Generalized Statistical Methods for Mixed Exponential Families, Part II: Applications

Cécile Levasseur, Uwe F. Mayer, and Kenneth Kreutz-Delgado

Abstract—This work considers the problem of both supervised and unsupervised classification for vector data of mixed types. An important subclass of graphical modeling techniques called Generalized Linear Statistics (GLS) is used to capture the underlying statistical structure of these complex data. The GLS methodology exploits the split between data space and natural parameter space for exponential family distributions, which are assumed to describe the data components, and constrains latent variables to a lower dimensional parameter subspace. It has the critical advantage of allowing one to transfer high-dimensional mixed-type data components to low-dimensional common-type latent variables, which are then, in turn, used to perform effective classification in a much simpler manner using well-known continuous-parameter classical linear techniques. We first demonstrate our ability to learn a GLS generative model in a controlled environment using synthetic data of mixed types. We then illustrate the benefits of making decisions in parameter space, with examples of categorical data (supervised and unsupervised) text categorization and mixed data-type classification and clustering, involving synthetic data and real data sets from the University of California, Irvine (UCI) machine learning repository.

Index Terms—Generalized Linear Statistics (GLS), exponential family distributions, latent variables, dimensionality reduction, text categorization.

1 Introduction

The complexity of data generally comes from the possible existence of a large number of components and from the fact that the components are often of different data types, i.e., some components might be continuous (with different underlying distributions) and some components might be discrete (categorical, count or Boolean). This is typically the case in drug discovery, health care, or fraud detection.

Graphical models, also referred to as Bayesian Networks when their graph is directed, are a powerful tool to encode and exploit the underlying statistical structure of complex data sets [1]. The Generalized Linear Statistics (GLS) framework represents a simple, yet useful, subclass of graphical model techniques and includes as special cases multivariate probabilistic approaches such as Principal Component Analysis (PCA), Generalized Linear Model (GLM) techniques and factor analysis [2], [3]. The GLS model is equivalent to a computationally tractable component-wise exponential families mixed data-type hierarchical Bayes graphical model with latent variables constrained to a low-dimensional parameter subspace. The use of exponential family distributions allows the data components to have different parametric forms and exploits the division between data space and parameter space specific to exponential families. In addition to giving a generative model that can be fitted to the data, it offers the advantage that problems can be attacked in a latent variable parameter subspace that is a continuous, Euclidean space, even when the data components are categorical or of varying exponential family types.

Although a variety of techniques exist for performing inference on graphical models, it is in general very difficult to learn the parameters which constitute the model, even if it is assumed that the graph structure is known [4], [5]. The main goal of this paper is to demonstrate our ability to learn a useful generative GLS graphical model that captures the statistical structure of vector data with components of differing data types, to then use this knowledge to gain insight into the problem domain, and perform effective classification. Text categorization and classification/clustering problems are presented as examples illustrating the benefits of both the GLS framework and making decisions in parameter space rather than in data space as with more classical approaches. Of course, Support Vector Machines (SVMs) also make decisions in a non-data space. However, although often promising the highest accuracy, the SVMs technique does not result in the construction of a generative model and will not generally provide any better understanding of the data. An advantage of learning a generative model of the data is that generating synthetic data for the purposes of developing and training classifiers with the same statistical structure as the original data becomes possible. This is particularly useful in cases where data

- *C. Levasseur and K. Kreutz-Delgado are with the Department of Electrical and Computer Engineering, University of California, San Diego (UCSD), La Jolla, CA, 92093.*
  E-mail: {clevasseur, kreutz}@ucsd.edu
- *Uwe F. Mayer is with the Department of Mathematics, University of Utah, Salt Lake City, UT, 84112.*
  E-mail: mayer@math.utah.edu

are very difficult or expensive to obtain, and when the original data are proprietary and cannot be directly used for publication purposes in open literature.

Building on a better understanding of previous work that first introduced Generalized Linear Statistics (GLS) and its applications [6], [2], [7], detailed classification results on both synthetic and real data sets are presented in this paper.

The paper is organized as follows. Section 2 presents a review of Generalized Linear Statistics, emphasizing how natural the GLS framework is for non-gaussian vector data of mixed vector-component types. In Section 3, synthetic data examples with data of mixed vector-component types, involving three different exponential families, illustrate the ability to correctly learn a GLS generative model. We introduce the angle between the estimated low-dimensional parameter subspace and the original low-dimensional parameter subspace used to generate the synthetic data as a natural measure of the quality of the estimated GLS model. Section 4 demonstrates the utility of the GLS approach, first with experiments on synthetic data, then with real-data experiments, where classification in parameter space often outperforms classification in data space. The synthetic data examples are a mixed vector components data-type unsupervised minority class detection problem and a mixed vector components data-type clustering problem involving the exponential family Principal Component Analysis technique of [8], the Semi-Parametric exponential family Principal Component Analysis technique of [9] and the Bregman soft clustering technique of [10]. These three techniques are special cases of GLS [3]. Finally, the GLS approach is applied to real data sets from the University of California, Irvine machine learning repository [11], namely the Twenty Newsgroups, the Reuters-21578 and the Abalone data sets, for the purposes of categorical-data text categorization and mixed vector components data-type classification.

2 Generalized Linear Statistics (GLS)

The Generalized Linear Statistics (GLS) framework is based on the hierarchical Bayes graphical model for hidden or latent variables shown in Figure 1 [3].

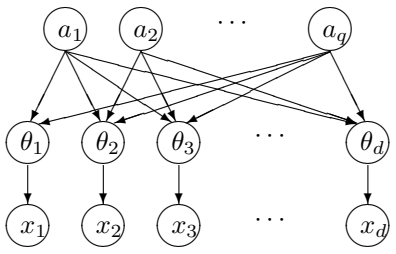

Fig. 1. Graphical model for the GLS framework.

The row vector $\mathbf{x} = [x_1, \ldots, x_d] \in \mathbb{R}^d$ consists of observed features of mixed data-type instances in a d-dimensional space. It is assumed that instances 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 parameter vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_d] \in \mathbb{R}^d$, the components of $\mathbf{x}$ are independent. 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. 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

$$p_i(x_i|\theta_i) = \exp\big(\theta_i x_i - G_i(\theta_i)\big),$$

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

$$G_i(\theta_i) = \log \int_{\mathcal{X}_i} \exp\big(\theta_i x_i\big) \nu_i(\mathrm{d}x_i),$$

with $\nu_i(\cdot)$ a σ-finite measure that generates the exponential family. It can be shown, using Fubini's theorem [12], that the cumulant generating function of the multivariate exponential family distribution $p(\mathbf{x}|\boldsymbol{\theta})$ in (1) is $G(\boldsymbol{\theta}) = \sum_{i=1}^{d} G_i(\theta_i)$.

It is further assumed that $\boldsymbol{\theta}$ can be written as

$$\boldsymbol{\theta} = \mathbf{a}\mathbf{V} + \mathbf{b} \tag{2}$$

with $\mathbf{V} \in \mathbb{R}^{q \times d}$ and $\mathbf{b} \in \mathbb{R}^d$ deterministic and unknown, and the hidden or latent variable $\mathbf{a} = [a_1, \ldots, a_q] \in \mathbb{R}^q$ unknown with $q < d$ (and ideally $q \ll d$). Note that the matrix $\mathbf{V}$ identifies a lower dimensional subspace in parameter space. The GLS framework both considers a Bayesian approach for which $\mathbf{a}$ is treated as a random vector and a classical approach where the vector $\mathbf{a}$ is deterministic. First, the vector $\mathbf{a}$ is assumed to be random. Then, in some way, the latent variable $\mathbf{a}$ explains part (or all) of the random behavior of the observed variables.

Since $\mathbf{a}$ (and hence $\boldsymbol{\theta}$) is treated as a random vector (Bayesian approach), the (non-conditional) probability density function $p(\mathbf{x})$ requires a generally intractable integration over the parameters,

$$p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \int \prod_{i=1}^{d} p_i(x_i|\theta_i)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{3}$$

with $\pi(\boldsymbol{\theta})$ the probability density function of $\boldsymbol{\theta}$. The maximum likelihood identification of this blind random effect model 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

$$p(\mathbf{X}) = \prod_{k=1}^{n} p\big(\mathbf{x}[k]\big) = \prod_{k=1}^{n} \int p\big(\mathbf{x}[k]|\boldsymbol{\theta}\big)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{4}$$

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

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

This difficulty can be avoided by Non-Parametric 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 [13], [14]. The NPML approach yields unknown point-mass support points $\underline{\mathbf{a}}[l]$, point-mass probability estimates π_l, and the linear predictor $\underline{\boldsymbol{\theta}}[l] = \underline{\mathbf{a}}[l]\mathbf{V} + \mathbf{b}$ for $l = 1, \dots, m$, with $m \leq n$. The single-sample likelihood (3) then becomes

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

and the data likelihood (4) is equal to

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

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 $\underline{\boldsymbol{\theta}}[l]$ in the lth mixture component. The combined problem of maximum likelihood estimation of the parameters $\mathbf{V}$, $\mathbf{b}$, the point-mass support points $\underline{\mathbf{a}}[l]$ and the point-mass probability estimates $\pi_l, l = 1, \dots, m$, can be attacked by using the Expectation-Maximization (EM) algorithm [15], [13], [14], [16], cf. in particular in the Semi-Parametric exponential family Principal Component Analysis technique [9].

However, a classical approach to the GLS estimation problem can also be considered and the vector $\mathbf{a}$ (and hence $\boldsymbol{\theta}$) is treated as a deterministic vector. Then, to each data point $\mathbf{x}[k]$, $k = 1, \dots, n$, corresponds a (generally different) parameter point, yielding a total of n points $\underline{\boldsymbol{\theta}}[k]$, $k = 1, \dots, n$, in parameter space (and hence n points $\underline{\mathbf{a}}[k]$, $k = 1, \dots, n$, in the parameter space low-dimensional subspace) as presented in the exponential family Principal Component Analysis technique [8]. The data likelihood is simply equal to

$$p(\mathbf{X}) = \prod_{k=1}^{n} p\big(\mathbf{x}[k]|\underline{\boldsymbol{\theta}}[k]\big) = \prod_{k=1}^{n} p\big(\mathbf{x}[k]|\underline{\mathbf{a}}[k]\mathbf{V} + \mathbf{b}\big). \quad (5)$$

Contrary to the Bayesian approach, no point-mass probabilities have to be estimated. For consistency of vocabulary throughout this paper, the points $\underline{\mathbf{a}}[k]$, $k = 1, \dots, n$, in the parameter space low-dimensional subspace are called latent variables for both Bayesian and classical approaches. Similarly, the parameter points $\underline{\boldsymbol{\theta}}[k]$, $k = 1, \dots, n$, are called atoms in both approaches. The classical approach can also be seen as an extreme case of the Bayesian approach for which the probability density function $\pi(\boldsymbol{\theta})$ is a delta function (one per data point) and the total number of atoms m equals the number of data points n, i.e., $m = n$. Note that while the $m < n$ parameter points of the Bayesian approach are shared by all the data points, the classical approach assigns one parameter point to each data point (hence $m = n$). This extreme case is the approach followed in Section 3 and in part of Section 4.

3 Fitting a GLS generative model

We now demonstrate our ability to learn a GLS generative model in a controlled environment using synthetic data of mixed types.

3.1 Angle between subspaces

Given a synthetic data set, the angle between the estimated low-dimensional paramater subspace $\mathcal{N}$ and the original low-dimensional parameter subspace $\mathcal{M}$ used to generate the synthetic data is proposed as a measure to assess the quality of the GLS model estimation. As stated in [17], defining the angle between subspaces in $\mathbb{R}^d$, $d \gg 1$, is not as straightforward as the visual geometry of $\mathbb{R}$ or $\mathbb{R}^3$ might suggest.

The *minimal angle* between nonzero subspaces $\mathcal{M}, \mathcal{N} \subseteq \mathbb{R}^d$ is defined to be the number $0 \leq \omega_{min} \leq \pi/2$ for which

$$\cos\omega_{min} = \max_{\substack{\mathbf{u}\in\mathcal{M},\mathbf{v}\in\mathcal{N} \\ \|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1}} \mathbf{v}^T\mathbf{u}. \quad (6)$$

If $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$ are the orthogonal projectors onto $\mathcal{M}$ and $\mathcal{N}$, respectively, then

$$\cos\omega_{min} = \|\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}}\|_2.$$

If $\mathcal{M}$ and $\mathcal{N}$ are complementary subspaces ($\mathcal{M} \oplus \mathcal{N} = \mathbb{R}^d$) and if $\mathbf{P}_{\mathcal{M}\mathcal{N}}$ is the oblique projector onto $\mathcal{M}$ along $\mathcal{N}$, then

$$\sin\omega_{min} = \frac{1}{\|\mathbf{P}_{\mathcal{M}\mathcal{N}}\|_2}.$$

$\mathcal{M}$ and $\mathcal{N}$ are complementary subspaces if and only if $\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}$ is invertible; in this case

$$\sin\omega_{min} = \frac{1}{\|(\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}})^{-1}\|_2}.$$

While the minimum angle works fine for complementary subspaces, it may not convey much information about the separation between non-complementary subspaces. For example, $\omega_{min} = 0$ whenever $\mathcal{M}$ and $\mathcal{N}$ have a nontrivial intersection, but there might be a nontrivial "gap" between $\mathcal{M}$ and $\mathcal{N}$ nevertheless.

The *maximal angle* between subspaces $\mathcal{M}, \mathcal{N} \subseteq \mathbb{R}^d$ is defined to be the number $0 \leq \omega_{max} \leq \pi/2$ for which

$$\sin\omega_{max} = \|\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}\|_2. \quad (7)$$

The maximum angle is chosen for the assessment of the lower dimensional subspace estimation performance.

Since $\mathbf{P}_{\mathcal{M}}$ is the orthogonal projector onto the lower dimensional subspace, and since the matrix $\mathbf{V} \in \mathbb{R}^{q \times d}$ defines this subspace, then

$$\mathbf{P}_{\mathcal{M}} = \mathbf{V}^T \mathbf{V}^+ = \mathbf{V}^T \left(\mathbf{V}\mathbf{V}^T\right)^{-1} \mathbf{V},$$

where the subscript $^+$ denotes a pseudo-inverse.

3.2 Mixed Binomial-Gaussian-Gamma data example

Figure 2 presents a mixture of two Binomial-Gaussian-Gamma mixed distributions in data space. The data are comprised of one Gamma attribute, one Gaussian attribute and one Binomial attribute with parameter $N = 10$ (the Binomial attribute counts the number of successes in N independent Bernoulli experiments).

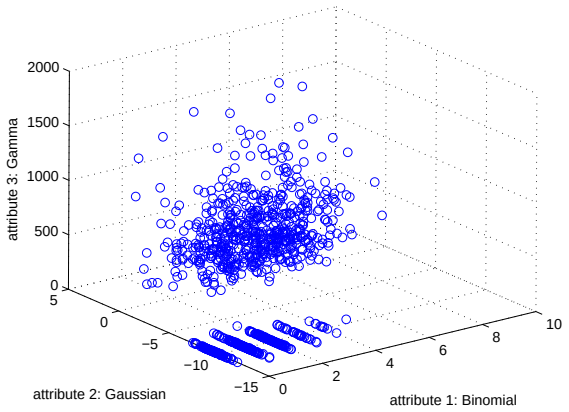

Fig. 2. Data space: a mixture of two Binomial-Gaussian-Gamma mixed distributions with parameters constrained on a 1-dimensional subspace.

The classical approach corresponding to exponential family Principal Component Analysis (exponential PCA) is assumed for GLS in this example. The sine of the angle between the original low-dimensional parameter subspace and the low-dimensional parameter subspace learned with GLS assuming a Binomial-Gaussian-Gamma mixed exponential family distribution is $\sin(\measuredangle_{GLS:BGG}) = 0.0326$. With a simple Gaussian GLS assumption, the sine becomes $\sin(\measuredangle_{GLS:G}) = 0.93375 > \sin(\measuredangle_{GLS:BGG})$. Since both a Binomial random variable and a Gamma random variable only take on positive values, we could only consider two more assumptions: a Binomial-Gaussian GLS assumption and a Gaussian-Gamma GLS assumption. The sine of the angle between the original subspace and the subspace learned with a Binomial-Gaussian GLS assumption is $\sin(\measuredangle_{GLS:BG}) = 0.93623 > \sin(\measuredangle_{GLS:BGG})$ and the sine of the angle between the original subspace and the subspace learned with a Gaussian-Gamma GLS assumption is $\sin(\measuredangle_{GLS:GG}) = 0.08310 > \sin(\measuredangle_{GLS:BGG})$. The results obtained with a Gaussian-Gamma GLS assumption are similar to the results obtained with the Binomial-Gaussian-Gamma GLS assumption. In this particular example, assuming a Gamma distribution for the last attribute seems to be essential for a good estimation performance.

We performed a similar experiment with a mixture of two Binomial-Gaussian-Gamma mixed distributions and the Binomial parameter N equal to 5. The results are similar to the ones obtained for the data with Binomial parameter N equal to 10 and are as follows: $\sin(\measuredangle_{GLS:G}) = 0.86571 > \sin(\measuredangle_{GLS:BG}) = 0.86359 > \sin(\measuredangle_{GLS:GG}) = 0.14198 > \sin(\measuredangle_{GLS:BGG}) = 0.098467$.

4 Data-driven decision making in parameter space

This section illustrates the benefits of making decisions in parameter space, with examples involving both synthetic and real data sets.

4.1 Unsupervised minority class detection on synthetic data

Minority class detection considers a binary class situation where a "minority class" is discriminated from a "majority class". It aims to differentiate rare key events belonging to the minority class from the remainder of the data belonging to the majority class.

The problem of unsupervised data-driven minority class (rare event) detection is one of relating property descriptors of a large unlabeled database of "objects" to measured properties of these objects, then using these empirically determined relationships to infer the properties of new objects. Here, the ultimate goal is to correctly characterize the new objects as either belonging to the minority class or not. This work assumes that minority class and majority class objects constitute two distinct, well-separated classes of objects in a latent variable subspace of the parameter space as described in Section 2. In the case of a rare occurrence of objects to be detected, it is believed that modeling the total unlabeled database allows one to discern the statistical structure of the majority class of objects. This experiment considers synthetic data sets of mixed types, involving both continuous and discrete data components.

Unsupervised methods for feature extraction, such as Principal Component Analysis (PCA), are commonly used to process data before using discriminative classifiers, such as Support Vector Machines (SVMs) or neural networks. However, methods such as Independent Component Analysis (ICA) and PCA assume the same form of the distribution for all components of the data. In contrast, the Generalized Linear Statistics framework developed in Section 2 allows each component to have its own parametric form. The minority class detection technique proposed here is based on the GLS framework, enabling the use of exponential family distributions to model the various mixed types of data measurements (continuous or discrete). A key aspect is that the parameters of the exponential family distributions are constrained to a lower dimensional latent variable subspace. The classical

approach to GLS corresponding to exponential PCA is considered here. The proposed minority class detection technique is performed in parameter space rather than in data space, as in more classical approaches, and exploits the low-dimensional information provided by the latent variables $\underline{\mathbf{a}}[k]$, $k = 1, \ldots, n$.

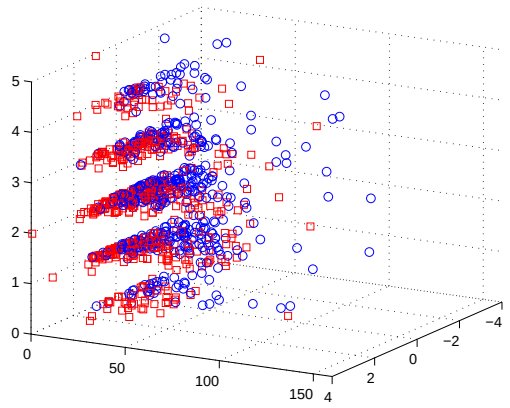

Fig. 3. Data space: data samples of a 3-dimensional mixed data set with Binomial, Exponential and Gaussian components (blue circles for one class and red squares for the other class).

Figure 3 shows an example of synthetic three-dimensional mixed data ($d = 3$), with each data sample comprised of a Binomial component with values between 0 and 5, an Exponential distribution component, and a unit-variance Gaussian component. The data are generated by two different classes, a minority one and a majority one, and for each class the parameters are assumed to be constrained to lie on a (different) one-dimensional subspace of the parameter space ($q = 1$). To assess the unsupervised minority class detection performance, we consider a situation where the minority class is a rare occurrence (1 percent of 10000 data samples), and the data are equally divided into a training set and a test set. The unsupervised minority class detection technique using the GLS information learned in parameter space works as follows: first, given the training set $\{\mathbf{x}[k]\}_{k=1}^{n}$, we learn the low-dimensional parameter subspace or direction of projection in parameter space, namely the matrix $\mathbf{V}$, by using the GLS modeling approach, and compute the training set mean-image on this low-dimensional parameter subspace, namely $(1/n)\sum_{k=1}^{n} \underline{\mathbf{a}}[k]$. The training set mean-image is then taken as an approximation to the training cluster mean of the majority class in the lower dimensional subspace. Then, to each test data sample corresponds a point in parameter space that was determined during the GLS model estimation. We compute its distance to the training set mean-image and compare the obtained distance to a given threshold λ to make a decision. The test point is declared to be part of the minority class if the distance is higher than λ, otherwise it is declared to be part of the majority class. This procedure is conducted for all of the test set samples, and the detection performance is assessed by plotting the ROC curve found from varying the value of λ. The ROC curve shows the probability of detection P_D versus the probability of false alarm P_{FA} as λ varies. The proposed technique is compared to classical PCA used in data space with a threshold test performed on new test data projected along the first principal axis, as well as to a supervised Bayes (minimum rate) detector for the sake of an optimal benchmark.

Data for which classical PCA will fail to provide accurate detection are easily created, using the knowledge that classical PCA defines the direction of projection as the direction of maximum variance in data space. The classical PCA approach will therefore give poor performance on data for which the direction of maximum variance is inappropriate for separating minority from majority class data. The Exponential distribution $p(x;\theta) = \beta \exp(-\beta x)$, with $\theta = -\beta$, is used as a component of the data. Because the inverse of the link function for this distribution is $f(x) = -1/x$, *the direction of maximum variance in data space is actually the direction of minimum variance in feature space,* and for this situation classical PCA is expected to perform poorly, and indeed it does.

Figure 4 shows a comparison between the supervised Bayes detector, the minority class detector based on GLS information and performed in parameter space, and the minority class detector based on classical PCA information and performed in data space. This illuminating example shows that there are domains for which classical PCA performs far from optimal.

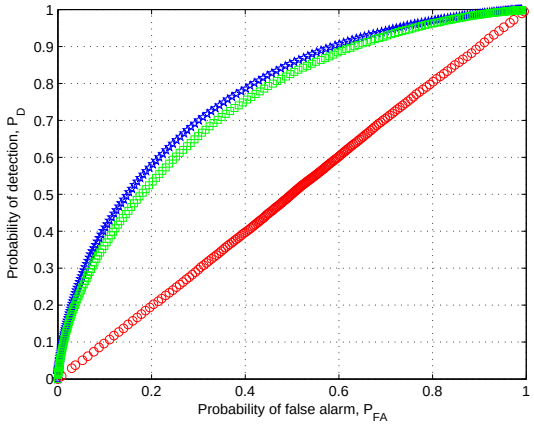

Fig. 4. Comparison of supervised Bayes optimal (top blue with pentagrams), proposed GLS technique (middle green with squares) and classical PCA (bottom red with circles) ROC curves.

4.2 Clustering results on synthetic data

This application compares the relative performances of exponential PCA, Semi-Parametric exponential family Principal Component Analysis (SP-PCA) and Bregman soft clustering in a mixed data set clustering problem

TABLE 1
Clustering results for a Poisson-Gaussian mixed data set.

	π_1; π_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]	0.058663
SP-PCA	0.6276	0.8355	[0.5644, 0.4081, 0.4614]	
modified Bregman	0.4069		[1.9317, 1.7162, 1.5585]	
soft clustering	0.5931		[-1.1061, -1.0802, -1.0304]	

TABLE 2
Clustering results for a Binomial-Gaussian mixed data set.

	π_1; π_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]	0.1455
SP-PCA	0.6022	-3.1821	[-3.0148, -0.3296, -0.9633]	
modified Bregman	0.3973		[0.82252, 0.144, 0.41004]	
soft clustering	0.6027		[-1.8072, -0.3089, -0.9816]	

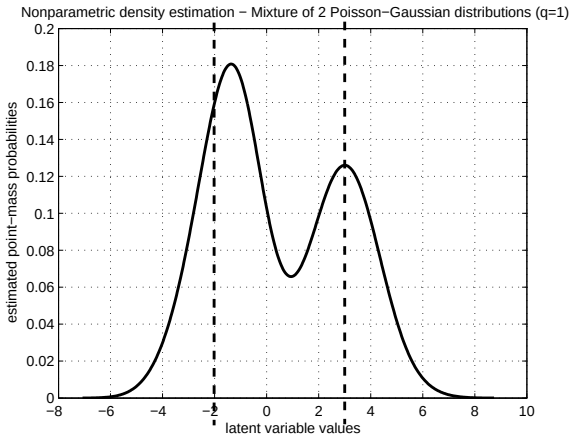

Fig. 5. Non-parametric estimation of the point-mass probabilities obtained with exponential PCA (dotted: correct cluster centers).

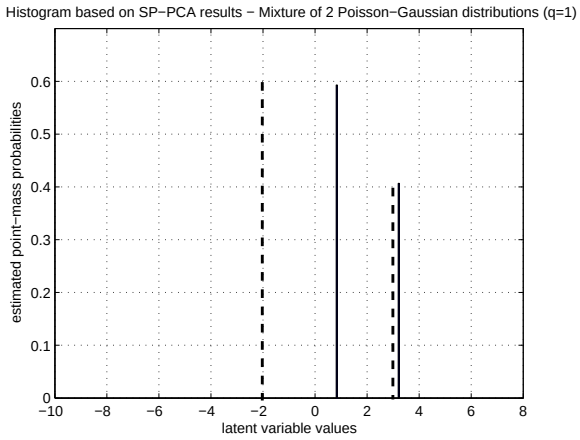

Fig. 6. Histogram of the estimated point-mass probabilities obtained with SP-PCA (dotted: correct cluster values).

with two data types and demonstrates how exponential PCA with the addition of a non-parametric estimation of the point-mass probabilities can exceed 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 $\underline{\boldsymbol{\theta}}[1]$ and $\underline{\boldsymbol{\theta}}[2]$ constrained to the lower dimensional subspace.

We first use exponential PCA. However, exponential PCA does not estimate point-mass probabilities. We use a non-parametric 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 5 shows that the non-parametric 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 K-means algorithm, with the $\pi_1 + \pi_2 = 1$ assumption.

Figure 6 presents the histogram of the estimated point-mass probabilities obtained with SP-PCA, $m = 2$.

Table 1 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 1 or Table 2. 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.

Results for a second experiment are shown in Table 2 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.

4.3 Text Categorization

The Twenty Newsgroups and the Reuters-21578 data sets [11] are used for most of the published experimental literature in text categorization, one example of information retrieval tasks. Text categorization is the activity of labeling natural language texts with thematic categories from a predefined set [18].

It has been acknowledged by the text categorization community that words seem to work well as features of a document for many classification tasks. In addition, it is usually assumed that the ordering of the words in a document does not matter. Hence, a document can simply be represented as a bag of words, i.e., as a vector for which each distinct word is a feature [19]. There are two ways to characterize the value of each feature that are commonly used in the literature: Boolean and *tf*×*idf* weighting schemes. In Boolean weighting, the weight of a word is considered to be 1 if the word appears in the document and 0 otherwise. We choose to characterize the value of each feature by using the *tf*×*idf* (term frequency × inverse document frequency) scheme as recently more commonly used for document representation [20], [18]. This scheme argues that terms (or words) appearing in documents should be weighted proportional to the term frequency and inversely proportional to the document frequency. The *tf*×*idf* weight is a statistical measure used to evaluate how important a word (or term) is to a corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. This weighting scheme and the combination of document length normalization have been shown to yield generally better retrieval results [20], [21], [22], [18]; interestingly, in practice, the Boolean approach does not always perform worse than the tf×idf approach [23]. The *term frequency tf* is the number of times a specific word occurs in a specific document. The *document frequency df* is the number of documents in which the specific word occurs at least once. The *inverse document frequency idf* is calculated from the document frequency, as follows:

$$idf = \log\left(\frac{\text{total \# of documents}}{df}\right),$$

yielding the *tf*×*idf* weight w_i for each feature i:

$$w_i = tf_i \cdot idf_i = tf_i \cdot \log\left(\frac{\text{total \# of documents}}{df_i}\right).$$

Hence, the tf×idf weighting with length normalization is for all i:

$$w_i = \frac{tf_i \cdot \log\left(\frac{\text{total \# of documents}}{df_i}\right)}{\sqrt{\sum_{j=1}^{|T|}\left[tf_j \cdot \log\left(\frac{\text{total \# of documents}}{df_j}\right)\right]^2}},$$

where $|T|$ is the length of the document, i.e., the number of distinct words in the document (after stopword removal and stemming is performed as explained below). Length normalization ensures that each document vector is of unit length, removing the advantage that long documents have over short documents with respect to information retrieval [18]. However, if a document is long, but has quite often a term that represents key information for a specific text categorization task, normalization would reduce the importance of the term as compared to a short document, where the term appears equally often in absolute term. Hence we decide to discard the length normalization step.

We choose to bin the weights and work with integer valued weights (5 bins are selected), i.e., categorical features.

4.3.1 Twenty Newsgroups data set

The Twenty Newsgroups data set consists of Usenet articles collected from twenty different newsgroups. Each newsgroup contains 1000 articles. We consider the three following newsgroups: sci.med, comp.sys.mac.hardware and comp.sys.ibm.pc.hardware. We decide on a text categorization problem with two distinct classes, the first class consisting of the newsgroup sci.med and the second class consisting of the two other newsgroups.

Following the text document representation preprocessing steps described in Figure 7, we first choose to discard all header fields such as Cc, Bcc, Message-ID, as well as the Subject field (this step is called parsing). Case-folding, which stands for converting all the characters in a document into the same case, is performed by converting all the characters into lower-case. We use a stop list, i.e., a list of words that will not be taken into account. Indeed, there are words such as pronouns, prepositions and conjunctions which are encountered very frequently but carry no useful information about the content of the document. We used the common stop list ftp://ftp.cs.cornell.edu/pub/smart/english.stop, made of 571 stop-words and commonly used in the literature. Then, some simple stemming is performed, such as

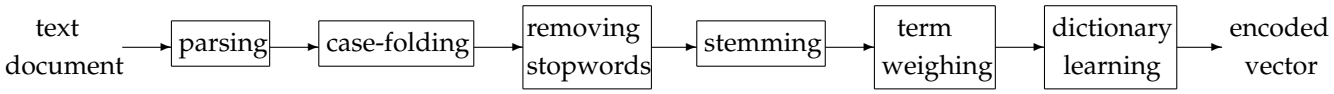

Fig. 7. Preprocessing and document representation for text categorization.

removing the third person and plural "s". In addition to removing very frequent words with the stop list, we remove rare words, i.e., words appearing less than 10 times in the corpus. It yields a drastic reduction in the number of features. The *tf* $\times$ *idf* weighting scheme is then used and we choose to bin the weights and work with integer valued weights (5 bins are selected), i.e., categorical features. At this point, each document is a vector in a 4383-dimensional space, i.e., 4383 distinct words were identified to represent the newsgroups documents.

Modified dictionary learning. We now construct a dictionary, and hence reduce the dimensionality of the feature space. There are various methods commonly applied for dimensionality reduction in document categorization [19]. We choose a conditional mutual information based approach to select a dictionary of $d = 150$ words. We modify the binary feature selection with conditional mutual information algorithm proposed in [24] to fit a categorical feature. The feature selection algorithm proposed in [24] is based on the Conditional Mutual Information Maximization (CMIM) criterion and selects features that maximize both the information about the class and the independence between features. The modification from binary to categorical is simple: following the definition of entropy and mutual information shown in [24], the summations are changed from summing over two values to summing over the total number of bins values.

We use this data set leaving out a randomly selected 40% of the instances of each class to use as a test set. The training set then consists of 1764 instances and the test set 1236. The dictionary is learned using the training set only. Table 3 presents the twenty first words of the dictionary. For the text categorization examples, the extreme case of GLS corresponding to exponential PCA is solely considered. Figure 8 represents the training set documents in the low-dimensional subspace of the parameter space learned with classical PCA for a dimension q of the subspace equal to 2. Similarly, Figure 9 represents the training set documents in the low-dimensional subspace of the parameter space learned with the GLS approach using a Binomial distribution for a dimension q of the subspace equal to 2.

Classification effectiveness is often measured in terms of *precision* and *recall* in the text categorization community [18]. Precision with respect to a class $\mathcal{C}_i$ (π_i) is defined as the probability that, if a random document is classified under $\mathcal{C}_i$, this decision is correct. Recall with respect to a class $\mathcal{C}_i$ (ρ_i) is defined as the probability that,

TABLE 3
Twenty Newsgroups data set: the twenty first words of the dictionary learned to differentiate the newsgroup sci.med from the newsgroups comp.sys.mac.hardware and comp.sys.ibm.pc.hardware.

doctor
card
mac
drive
disease
medical
treatment
food
patient
effect medicine
drug
skepticism
pc
body
health
blood
study
hardware
infection

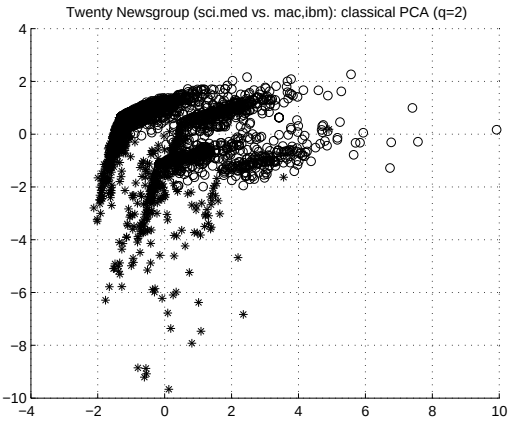

Fig. 8. Twenty Newsgroup data set: training documents in the lower-dimensional subspace of the parameter space learned with classical PCA, $q = 2$ (sci.med: *, others: $\circ$).

if a random document ought to be classified under $\mathcal{C}_i$, this decision is taken. These probabilities are estimated in terms of the contingency table for $\mathcal{C}_i$ on a given test set as follows:

$$\widehat{\pi}_i = \frac{TP_i}{TP_i + FP_i} \quad \text{and} \quad \widehat{\rho}_i = \frac{TP_i}{TP_i + FN_i},$$

where TP_i, FP_i and FN_i refer to the sets of *true positives with respect to* $\mathcal{C}_i$ (documents correctly deemed to belong

TABLE 4
Averaging precision, recall and F_1 measure across different classes.

	microaveraging (μ)	macroaveraging (M)
precision (π)	$\widehat{\pi}^{\mu} = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} TP_i}{\sum_{i=1}^{\lvert\mathcal{C}\rvert}(TP_i+FP_i)}$	$\widehat{\pi}^{M} = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \pi_i}{\lvert\mathcal{C}\rvert} = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \frac{TP_i}{TP_i+FP_i}}{\lvert\mathcal{C}\rvert}$
recall (ρ)	$\widehat{\rho}^{\mu} = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} TP_i}{\sum_{i=1}^{\lvert\mathcal{C}\rvert}(TP_i+FN_i)}$	$\widehat{\rho}^{M} = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \pi_i}{\lvert\mathcal{C}\rvert} = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \frac{TP_i}{TP_i+FN_i}}{\lvert\mathcal{C}\rvert}$
F_1	$F_1^{\mu}=\frac{2\cdot\sum_{i=1}^{\lvert\mathcal{C}\rvert} TP_i}{2\cdot\sum_{i=1}^{\lvert\mathcal{C}\rvert} TP_i+\sum_{i=1}^{\lvert\mathcal{C}\rvert} FP_i+\sum_{i=1}^{\lvert\mathcal{C}\rvert} FN_i}$	$F_1^{M}=\frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} F_{1,i}}{\lvert\mathcal{C}\rvert} = \frac{\sum_{i=1}^{\lvert\mathcal{C}\rvert} \frac{2\cdot TP_i}{2\cdot TP_i+FP_i+FN_i}}{\lvert\mathcal{C}\rvert}$

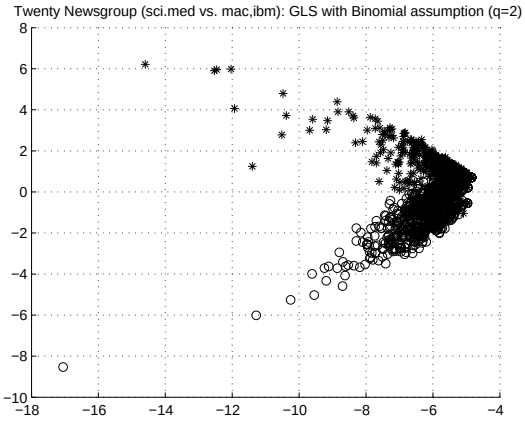

Fig. 9. Twenty Newsgroup data set: training documents in the lower-dimensional subspace of the parameter space learned with GLS (Binomial), $q = 2$ (sci.med: *, others: ∘).

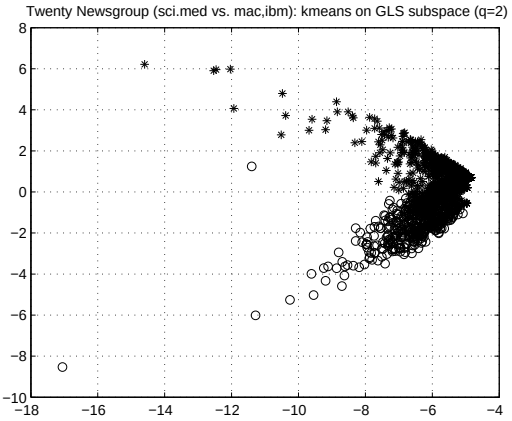

Fig. 10. Twenty Newsgroup data set: K-means results for a two-class classification of the training documents in the GLS subspace (Binomial, $q = 2$).

to class $\mathcal{C}_i$), *false positives with respect to* $\mathcal{C}_i$ (documents incorrectly deemed to belong to class $\mathcal{C}_i$), and *false negatives with respect to* $\mathcal{C}_i$ (documents incorrectly deemed not to belong to class $\mathcal{C}_i$). Then, the F_1 measure combines precision and recall, attributing equal importance to π and ρ:

$$F_1 = \frac{2 \cdot \pi\rho}{\pi + \rho}.$$

When effectiveness is computed for several classes, the results for individual classes can be averaged in two ways: *microaveraging*, where π and ρ are obtained by summing over all individual classes (the subscript "μ" indicates microaveraging), and *macroaveraging*, where π and ρ are first evaluated "locally" for each class and then "globally" by averaging over the results of the different classes (the subscript "M" indicates macroaveraging) [18].

Supervised text categorization. Table 5 compares classification performances on (a) the q-dimensional latent variable subspace learned with GLS using a Binomial distribution assumption and (b) the q-dimensional classical Principal Component Analysis (PCA) subspace learned in data space in terms of precision, recall and F_1 measure, for several values of q. The classifier is a simple linear discriminant. The classification performances are often very similar, at times at the advantage of GLS ($q = 4$ and $q = 10$). These results were obtained by using the MatlabArsenal toolbox, a package for classification algorithms [25]. The operating point is defined as the one maximizing the F_1 measure.

Unsupervised text categorization. The K-means algorithm is used to cluster the training documents into two distinct classes. Figure 10 represents the two clusters learned with K-means on the training set documents in the low-dimensional parameter subspace ($q = 2$). Comparing Figure 10 with Figure 9 shows how effective the K-means clustering algorithm is in this parameter subspace. Based on this clustering information, a linear discriminant is learned on the training documents and used to classify the test documents. Figure 11 presents the corresponding ROC curve for this unsupervised approach performed on both the GLS parameter subspace and the classical PCA data subspace ($q = 2$). The performance is best when the unsupervised approach is used on the GLS subspace rather than on the classical PCA subspace.

4.3.2 Reuters-21578 data set

The Reuters-21578 text categorization test collection Distribution 1.0 is considered as the standard benchmark for

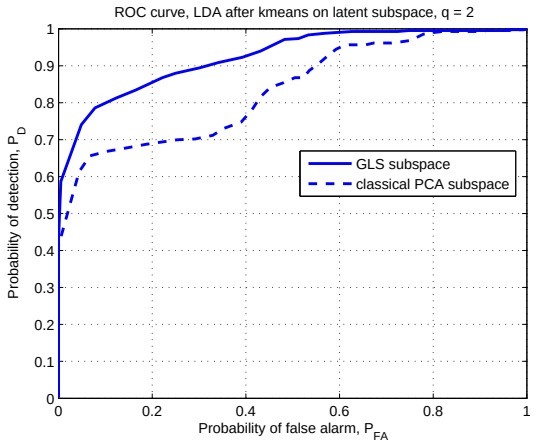

Fig. 11. Twenty Newsgroups data set: ROC curve for the unsupervised approach learned on the GLS subspace (solid line) and the classical PCA subspace (dashed line) ($q = 2$).

TABLE 6
The ten topics with the highest number of training documents in the Reuters-21578 data set with the number of their documents in the training and test sets.

topics	training set	test set
earn	2877	1087
acq	1650	719
money-fx	538	179
grain	433	149
crude	389	189
trade	369	118
interest	347	131
wheat	212	71
ship	197	89
corn	181	56

automatic document organization systems and consists of documents that appeared on the Reuters newswire in 1987. This corpus contains 21578 documents assigned to 135 different economic subject categories called *topics*. The topics are not disjoint. For the training test division of the data, the "Modified Apte" (ModApte) split is used, dividing the corpus into a training set of 9603 documents and a test set of 3299 documents. We reduce the size of the training test sets by only considering the ten topics that have the highest number of training documents as as suggested in [18], [26]. These topics are given in Table 6 and yield a training set of 6490 documents and a test set of 2545 documents. They cover almost all of the data, hence, researchers are able to restrict their work to them and still capture the essence of the data set. The data are preprocessed as for the previous data set: parsing, case-folding, elimination of stopwords, stemming by using Porter's stemming algorithm commonly used for word stemming in English [27], elimination of words that appear less than 20 times in the corpus, $tf \times idf$ weighting. At this point, 3613 distinct words were identified to represent the Reuters-21578 documents. Then, we choose to bin the weights and work with integer valued weights (5 bins are selected), i.e., categorical features. A dictionary of $d = 50$ words is learned using the following approach. The dictionary is learned on the training set only and built independently for each of the ten classes. Feature selection was incremental purely out of computational-runtime reasons. First we do a backward selection to 300 features with linear regression. From these 300 features, we use a logistic regression with a number of iterations reduced down to 5 for convergence, and do a backward selection down to 100 features. Finally, we do a standard full-convergence logistic regression from those 100 features down to 50 features.

Table 7 compares linear discriminant classification performances on the q-dimensional latent variable space learned with classical PCA and GLS with a Binomial distribution (1087 positive test instances) for the first and the tenth categories of the Reuters-21578 data set. Table 8 compares classification performances micro- and macroaveraged over the top ten categories of the Reuters-21578 data set using a linear discriminant classifier on (a) the latent q-dimensional variable subspace learned with GLS using a Binomial distribution assumption and (b) the classical PCA q-dimensional subspace learned in data space. The averaging is performed as explained in Table 4. Microaveraging and macroaveraging methods give quite different results: the linear discriminant classifier performs better based on the GLS information than on classical PCA information when the macroaveraging method is used, while microaveraging emphasizes how similar the two results are. It is known that the ability of a classifier to behave well on categories with few positive training instances will be highlighted by macroaveraging compared to microaveraging [18]. The linear discriminant classifier based on GLS information performs very well for the categories with fewer positive training instances yielding a better macroaveraged performance than the microaveraged one, cf. Table 6.

4.4 Abalone data set

The task is to predict the age of an abalone based on physical measurements. The Abalone data set consists of 4177 instances with 8 attributes. The problem can be seen as a classification problem aiming to distinguish three classes (number of rings $= 1$ to 8, number of rings $= 9$ to 10, number of rings $= 11$ and higher). The number of rings approximately corresponds to the age of the abalone. We use this data set leaving out a randomly selected 40% of the instances to use as a test set (2506 training points and 1671 test points).

Figure 12 represents the histograms of the complete data set for each attribute. Attribute 4 has two outliers not shown in its histogram below; with the exception of these two data points, all histograms show the full data.

Attribute 1 (sex, defined as infant, male or female) is the only noncontinuous attribute. We choose to model

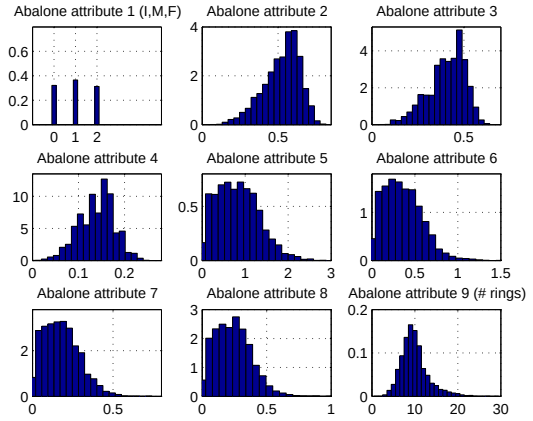

Fig. 12. Histograms performed on each attribute of the Abalone data set.

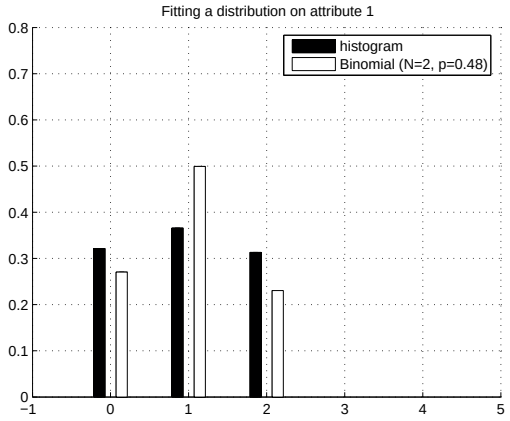

Fig. 13. Abalone data set: distribution fitting on attribute 1.

this attribute with a Binomial ($N = 2$) distribution, hence choosing a Gaussian-Binomial mixed-data GLS assumption. The extreme GLS case corresponding to exponential PCA is used here. Figure 13 presents a distribution fitting option for attribute 1. Table 9 compares micro- and macroaveraged classification performances using a linear discriminant classifier on (a) the latent q-dimensional variable subspace learned with GLS using a mixed Gaussian-Binomial distribution assumption and (b) the classical PCA q-dimensional subspace learned in data space. Performances are best when classification is performed on the GLS parameter subspace.

Then, we try to fit a distribution to the attributes 5, 6, 7 and 8. Possible distributions are the Weibull distribution, the Gamma distribution, the Beta distribution, the Chi-square distribution and the Non-central Chi-square distribution. The Beta distribution has a special constraint that the data should be greater than 0 and smaller than 1; only attributes 7 and 8 verify this constraint. Figure 14 presents distribution fitting options for attribute 5. The Gamma distribution is chosen as a good candidate to fit attributes 5, 6, 7 and 8.

Hence, we choose a Binomial-Gaussian-Gamma

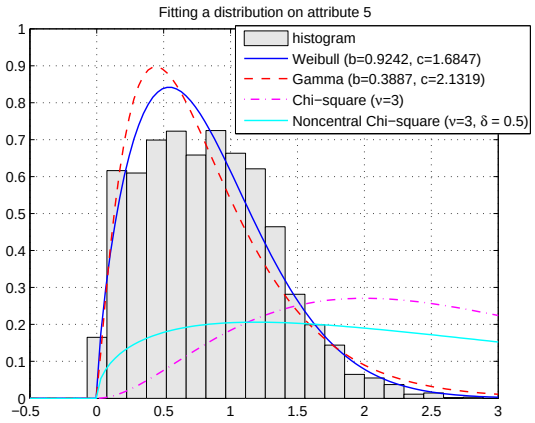

Fig. 14. Abalone data set: distribution fitting on attribute 5.

mixed-data assumption for the data set, with attribute 1 modeled as a Binomial variable, attributes 2, 3 and 4 modeled as Gaussian variables and attributes 5, 6, 7 and 8 modeled as Gamma variables. Table 10 compares the micro- and macroaveraged classification performances using a linear discriminant classifier on (a) the latent q-dimensional variable subspace learned with GLS using a mixed Binomial-Gaussian-Gamma distribution assumption and (b) the classical PCA q-dimensional subspace learned in data space. There are no statistically significant differences between the performances obtained with a mixed Binomial-Gaussian GLS assumption and performances obtained with a mixed Binomial-Gaussian-Gamma GLS assumption. As a conclusion, using a Gamma modeling assumption for the last four attributes, while it did not hurt, was not useful to the linear discriminant classifier.

5 Conclusion

As with Bayesian Networks in general, the specialized Generalized Linear Statistics (GLS) framework offers important insight into the underlying statistical structure of complex data of mixed types, both creating a generative model of vector data and enabling effective classification. We first demonstrated our ability to learn a GLS generative model using synthetic data examples with data components of varying exponential family types. The angle between the estimated low-dimensional parameter subspace and the original low-dimensional parameter subspace used to generate the synthetic data was proposed to assess the quality of the estimated GLS model. The benefits of making decisions in parameter space rather than in data space as with more classical approaches have been clearly illustrated with examples of Binomial data supervised and unsupervised text categorization and several mixed-data supervised and unsupervised classification examples, involving up to three different exponential family distributions to describe the data components. For the text categorization situation, the conditional mutual information maximization based

feature selection algorithm was modified to fit categorical data. It has been shown previously that GLS contains as special cases exponential family Principal Component Analysis, Semi-Parametric exponential family Principal Component Analysis and Bregman soft clustering. We compared the relative performances of the three algorithms in a clustering setting for mixed data sets, showing that GLS in general achieves comparable, and at times superior performance to established methods.

REFERENCES

[1] M. I. Jordan and T. J. Sejnowski, *Graphical Models: Foundations of Neural Computation*. Computational Neuroscience, The MIT Press, 1st edition, 2001.

[2] C. Levasseur, B. Burdge, K. Kreutz-Delgado, and U. F. Mayer, "A unifying viewpoint of some clustering techniques using Bregman divergences and extensions to mixed data sets," *Proceedings of the First IEEE Int'l Workshop on Data Mining and Artificial Intelligence (DMAI)*, pp. 56–63, 2008.

[3] C. Levasseur, K. Kreutz-Delgado, and U. F. Mayer, "Generalized statistical methods for mixed exponential families, part I: theoretical foundations," *submitted to IEEE Transactions of Pattern Analysis and Machine Intelligence (PAMI), available at http://dsp.ucsd.edu/∼cecile/*, 2009.

[4] G. F. Cooper, "Probabilistic inference using belief networks is NP-hard," *Technical report KSL 87-27, Stanford Knowledge Systems Laboratory*, 1987.

[5] D. M. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of Bayesian networks is NP-hard," *Journal of Machine Learning Research*, vol. 5, pp. 1287–1330, 2004.

[6] C. Levasseur, U. F. Mayer, B. Burdge, and K. Kreutz-Delgado, "Generalized statistical methods for unsupervised minority class detection in mixed data sets," *Proceedings of the First IAPR Workshop on Cognitive Information Processing*, pp. 126–131, 2008.

[7] C. Levasseur, U. F. Mayer, and K. Kreutz-Delgado, "Classifying non-gaussian and mixed data sets in their natural parameter space," *Proceedings of the Nineteenth IEEE Int'l Workshop on Machine Learning for Signal Processing (MLSP)*.

[8] M. Collins, S. Dasgupta, and R. E. Schapire, "A generalization of principal components analysis to the exponential family," in *Advances in Neural Information Processing Systems*, 2001, vol. 14.

[9] Sajama and A. Orlitsky, "Semi-parametric exponential family PCA," *Advances in Neural Information Processing Systems*, vol. 17, 2004.

[10] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[11] A. Asuncion and D. Newman, "UCI machine learning repository, http://www.ics.uci.edu/∼mlearn/MLRepository.html," University of California, Irvine, School of Information and Computer Sciences, 2007.

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

[13] N. Laird, "Nonparametric maximum likelihood estimation of a mixing distribution," *Journal of the American Statistical Society*, vol. 73, 1978.

[14] B. G. Lindsay, "The geometry of mixture likelihoods: a general theory," *Annals of Statistics*, vol. 11, no. 1, pp. 86–94, 1983.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *Journal of Royal Statistical Society, B*, vol. 39, pp. 1–38, 1977.

[16] D. Boehning, *Computer-Assited Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping, and Others*. Monographs on Statistics and Applied Probability 81, Chapman and Hall/CRC, New York, 2000.

[17] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra Book and Solutions Manual*. Society for Industrial and Applied Mathematics (SIAM); Package edition, 2001.

[18] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[19] A. Özgür, *Supervised and Unsupervised Machine Learning Techniques for Text Document Categorization*. M.S. thesis, Computer Engineering, Bogazici University, Istanbul, Turkey, 2004.

[20] G. Salton, *Introduction to Modern Information Retrieval*. Computer Science Series, McGraw-Hill, 1983.

[21] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988.

[22] D. Hiemstra, "A probabilistic justification for using the tf×idf term weighting in information retrieval," *International Journal on Digital Libraries*, 2000.

[23] A. Özgür and T. Güngör, "Classification of skewed and homogenous document corpora with class-based and corpus-based keywords," *Lecture Notes in Artificial Intelligence*, vol. 4314, pp. 91–101, 2006.

[24] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5, pp. 1531–1555, 2004.

[25] R. Yan, "MATLABArsenal, a MATLAB package for classification algorithms," Informedia, School of Computer Science, Carnegie Mellon University, 2006.

[26] F. Debole and F. Sebastiani, "An analysis of the relative hardness of Reuters-21578 subsets," *Journal of the American Society for Information Science and Technology*, vol. 56, pp. 584–596, 2005.

[27] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

TABLE 5
Twenty Newsgroups data set: linear discriminant classification performances on the q-dimensional latent variable space learned with classical PCA and GLS with a Binomial distribution (distinguishing newsgroup sci.med from newsgroups comp.sys.mac.hardware and comp.sys.ibm.pc.hardware).

	PCA - Precision	PCA - Recall	PCA - F_1	GLS - Precision	GLS - Recall	GLS - F_1
$q = 1$	0.5045	0.8149	**0.6232**	0.3677	0.6603	0.4744
$q = 2$	0.7843	0.9351	**0.8531**	0.7844	0.8918	0.8346
$q = 3$	0.9388	0.8846	**0.9109**	0.8641	0.8558	0.8599
$q = 4$	0.9389	0.8870	0.9122	0.8830	0.9615	**0.9206**
$q = 5$	0.9038	0.9712	**0.9363**	0.8931	0.9639	0.9272
$q = 6$	0.9038	0.9712	**0.9363**	0.8914	0.9663	0.9273
$q = 8$	0.9040	0.9736	**0.9375**	0.8813	0.9639	0.9208
$q = 10$	0.8904	0.9760	0.9312	0.9691	0.9038	**0.9353**

TABLE 7
Reuters-21578 data set: linear discriminant classification performances on the q-dimensional latent variable space learned with classical PCA and GLS with a Binomial distribution for two of the top ten topics.

(a) Performances for topic named "earn"

earn	PCA - Precision	PCA - Recall	PCA - F_1	GLS - Precision	GLS - Recall	GLS - F_1
$q = 1$	0.8893	0.8868	**0.8881**	0.4483	0.3707	0.4058
$q = 2$	0.9682	0.8951	**0.9302**	0.9834	0.8712	0.9239
$q = 3$	0.9644	0.8960	0.9289	0.9907	0.8804	**0.9323**
$q = 4$	0.9709	0.8905	0.9290	0.9866	0.8822	**0.9315**
$q = 5$	0.9630	0.9108	**0.9362**	0.9768	0.8914	0.9322

(b) Performances for topic named "corn"

corn	PCA - Precision	PCA - Recall	PCA - F_1	GLS - Precision	GLS - Recall	GLS - F_1
$q = 1$	0.0220	0.2500	0.0404	0.1542	0.6250	**0.2473**
$q = 2$	0.0900	0.5000	0.1526	0.1757	0.6964	**0.2806**
$q = 3$	0.2394	0.8036	**0.3689**	0.2009	0.7679	0.3185
$q = 4$	0.3659	0.8036	0.5028	0.4272	0.7857	**0.5535**
$q = 5$	0.3600	0.8036	0.4972	0.4175	0.7679	**0.5409**

TABLE 8
Reuters-21578 data set: linear discriminant classification performances (micro- and macroaveraged) on the q-dimensional latent variable space learned with classical PCA and GLS with a Binomial distribution.

(a) Microaveraged performances

	PCA - Precision$^\mu$	PCA - Recall$^\mu$	PCA - F_1^μ	GLS - Precision$^\mu$	GLS - Recall$^\mu$	GLS - F_1^μ
$q = 1$	0.2408	0.7306	0.3622	0.2845	0.5653	**0.3785**
$q = 2$	0.3704	0.8303	0.5123	0.4087	0.7665	**0.5331**
$q = 3$	0.4553	0.8296	**0.5880**	0.4239	0.8099	0.5565
$q = 4$	0.4709	0.8260	**0.5998**	0.4743	0.8128	0.5990
$q = 5$	0.6178	0.8275	**0.7075**	0.6233	0.7895	0.6966
$q = 6$	0.6265	0.8364	0.7164	0.6484	0.8056	**0.7185**

(b) Macroaveraged performances

	PCA - Precision^M	PCA - Recall^M	PCA - F_1^M	GLS - Precision^M	GLS - Recall^M	GLS - F_1^M
$q = 1$	0.2200	0.6403	0.2905	0.3040	0.6274	**0.3751**
$q = 2$	0.3763	0.7717	0.4475	0.4006	0.7174	**0.4757**
$q = 3$	0.4342	0.7842	0.5184	0.4662	0.7552	**0.5267**
$q = 4$	0.4594	0.7820	0.5423	0.5138	0.7611	**0.5804**
$q = 5$	0.4988	0.7673	0.5870	0.5307	0.7386	**0.6007**
$q = 6$	0.5306	0.7809	**0.6150**	0.5471	0.7373	0.6134

TABLE 9
Abalone data set: linear discriminant classification performances (micro- and macroaveraged) on the q-dimensional latent variable space learned with classical PCA and GLS with a Binomial-Gaussian distribution.

(a) Microaveraged performances

	PCA Precision$^{\mu}$	PCA Recall$^{\mu}$	PCA F_1^{μ}	GLS Precision$^{\mu}$	GLS Recall$^{\mu}$	GLS F_1^{μ}
$q = 1$	0.5036	0.7120	0.5899	0.5043	0.7409	**0.6001**
$q = 2$	0.5085	0.7337	0.6007	0.5178	0.7385	**0.6088**
$q = 3$	0.5523	0.7331	**0.6300**	0.5103	0.6954	0.5887

(b) Macroaveraged performances

	PCA Precision^{M}	PCA Recall^{M}	PCA F_1^{M}	GLS Precision^{M}	GLS Recall^{M}	GLS F_1^{M}
$q = 1$	0.5204	0.7126	0.5952	0.5208	0.7415	**0.6058**
$q = 2$	0.5242	0.7335	0.6062	0.5337	0.7380	**0.6141**
$q = 3$	0.5667	0.7336	**0.6355**	0.5220	0.6958	0.5925

TABLE 10
Abalone data set: linear discriminant classification performances (micro- and macroaveraged) on the q-dimensional latent variable space learned with classical PCA and GLS with a Binomial-Gaussian-Gamma distribution.

(a) Microaveraged performances

	PCA Precision$^{\mu}$	PCA Recall$^{\mu}$	PCA F_1^{μ}	GLS Precision$^{\mu}$	GLS Recall$^{\mu}$	GLS F_1^{μ}
$q = 1$	0.5036	0.7120	0.5899	0.5037	0.7394	**0.5992**
$q = 2$	0.5085	0.7337	0.6007	0.5191	0.7382	**0.6096**
$q = 3$	0.5523	0.7331	**0.6300**	0.5119	0.6960	0.5899

(b) Macroaveraged performances

	PCA Precision^{M}	PCA Recall^{M}	PCA F_1^{M}	GLS Precision^{M}	GLS Recall^{M}	GLS F_1^{M}
$q = 1$	0.5204	0.7126	0.5952	0.5199	0.7400	**0.6046**
$q = 2$	0.5242	0.7335	0.6062	0.5348	0.7375	**0.6146**
$q = 3$	0.5667	0.7336	**0.6355**	0.5243	0.6967	0.5939