+1)}[l]\mathbf{V}^{(t)} + \mathbf{b}^{(t)}\big) - \widetilde{\mathbf{x}}[k]^T \big\} \right).\end{aligned} \tag{11}$$

And finally for the last step, the update equations can then be used for $\mathbf{b}^{(\mathbf{1})}$ and $\mathbf{b}^{(\mathbf{2})}$ by changing $\mathbf{b}$ to $\mathbf{b}^{(\mathbf{1})}$, respectively to $\mathbf{b}^{(\mathbf{2})}$, $\mathbf{V}$ to $\mathbf{V}^{(\mathbf{1})}$, respectively to $\mathbf{V}^{(\mathbf{2})}$, $G(\cdot), G'(\cdot)$, and $G''(\cdot)$ to $G^{(\mathbf{1})}(\cdot), G^{(\mathbf{1})'}(\cdot)$, and $G^{(\mathbf{1})''}(\cdot)$, respectively to $G^{(\mathbf{2})}(\cdot), G^{(\mathbf{2})'}(\cdot)$, and $G^{(\mathbf{2})''}(\cdot)$.

$$\begin{aligned}\mathbf{b}^{(t+1),T} &= \mathbf{b}^{(t),T} \\ &= \mathbf{b}^{(t),T} - \alpha_{\mathbf{b}}^{(t+1)} \left(\sum_{l=1}^{m} G''\big(\underline{\mathbf{a}}^{(t+1)}[l]\mathbf{V}^{(t+1)} + \mathbf{b}\big) \right)^{-1} \\ &\cdot \left(\sum_{l=1}^{m} \big\{ G'\big(\underline{\mathbf{a}}^{(t+1)}[l]\mathbf{V}^{(t+1)} + \mathbf{b}\big) - \widetilde{\mathbf{x}}[k]^T \big\} \right).\end{aligned} \tag{12}$$

Table 1 summarizes the SP-PCA algorithm.

4.2. Exponential family PCA

The work proposed in [12] is based on the generalization of Principal Component Analysis to the exponential family technique presented in [7], often referred to as exponential Principal Component Analysis.

As stated earlier, instead of the fastidious estimation of the point-mass probabilities, exponential PCA considers the special uniform case where $\pi_l = 1/m$ for $l = 1, \ldots, m$, for which the number of support points equals the number of data samples ($m = n$). Hence, the point-mass probabilities do not need to be estimated and the EM algorithm is unnecessary. Then, to each vector $\mathbf{x}$ corresponds a vector $\underline{\mathbf{a}}$, i.e., a vector $\underline{\boldsymbol{\theta}}$, and they can share the same index $k = 1, \ldots, n$.

The update equations were presented in [12]. It is easily noticed that they are the same as equations (10), (11) and (12), the only difference being that the exponential PCA update equations use $\mathbf{x}$ instead of $\widetilde{\mathbf{x}}$, i.e., data points instead of mixture component centers. Each data point $\mathbf{x}$ is its own mixture component center.

As noted in [7], it is possible for the atoms obtained with exponential PCA to diverge since the optimum may be at infinity. To avoid such behavior, we introduce a penalty function that defines and places a set of constraints into the loss function via a penalty parameter in a way that penalizes any divergence to infinity. The penalty function

$$\begin{aligned}\psi(\underline{\boldsymbol{\theta}}) = \sum_{i=1}^{d} \Big\{ &\exp\big(-\beta_{min}(\underline{\theta}_i - \underline{\theta}_{min,i}) \big) \\ &+ \exp\big(\beta_{max}(\underline{\theta}_i - \underline{\theta}_{max,i}) \big) \Big\},\end{aligned}$$

is designed so that $\psi(\underline{\boldsymbol{\theta}})$ is close to zero for $\underline{\boldsymbol{\theta}}_{min} \leq \underline{\boldsymbol{\theta}} \leq \underline{\boldsymbol{\theta}}_{max}$ and reaches infinity otherwise. The penalty function modifications on the update equations (10), (11) and (12), i.e., an additional additive term within each of the large parentheses, are omitted here for sake of brevity.

Table 2 summarizes exponential PCA algorithm.

Algorithm: Semi-Parametric exp. family PCA [20]

Input: a set of observations $\{\mathbf{x}[k]\}_{k=1}^{n} \subseteq \mathbb{R}^d$, two exponential family distribution $p^{(\mathbf{1})}, p^{(\mathbf{2})}$ defined by their cumulant generating functions $G^{(\mathbf{1})}, G^{(\mathbf{2})}$, a number of atoms m, $q \ll d$ the dimension of the latent variable lower dimensional subspace.

Output: the NPML estimator that maximizes the complete log-likelihood function $L^{(c)}\big(\mathcal{Q}, \{\mathbf{z}_k\}_{k=1}^{n}\big)$: $\widehat{\mathcal{Q}} = \big\{\widehat{\underline{\boldsymbol{\theta}}}[l], \widehat{\pi}_l\big\}_{l=1}^{m}$ with $\widehat{\underline{\boldsymbol{\theta}}}[l] = \widehat{\underline{\mathbf{a}}}[l]\widehat{\mathbf{V}} + \widehat{\mathbf{b}}$ for all l, $\{\widehat{\underline{\mathbf{a}}}[l]\}_{l=1}^{m} \in \mathbb{R}^q$, $\widehat{\mathbf{V}} \in \mathbb{R}^{q \times d}$ and $\widehat{\mathbf{b}} \in \mathbb{R}^d$.

Method:

Initialize $\mathbf{V}$, $\mathbf{b}$ and $\big\{\underline{\mathbf{a}}[l], \pi_l\big\}_{l=1}^{m}$ with $\pi_l \geq 0$ for all l and $\sum_{l=1}^{m} \pi_l = 1$; $\underline{\boldsymbol{\theta}}[l] = \underline{\mathbf{a}}[l]\mathbf{V} + \mathbf{b} \in \boldsymbol{\Theta}$ for all l; $p\big(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[l]\big)$ as defined in (7) for all k and l;

repeat

{The Expectation Step}

for $k = 1$ to n **do**

for $l = 1$ to m **do**

$\widehat{z}_{kl} \longleftarrow p\big(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[l]\big)\pi_l / \sum_{r=1}^{m} p\big(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[r]\big)\pi_r$

end for

end for

{The Maximization Step}

for $l = 1$ to m **do**

$\pi_l \longleftarrow (1/n) \sum_{k=1}^{n} \widehat{z}_{kl}$

end for

{The Newton-Raphson iterative algorithm}

for $l = 1$ to m **do**

$\underline{\mathbf{a}}[l] \longleftarrow$ update equation (10)

end for

for $j = 1$ to q **do**

$\mathbf{v}_j \longleftarrow$ update equation (11)

end for

$\mathbf{b} \longleftarrow$ update equation (12)

until *convergence*;

return $\widehat{\mathcal{Q}} = \big\{\widehat{\underline{\boldsymbol{\theta}}}[l] = \widehat{\underline{\mathbf{a}}}[l]\widehat{\mathbf{V}} + \widehat{\mathbf{b}}, \widehat{\pi}_l\big\}_{l=1}^{m}$.

Table 1. Semi-Parametric PCA algorithm.

Algorithm: Exponential PCA [7]

Input: a set of observations $\{\mathbf{x}[k]\}_{k=1}^{n} \subseteq \mathbb{R}^d$, two exponential family distribution $p^{(\mathbf{1})}, p^{(\mathbf{2})}$ defined by their cumulant generating functions $G^{(\mathbf{1})}, G^{(\mathbf{2})}$, a number of atoms n, $q \ll d$ the dimension of the latent variable lower dimensional subspace.

Output: the NPML estimator that maximizes the log-likelihood function $L(\mathcal{Q})$: $\widehat{\mathcal{Q}} = \{\widehat{\underline{\boldsymbol{\theta}}}[k]\}_{k=1}^{n}$ with $\widehat{\underline{\boldsymbol{\theta}}}[k] = \widehat{\underline{\mathbf{a}}}[k]\widehat{\mathbf{V}} + \widehat{\mathbf{b}}$ for all k, $\{\widehat{\underline{\mathbf{a}}}[k]\}_{k=1}^{n} \in \mathbb{R}^q$, $\widehat{\mathbf{V}} \in \mathbb{R}^{q\times d}$ and $\widehat{\mathbf{b}} \in \mathbb{R}^d$.

Method:

Initialize $\mathbf{V}$, $\mathbf{b}$ and $\{\underline{\mathbf{a}}[k]\}_{k=1}^{n}$; $\underline{\boldsymbol{\theta}}[k] = \underline{\mathbf{a}}[k]\mathbf{V}+\mathbf{b} \in \boldsymbol{\Theta}$ for all k; $p(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[k])$ as defined in (7) for all k;
repeat
{The Newton-Raphson iterative algorithm}
for $k = 1$ to n **do**
$\underline{\mathbf{a}}[l] \longleftarrow$ penalty-modified update equation (10) with $\mathbf{x}$ instead of $\widetilde{\mathbf{x}}$
end for
for $j = 1$ to q **do**
$\mathbf{v}_j \longleftarrow$ penalty-modified update equation (11) with $\mathbf{x}$ instead of $\widetilde{\mathbf{x}}$
end for
$\mathbf{b} \longleftarrow$ penalty-modified update equation (12) with $\mathbf{x}$ instead of $\widetilde{\mathbf{x}}$
until *convergence*;
return $\widehat{\mathcal{Q}} = \{\widehat{\underline{\boldsymbol{\theta}}}[k] = \widehat{\underline{\mathbf{a}}}[k]\widehat{\mathbf{V}} + \widehat{\mathbf{b}}\}_{k=1}^{n}$.

Table 2. Exponential PCA algorithm.

4.3. Bregman soft clustering

The Bregman soft clustering approach presented in [4] utilizes an alternative interpretation of the EM algorithm for learning models involving mixtures of exponential family distributions. It is a simple soft clustering algorithm for all Bregman divergences, i.e., for all exponential family distributions. We choose here to present this technique without referring to the Bregman divergence as done in [4] but by using its corresponding exponential family probability distribution for the sake of comparison with SP-PCA and exponential PCA.

Given a data set of observations $\{\mathbf{x}[k]\}_{k=1}^{n}$, Bregman soft clustering aims at modeling the statistical structure of the data as a mixture of m densities of the same exponential family. The clusters correspond to the components of the mixture model and the soft membership of a data point in each cluster is proportional to the probability of the data point being generated by the corresponding density function. The Bregman soft clustering problem is based on a

Algorithm: Bregman Soft Clustering [4]

Input: a set of observations $\{\mathbf{x}[k]\}_{k=1}^{n} \subseteq \mathbb{R}^d$, two exponential family distribution $p^{(\mathbf{1})}, p^{(\mathbf{2})}$ defined by their cumulant generating functions $G^{(\mathbf{1})}, G^{(\mathbf{2})}$, a number of atoms m.

Output: the NPML estimator that maximizes the complete log-likelihood function $L^{(c)}(\mathcal{Q}, \{\mathbf{z}_k\}_{k=1}^{n})$: $\widehat{\mathcal{Q}} = \{\widehat{\underline{\boldsymbol{\theta}}}[l], \widehat{\pi}_l\}_{l=1}^{m}$.

Method:

Initialize $\{\underline{\boldsymbol{\theta}}[l], \pi_l\}_{l=1}^{m}$ with $\pi_l \geq 0$ for all l and $\sum_{l=1}^{m} \pi_l = 1$; $p(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[l])$ as defined in (7) for all k and l; $\underline{\boldsymbol{\theta}}[l] \in \boldsymbol{\Theta}$ for all l;
repeat
{The Expectation Step}
for $k = 1$ to n **do**
for $l = 1$ to m **do**
$\widehat{z}_{kl} \longleftarrow p(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[l])\pi_l / \sum_{r=1}^{m} p(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[r])\pi_r$
end for
end for
{The Maximization Step}
for $l = 1$ to m **do**
$\pi_l \longleftarrow (1/n)\sum_{k=1}^{n} \widehat{z}_{kl}$
$\underline{\boldsymbol{\theta}}[l] \longleftarrow$ solve for $\underline{\boldsymbol{\theta}}[l]$:
$G'(\underline{\boldsymbol{\theta}}[l]) = \sum_{k=1}^{n} \widehat{z}_{kl}\mathbf{x}[k] / \sum_{k=1}^{n} \widehat{z}_{kl}$
end for
until *convergence*;
return $\widehat{\mathcal{Q}} = \{\widehat{\underline{\boldsymbol{\theta}}}[l], \widehat{\pi}_l\}_{l=1}^{m}$.

Table 3. Bregman soft clustering algorithm.

maximum likelihood estimation of the cluster parameters $\{\underline{\boldsymbol{\theta}}[l], \pi_l\}_{l=1}^{m}$ satisfying the following mixture structure:

$$p(\mathbf{x}) = \sum_{l=1}^{m} p(\mathbf{x}|\underline{\boldsymbol{\theta}}[l])\pi_l,$$

where $p(\mathbf{x}|\cdot)$ is an exponential family distribution. The data likelihood function takes the following form:

$$p(\mathbf{X}) = \prod_{k=1}^{n}\sum_{l=1}^{m} p(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[l])\pi_l. \tag{13}$$

The data likelihood function in (13) is similar to the data likelihood function in (6) without the linear constraint $\underline{\boldsymbol{\theta}}[l] = \underline{\mathbf{a}}[l]\mathbf{V}+\mathbf{b}$ for $l = 1, \ldots, m$. Hence, the Bregman soft clustering problem is similar to the SP-PCA problem without the lower dimensional subspace constraint and a simple EM algorithm is used to estimate the cluster parameters. We consider again the mixed data type case. The E-step and the first part of the M-step yield the same results as for SP-PCA. In the second part of the M-step, the component parameters

$\underline{\boldsymbol{\theta}}[l], l = 1, \ldots, m$, are estimated in the following way:

$$\begin{aligned}\underline{\boldsymbol{\theta}}[l]^{(new)} &= \arg\max_{\underline{\boldsymbol{\theta}}[l]} \sum_{k=1}^{n}\sum_{r=1}^{m} \widehat{z}_{kr} \log p\big(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[r]\big) \\ &= \arg\max_{\underline{\boldsymbol{\theta}}[l]} \Big\{ \sum_{k=1}^{n}\sum_{r=1}^{m} \widehat{z}_{kr} \log p^{(\mathbf{1})}\big(\mathbf{x}^{(\mathbf{1})}[k]|\underline{\boldsymbol{\theta}}^{(\mathbf{1})}[r]\big) \\ &\qquad + \sum_{k=1}^{n}\sum_{r=1}^{m} \widehat{z}_{kr} \log p^{(\mathbf{2})}\big(\mathbf{x}^{(\mathbf{2})}[k]|\underline{\boldsymbol{\theta}}^{(\mathbf{2})}[r]\big) \Big\},\end{aligned}$$

with $\log p\big(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[r]\big) = \underline{\boldsymbol{\theta}}[r]\mathbf{x}[k]^T - G\big(\underline{\boldsymbol{\theta}}[r]\big)$. Using the convexity properties of $G(\cdot)$, it is easily shown that:

$$G'(\underline{\boldsymbol{\theta}}[l]^{(new)}) = \Big(\sum_{k=1}^{n} \widehat{z}_{kl}\mathbf{x}[k]\Big) \Big/ \Big(\sum_{k=1}^{n} \widehat{z}_{kl}\Big)$$

can be solved for $\underline{\boldsymbol{\theta}}[l]^{(new),(\mathbf{1})}$ and $\underline{\boldsymbol{\theta}}[l]^{(new),(\mathbf{2})}$ by changing $\mathbf{x}$ to $\mathbf{x}^{(\mathbf{1})}$, respectively to $\mathbf{x}^{(\mathbf{2})}$, $G'(\cdot)$ to $G^{(\mathbf{1})'}(\cdot)$, respectively to $G^{(\mathbf{2})'}(\cdot)$.

Table 3 summarizes the Bregman soft clustering algorithm.

5. Experimental results on synthetic data

We compare the relative performances of exponential PCA, SP-PCA and Bregman soft clustering in a mixed data set clustering problem with two data types and demonstrate how exponential PCA with the addition of a nonparametric estimation of the point-mass probabilities exceeds SP-PCA in performance.

We first consider a synthetic $d = 3$-dimensional data set with a lower dimensional subspace of dimension $q = 1$. The first data feature is Poisson distributed, the second and third features are Gaussian distributed. The data has $n = 500$ points and is composed of two mixture components with parameters $\boldsymbol{\theta}[1]$ and $\boldsymbol{\theta}[2]$ constrained to the lower dimensional subspace.

We first use exponential PCA. However, exponential PCA does not estimate point-mass probabilities. We use a nonparametric density estimation technique based on a kernel smoothing method to estimate the point-mass probabilities using the support points values $\underline{\mathbf{a}}[k]$, $k = 1, \ldots, n$, obtained by exponential PCA. Figure 3 shows that the nonparametric density estimation exhibits a definite two-component shape. The dotted lines represent the correct values $\underline{\mathbf{a}}[1]$ and $\underline{\mathbf{a}}[2]$. We can then estimate the values of $\underline{\mathbf{a}}[1]$ and $\underline{\mathbf{a}}[2]$ as well as their mixing distributions π_1 and π_2 using a simple `kmeans` algorithm, with the $\pi_1 + \pi_2 = 1$ assumption.

Figure 4 presents the histogram of the estimated point-mass probabilities obtained with SP-PCA, $m = 2$.

Table 4 shows detailed results for this synthetic data setting ("modified" means the extension to mixed data sets of the algorithm): the mixing distributions or point-mass probabilities π_1 and π_2, the latent variable or point of support values $\underline{\mathbf{a}}[1]$ and $\underline{\mathbf{a}}[2]$, the parameter values $\underline{\boldsymbol{\theta}}[1]$ and $\underline{\boldsymbol{\theta}}[2]$ as well as the sine of the angle between the estimated lower dimensional subspace and the correct subspace. Bregman soft clustering does not have the lower dimensional subspace constraint, and hence does not exhibit a sine or the latent variables values in Table 4. The estimation quality of the $\underline{\boldsymbol{\theta}}[1]$, $\underline{\boldsymbol{\theta}}[2]$ and π_1, π_2 values defines the clustering performance. For this simple Poisson-Gaussian mixed data setting, both exponential PCA and Bregman soft clustering seem to perform better than SP-PCA: the SP-PCA obtained parameter values for $\underline{\boldsymbol{\theta}}[2]$ are far from the original values, contrary to exponential PCA and Bregman soft clustering.

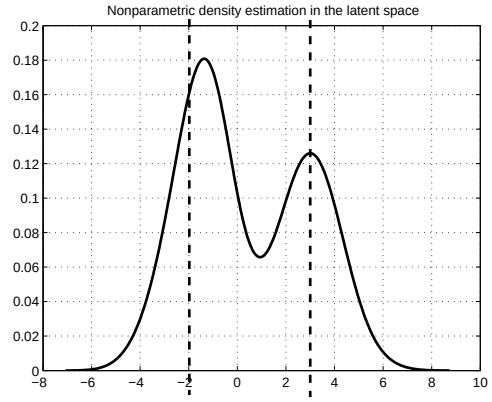

Figure 3. Nonparametric estimation of the point-mass probabilities obtained with exponential PCA (dotted: correct cluster centers).

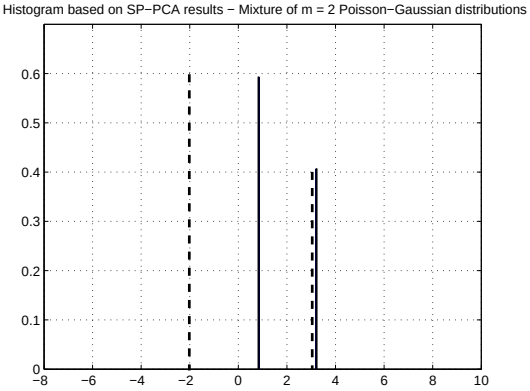

Figure 4. Histogram of the estimated point-mass probabilities obtained with SP-PCA (dotted: correct cluster values).

Results for a second experiment are shown in Table 5 for a Binomial-Gaussian mixed data set created in a similar fashion as the Poisson-Gaussian mixed data set (the parameter N is set to 10 for the Binomial component). Again, exponential PCA exceeds SP-PCA in clustering performance.

	$\pi_1; \pi_2$	$\underline{\mathbf{a}}[1]; \underline{\mathbf{a}}[2]$	$\underline{\boldsymbol{\theta}}[1]; \underline{\boldsymbol{\theta}}[2]$	sin
correct	0.4	3	[1.9404, 1.6148, 1.6210]	
model values	0.6	−2	[−1.2936, −1.0765, −1.0806]	
modified	0.4107	3.0009	[1.6235, 1.8648, 1.7007]	0.1368
exponential PCA	0.5893	−1.3725	[−0.7425, −0.8529, −0.7778]	
modified	0.3724	3.2170	[2.1732, 1.5715, 1.7768]	
SP-PCA	0.6276	0.8355	[0.5644, 0.4081, 0.4614]	0.058663
modified Bregman	0.4069		[1.9317, 1.7162, 1.5585]	
soft clustering	0.5931		[−1.1061, −1.0802, −1.0304]	

Table 4. Clustering results for a Poisson-Gaussian mixed data set.

	$\pi_1; \pi_2$	$\underline{\mathbf{a}}[1]; \underline{\mathbf{a}}[2]$	$\underline{\boldsymbol{\theta}}[1]; \underline{\boldsymbol{\theta}}[2]$	sin
correct	0.4	1	[0.8914, 0.1688, 0.4206]	
model values	0.6	−2	[−1.7828, −0.3375, −0.8412]	
modified	0.4475	0.8559	[0.7796, 0.1166, 0.3334]	0.049038
exponential PCA	0.5525	−1.9972	[−1.8193, −0.2721, −0.7779]	
modified	0.3978	−0.9548	[−0.9046, −0.0989, −0.2890]	
SP-PCA	0.6022	−3.1821	[−3.0148, −0.3296, −0.9633]	0.1455
modified Bregman	0.3973		[0.82252, 0.144, 0.41004]	
soft clustering	0.6027		[−1.8072, −0.3089, −0.9816]	

Table 5. Clustering results for a Binomial-Gaussian mixed data set.

6. Conclusion

We presented a mixed data-type hierarchical Bayes graphical model framework that adds clarity and perspective to our understanding of exponential PCA, SP-PCA and Bregman soft clustering. We demonstrated that these techniques are not separate unrelated algorithms but different manifestations of parameter choices taken within a common framework. Because of this insight, we were able to extend the algorithms to readily derive novel extensions that deal with the important mixed data type case. Our framework has the critical advantage of allowing one to go from high-dimensional mixed-type data components to low-dimensional common-type latent variables that are then used to perform clustering in a much simpler manner using well-known continuous-parameter clustering techniques.

References

[1] M. Aitkin, A general maximum likelihood analysis of overdispersion in generalized linear models, *Statistics and Comput.*, vol. 6, pp. 251-262, 1996.

[2] M. Aitkin, A maximum likelihood analysis of variable components in generalized linear models, *Biometrics*, vol. 55, pp. 117-128, 1999.

[3] K. S. Azoury and M. K. Warmuth, Relative loss bounds for online density estimation with the exponential family of distributions, *Mach. Learning*, vol. 43, pp. 211-246, 2001.

[4] A. Banerjee, S. Merugu, I. Dhillon and J. Ghosh, Clustering with Bregman divergences, *J. Mach. Learning Research*, vol. 6, pp. 1705-1749, 2005.

[5] D. Boehning, *Computer-Assited Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping, and Others*, Chapman and Hall/CRC, 2000.

[6] L. Cayton, Fast nearest neighbor retrieval for Bregman divergences, *25th Int. Conf. Mach. Learning*, 2008.

[7] M. Collins, S. Dasgupta and R. Shapire, A generalization of principal component analysis to the exponential family, *NIPS*, 2001.

[8] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood estimation from incomplete data via the EM algorithm, *J. Royal Statistical Soc., B*, vol. 39, pp. 1-38, 1977.

[9] J. Kiefer and J. Wolfowitz, Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *The Annals of Mathematical Statistics*, vol. 27, pp. 887-906, 1956.

[10] N. Laird, Nonparametric Maximum Likelihood Estimation of a Mixing Distribution, *J. American Statistical Soc.*, vol. 73, 1978.

[11] E. L. Lehmann and G. Castella, *Theory of Point Estimation*, Springer, 2nd edition, 1998.

[12] C. Levasseur, U. F. Mayer, B. Burdge and K. Kreutz-Delgado, Generalized statistical methods for unsupervised minority class detection in mixed data sets, *IAPR Workshop on Cognitive Inf.*, 2008.

[13] B. G. Lindsay, The geometry of mixture likelihoods: a general theory, *Annals of Stat.*, vol. 11, no. 1, pp. 86-94, 1983.

[14] B. G. Lindsay and M. L. Lesperance, A review of semiparametric mixture models, *J. Statistical Planning and Inference*, vol. 47, pp. 29-99, 1995.

[15] A. Mallet, A maximum likelihood estimation method for random coefficient regression models, *Biometrika*, vol. 73, no. 3, pp. 645-656, 1986.

[16] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley-Interscience, 2000.

[17] F. Nielsen, J.-. Boissonnat and R. Nock, On Visualizing Bregman Voronoi Diagrams, *Proc. of the 23rd ACM Conf. on Comput. Geometry*, 2007.

[18] R. S. Pilla and B. Lindsay, Alternative EM methods for nonparametric finite mixture models, *Biometrika*, vol. 88, no. 2, pp. 535-550, 2001.

[19] R. A. Redner and H. F. Walker, Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review*, vol. 26, no. 2, pp. 195-239, 1984.

[20] Sajama and A. Orlitsky, Semi-parametric exponential family PCA, *NIPS*, 2004.