GENERALIZED STATISTICAL METHODS FOR UNSUPERVISED MINORITY CLASS DETECTION IN MIXED DATA SETS

Cécile Levasseur

Jacobs School of Engineering
University of California, San Diego
La Jolla, CA, USA
clevasseur@ucsd.edu

Uwe F. Mayer

Fair Isaac Corporation
San Diego, CA, USA
uwemayer@fairisaac.com

Brandon Burdge, Ken Kreutz-Delgado

Jacobs School of Engineering
University of California, San Diego
La Jolla, CA, USA
{bburdge,kreutz}@ucsd.edu

ABSTRACT

Minority class detection is the problem of detecting the occurrence of rare key events differing from the majority of a data set. This paper considers the problem of unsupervised minority class detection for multidimensional data that are highly nongaussian, mixed (continuous and/or discrete), noisy, and nonlinearly related, such as occurs, for example, in fraud detection in typical financial data. A statistical modeling approach is proposed which is a subclass of graphical model techniques. It exploits the properties of exponential family distributions and generalizes techniques from classical linear statistics into a framework referred to as Generalized Linear Statistics (GLS). The methodology exploits the split between the data space and the parameter space for exponential family distributions and solves a nonlinear problem by using classical linear statistical tools applied to data that has been mapped into the parameter space. A fraud detection technique utilizing low-dimensional information learned by using an Iteratively Reweighted Least Squares (IRLS) based approach to GLS is proposed in the parameter space for data of mixed type. ROC curves for an initial simulation on synthetic data are presented, which gives predictions for results on actual financial data sets.

Index Terms— Minority class detection, generalized linear models, exponential family distributions, graphical models, dimensionality reduction.

1. INTRODUCTION

Minority class detection considers a binary class situation where a "minority class" is discriminated from a "majority class". It aims at differentiating rare key events belonging to the minority class from the remainder of the data belonging to the majority class. Many important risk assessment system applications depend on the ability to accurately detect the occurrence of rare key events given a large data set of observations. This problem arises in drug discovery ("Do the molecular descriptors associated with known drugs suggest that a new, candidate drug will have low toxicity and high effectiveness?"); and health care ("Do the descriptors associated with a medical doctor professional behavior suggest that he/she is an outlier in the category he/she was assigned to?"). The work proposed here is specifically concerned with the problem of minority class detection; for example, in credit card fraud detection ("Given the data for a large set of credit card users does the usage pattern of this particular card indicate that it might have been stolen?"). In many domains, no or little *a priori* knowledge exists regarding the true sources of any causal relationships that may occur between variables of interest. In these situations, meaningful information regarding the key events must be extracted from the data itself.

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, and then using these empirically determined relationships to detect 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 space (the "parameter space") to be defined below. In the case of a rare occurrence of objects to be detected, as is typically the case in credit card fraud detection, there is the belief that modeling the total unlabeled database allows one to discern the statistical structure of the majority class of objects. This work considers measured object properties that are nongaussian, mixed (comprised of continuous and discrete data), very noisy, and highly nonlinearly related for which the resulting minority class detection problem is very difficult. The difficulties are further compounded because the descriptor space is of high dimension.

Many of the classical tools for unsupervised feature extraction and analysis such as Principle Components Analysis (PCA), Independent Component Analysis (ICA) and classical Factor Analysis (FA) are all tied together by sharing a common general directed graph structure, and differ only in

certain assumptions about the type (discrete or continuous) of latent variables and the form of node probability distributions [1]. One of the key assumptions made by these approaches is that the components of the observed node all share the same form of conditional probability distribution. In contrast, the proposed approach allows for the components to have differing parametric forms; using exponential family distributions the components can model a large variety of mixed data types. A key aspect of our method, referred to as Generalized Linear Statistics (GLS), is that the parameter of the exponential family distributions is constrained to a lower dimensional latent variable subspace. This models the belief that the intrinsic dimensionality of the data is smaller than the observed dimensionality of the data space.

The unsupervised minority class detection technique proposed here is performed in the parameter space rather than in the data space as done in more classical approaches. As an example, a synthetic data set is investigated, where a single latent variable subspace is learned by using the GLS based statistical modeling on an unlabeled training set. Given a new data point, that point is projected to its image in the parameter space on the learned subspace and minority class detection is performed by comparing its distance from the training set mean-image to a threshold. The presented example shows that there are domains for which the classical linear techniques, such as PCA, used in the data space perform far from optimal compared to the new proposed parameter space techniques. ROC curves are generated to assess the performance of the proposed minority class detection method.

2. GENERALIZED LINEAR STATISTICS (GLS)

The proposed statistical modeling approach is a subclass of graphical model techniques. It is a generalization and amalgamation of techniques from classical linear statistics, Logistic Regression, Principal Component Analysis (PCA), latent variable analysis and Generalized Linear Models (GLMs) as well as our previous work [2] into a unified framework we refer to (analogously to GLMs theory) as *Generalized Linear Statistics* (GLS). This is actually a *nonlinear* methodology which exploits the split that occurs for exponential family distributions between the *data space* (also known as the *expected value space*) and the *(natural) parameter space* as soon as one leaves the domain of purely Gaussian random variables. The point is that although the problem is now nonlinear, it can be attacked by using classical linear (and other standard) statistical tools applied to data which has been *mapped into the parameter space*, which still has a natural, flat Euclidean space structure. For example, in the parameter space one can perform regression (resulting in the technique of logistic regression and other GLMs methods [3]), PCA (resulting in a variety of "generalized PCA" methods [4]), or clustering [5].

This framework is used to develop algorithms capable of minority class detection in domains involving highly heterogeneous data types and unlabeled data sets. Specifically, this work considers data records which have both continuous (e.g., exponential and Gaussian) and discrete (e.g., count and binary) components. It focusses on the development of *unsupervised* minority class detection algorithms which can be trained using unlabeled training data sets. The unsupervised case is very difficult and takes one out of the domain of the standard supervised approaches, such as neural networks.

To motivate theoretical developments, a general graphical model for hidden variables is considered, cf. Fig. 1. The row vector $\mathbf{x} = [x_1, \ldots, x_d] \in \mathbb{R}^d$ consists of observed features in a d-dimensional space. categorical or count data) and continuous values. Following the probabilistic Generalized Latent Variable (GLV) formalism described in [6], 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

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

with the hidden or latent variable $\mathbf{a} = [a_1, \ldots, a_q] \in \mathbb{R}^q$ random with $q < d$ (and ideally $q \ll d$), $\mathbf{V} \in \mathbb{R}^{q \times d}$ and $\mathbf{b} \in \mathbb{R}^d$ deterministic. Conditioning on the random vector $\boldsymbol{\theta}$ is equivalent to conditioning on the low-dimensional random vector $\mathbf{a}$. In a probabilistic sense, all of the information which is mutually contained in the data vector $\mathbf{x}$ must be contained in the latent variable $\mathbf{a}$. As noted in [6], equations (1) and (2) generalize the classical factor analysis model to the case when the marginal densities $p_i(x_i|\theta_i)$ are nongaussian. Indeed, 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. It is further assumed that the marginal densities are each one-

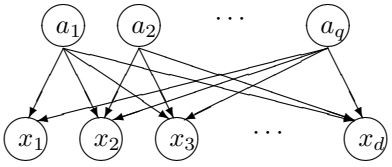

Fig. 1. Graphical model for GLS

parameter exponential family densities, allowing the use of a rich and powerful theory of such densities to be utilized, and it is commonly the case that 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

$$p(x_i|\theta_i) = \exp\big(\theta_i x_i - G(\theta_i)\big), \tag{3}$$

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

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

with $\nu(\cdot)$ a σ-finite measure that generates the exponential family. It can be shown, using Fubini's theorem [10], that $G(\boldsymbol{\theta}) = \sum_{i=1}^{d} G(\theta_i)$.

In both the GLV theory described in [6] and the random- and Mixed-effects Generalized Linear Models (MGLMs) literature [3], $\mathbf{V}$ and $\mathbf{b}$ are deterministic while $\mathbf{a}$ (and hence $\boldsymbol{\theta}$) is treated as a random vector. The difference is that in GLV, *all* of the quantities $\mathbf{V}$, $\mathbf{b}$, and $\mathbf{a}$ are unknown, and hence need to be identified, whereas in MGLMs, $\mathbf{V}$ is a known matrix of regressor variables and only the deterministic vector $\mathbf{b}$ and the unknown realizations of the *random effect* vector $\mathbf{a}$ must be estimated. In both GLV and MGLMs, it is assumed that in the $\boldsymbol{\theta}$-parameter space the linear relationship (2) holds and (at least conceptually) that the tools of linear and statistical inverse theory are applicable or insightful. The MGLMs theory is a generalization of the classical theory of linear regression, while the GLV theory is a generalization of the classical theory of factor analysis and PCA. In both cases the generalization is based on a move from the data/description space containing the measurement vector $\mathbf{x}$ to the parameter space containing $\boldsymbol{\theta}$ (via a generally nonlinear transformation known as a link function [3]), and it is in the latter space that the linear relationship (2) is assumed to hold. Because both the Generalized Linear Models (GLMs) and the Generalized Latent Variable (GLV) methodologies exploit the linear structure (2), they can be viewed as special cases of a Generalized Linear Statistics (GLS) approach to data analysis.

With the observational conditional distributions described, attention turns to the marginal distribution of parent nodes $a_1, \ldots, a_q$. For motivation, note that the (nonconditional) density $p(\mathbf{x})$ requires a generally intractable integration over the parameters,

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

where $\pi(\boldsymbol{\theta})$ is the probability density function of $\boldsymbol{\theta} = \mathbf{aV}+\mathbf{b}$. Given the observation matrix $\mathbf{X} = \big[\mathbf{x}[1]^T, \ldots, \mathbf{x}[n]^T\big]^T$ in $\mathbb{R}^{n\times d}$ composed of n iid statistical samples, each assumed to be stochastically equivalent to the random row vector $\mathbf{x}$,

$$p(\mathbf{X}) = \prod_{k=1}^{n} p(\mathbf{x}[k]) = \prod_{k=1}^{n} \int \prod_{i=1}^{d} p_i(x_i[k]|\theta_i)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \quad (5)$$

with $\boldsymbol{\theta} = \mathbf{aV} + \mathbf{b}$. For specified exponential family densities $p_i(\cdot|\cdot)$, $i = 1, \ldots, d$, maximum likelihood identification of the model (4) corresponds to identifying $\pi(\boldsymbol{\theta})$, which, under the condition $\boldsymbol{\theta} = \mathbf{aV} + \mathbf{b}$, corresponds to identifying the matrix $\mathbf{V}$, the vector $\mathbf{b}$, and a density function, $\mu(\mathbf{a})$, on $\mathbf{a}$ via a maximization of the data likelihood function $p(\mathbf{X})$ with respect to $\mathbf{V}$, $\mathbf{b}$, and $\mu(\mathbf{a})$. This is generally a quite difficult problem [3] and is usually attacked using approximation methods which correspond to replacing the integral in (4)/(5) by a sum [7]:

$$p(\mathbf{x}) = \sum_{j=1}^{m} p(\mathbf{x}|\underline{\boldsymbol{\theta}}_j)\pi_j = \sum_{j=1}^{m}\prod_{i=1}^{d} p_i(x_i|\underline{\theta}_{j,i})\pi_j \quad (6)$$

$$p(\mathbf{X}) = \prod_{k=1}^{n}\sum_{j=1}^{m}\prod_{i=1}^{d} p_i(x_i[k]|\underline{\theta}_{j,i})\pi_j \quad (7)$$

over a finite number of support points $\underline{\boldsymbol{\theta}}_j$, (equivalently, $\underline{\mathbf{a}}_j$), $j = 1, \ldots, m$, with point-mass probabilities

$$\pi_j \triangleq \pi(\boldsymbol{\theta} = \underline{\boldsymbol{\theta}}_j) = \pi(\mathbf{a} = \underline{\mathbf{a}}_j).$$

As clearly described in [7], this approximation is justified either as a Gaussian quadrature approximation to the integral in (5) [3] or by appealing to the fact that the *NonParametric Maximum Likelihood* (NPML) estimate of the mixture density $\pi(\boldsymbol{\theta})$ yields a solution which takes a finite number (m) of points of support [5, 8].

With $\boldsymbol{\theta} = \mathbf{aV} + \mathbf{b}$, with $\mathbf{V}$, $\mathbf{b}$ fixed and $\mathbf{a}$ random, the likelihood function (6) is equal to

$$p(\mathbf{x}) = \sum_{j=1}^{m} p(\mathbf{x}|\underline{\boldsymbol{\theta}}_j)\pi_j = \sum_{j=1}^{m} p(\mathbf{x}|\underline{\mathbf{a}}_j\mathbf{V} + \mathbf{b})\pi_j, \quad (8)$$

and the data likelihood function (7) is equal to

$$p(\mathbf{X}) = \prod_{k=1}^{n}\sum_{j=1}^{m} p(\mathbf{x}[k]|\underline{\mathbf{a}}_j\mathbf{V} + \mathbf{b})\pi_j. \quad (9)$$

The combined problem of maximum likelihood estimation (MLE) of the parameters $\mathbf{V}$, $\mathbf{b}$, the support points $\underline{\mathbf{a}}_j$ and the point-mass probability estimates π_j, $j = 1, \ldots, m$, (as approximations to the unknown, and possibly continuous density $\mu(\mathbf{a})$) is known as the NPML estimation problem [8]. It can be attacked by using the Expectation-Maximization (EM) algorithm [6] as done in [5], or, as done in [4], by simply considering the special case of uniform point-mass probabilities, i.e., $\pi_j = 1/m$ for j = 1, . . . , m, for which the number of support points equals the number of data samples, i.e., $m = n$.

The goal here is to fit a probability model of the form (9) using exponential family densities to labeled (when available) or unlabeled data to develop algorithms for deciding if a new measurement belongs to the minority class or not. For example, if an adequate fit of a parameterized probability distribution has only been found to the single, labeled majority class, the question is whether the new data point fits well with this distribution or whether it should be flagged as a potential member of the minority class. Alternatively, if class-conditional distributions can be fitted to minority and majority class labeled data, a Bayes-optimal likelihood ratio test can

be computed [9]. Class-conditional density-based tests can be equivalently posed as discriminant functions which are functions of sufficient statistics of the densities (when they exist) and which, in turn, define decision surfaces in feature space. Of course, the most difficult situation arises when the training samples are unlabeled. However, even in this case, sometimes the single-class model can still be effective for minority class detection. For example, if the ratio of minority class data to majority class data is very small, then the unlabeled data points are approximately distributed like the majority class data, and the simpler single-class model might be effectively assumed and utilized. This condition can be satisfied in practice; fraudulent credit card transactions are typically approximately one tenth of one percent of all transactions.

3. MINORITY CLASS DETECTION

The minority class detection technique proposed here is performed in the parameter space rather than in the data space as done in more classical approaches, and exploits the low dimensional information provided by the latent variables $\underline{\mathbf{a}}_j, j = 1, \ldots, m$. The proposed technique considers the special case of uniform point-mass probabilities, $\pi_j = 1/m$ for $j = 1, \ldots, m$, for which the number of support points equals the number of data samples, i.e., $m = n$. 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 $\underline{\mathbf{a}}$ and they can share the same index $k = 1, \ldots, n$.

For sake of simplicity, the $\mathbf{b}$-term in (2) is absorbed in the standard manner into the matrix $\mathbf{V}$ using the homogenous coordinates. The simultaneous estimation of the parameters $\mathbf{V} \in \mathbb{R}^{q \times d}$ and $\mathbf{A} = \left[\underline{\mathbf{a}}_1^T, \ldots, \underline{\mathbf{a}}_n^T\right]^T \in \mathbb{R}^{n \times q}$ is performed by minimizing the negative log-likelihood function. Using (9), the loss function is expressed as

$$\begin{aligned} L(\mathbf{V}, \mathbf{A}) &= -\log p(\mathbf{X}) = -\sum_{k=1}^{n} \log p(\mathbf{x}[k] | \underline{\mathbf{a}}_k \mathbf{V}) \\ &= \sum_{k=1}^{n} \left\{ G(\underline{\mathbf{a}}_k \mathbf{V}) - \underline{\mathbf{a}}_k \mathbf{V} \mathbf{x}[k]^T \right\} = \sum_{k=1}^{n} L(\underline{\mathbf{a}}_k, \mathbf{V}), \end{aligned} \tag{10}$$

using the exponential family definition in (3).

It can be shown that the loss function (10) is convex in either of its arguments with the others fixed [4]. Hence, its minimization is attacked by using an iterative approach. The Newton-Raphson method is used for the iterative minimization. The first step in the $(l+1)^{\text{th}}$ iteration consists of the update $\mathbf{A}^{(l+1)} = \arg\min_{\mathbf{A}} L(\mathbf{A}, \mathbf{V}^{(l)})$, with $\mathbf{V}^{(l)}$ the update obtained at the end of the l^{th} iteration. The Newton-Raphson technique solves this problem by using the update

$$\begin{aligned} \underline{\mathbf{a}}_k^{(l+1)} =& \underline{\mathbf{a}}_k^{(l)} - \alpha_{\mathbf{a}}^{(l+1)} \left(\nabla_{\mathbf{a}}^2 L\big(\underline{\mathbf{a}}_k^{(l)}, \mathbf{V}^{(l)}\big) \right)^{-1} \\ & \cdot \nabla_{\mathbf{a}} L\big(\underline{\mathbf{a}}_k^{(l)}, \mathbf{V}^{(l)}\big) \end{aligned} \tag{11}$$

for $k = 1, \ldots, n$, where $\nabla L(\cdot)$ is the gradient of the function $L(\cdot)$, $\nabla^2 L(\cdot)$ its Hessian matrix and $\alpha^{(l+1)}$ the so-called step size. It is easily shown that, for $k = 1, \ldots, n$,

$$\nabla_{\mathbf{a}} L\big(\underline{\mathbf{a}}_k^{(l)}, \mathbf{V}^{(l)}\big) = \mathbf{V}^{(l)} \Big(G'\big(\underline{\mathbf{a}}_k^{(l)} \mathbf{V}^{(l)}\big) - \mathbf{x}[k]^T \Big),$$

where

$$G'\big(\underline{\mathbf{a}}_k \mathbf{V}^{(l)}\big) = \left. \frac{\partial G(\underline{\boldsymbol{\theta}}_k)}{\partial \underline{\boldsymbol{\theta}}_k} \right|_{\underline{\boldsymbol{\theta}}_k = \underline{\mathbf{a}}_k \mathbf{V}^{(l)}}, \quad \frac{\partial \underline{\boldsymbol{\theta}}_k}{\partial \underline{\mathbf{a}}_k} = \mathbf{V}^{(l)}.$$

Furthermore, for $k = 1, \ldots, n$,

$$\nabla_{\mathbf{a}}^2 L\big(\underline{\mathbf{a}}_k^{(l)}, \mathbf{V}^{(l)}\big) = \mathbf{V}^{(l)} G''\big(\underline{\mathbf{a}}_k^{(l)} \mathbf{V}^{(l)}\big) \mathbf{V}^{(l),T},$$

where $G''\big(\underline{\mathbf{a}}_k^{(l)} \mathbf{V}^{(l)}\big)$ is a $(d \times d)$-diagonal matrix with the diagonal terms equal $\partial^2 G(\underline{\boldsymbol{\theta}}_k)/\partial \underline{\theta}_{k,i}^2$, $i = 1, \ldots, d$. (Note that the diagonal structure of $G''(\cdot)$ is exact and *not* an approximation.) Similarly, the second step in the iterative minimization method consists of the update $\mathbf{V}^{(l+1)} = \arg\min_{\mathbf{V}} L(\mathbf{A}^{(l+1)}, \mathbf{V})$. This update takes the form

$$\begin{aligned} \mathbf{v}_r^{(l+1)} =& \mathbf{v}_r^{(l)} - \alpha_{\mathbf{v}}^{(l+1)} \left(\nabla_{\mathbf{v}}^2 L\big(\underline{\mathbf{a}}_k^{(l+1)}, \mathbf{V}^{(l)}\big) \right)^{-1} \\ & \cdot \nabla_{\mathbf{v}} L\big(\underline{\mathbf{a}}_k^{(l+1)}, \mathbf{V}^{(l)}\big) \end{aligned} \tag{12}$$

for $r = 1, \ldots, q$, where

$$\nabla_{\mathbf{v}} L\big(\underline{\mathbf{a}}_k^{(l+1)}, \mathbf{V}^{(l)}\big) = \sum_{k=1}^{n} \underline{\mathbf{a}}_k^{(l+1)} \big\{ G'\big(\underline{\mathbf{a}}_k^{(l+1)} \mathbf{V}^{(l)}\big) - \mathbf{x}[k]^T \big\},$$

$$\nabla_{\mathbf{v}}^2 L\big(\underline{\mathbf{a}}_k^{(l+1)}, \mathbf{V}^{(l)}\big) = \sum_{k=1}^{n} \big(\underline{\mathbf{a}}_k^{(l+1)}\big)^2 G''\big(\underline{\mathbf{a}}_k^{(l+1)} \mathbf{V}^{(l)}\big).$$

For exponential family distributions and canonical link functions, it can be shown that the update equations correspond to normal equations in a least squares environment [11]. Therefore the minimization problem corresponds to an Iteratively Reweighted Least Squares (IRLS) algorithm. developed for a large set of exponential family distributions (Gaussian, exponential for the continuous distributions, Bernoulli, binomial and Poisson for the discrete distributions). IRLS-based iterative updates exist for a large set of exponential family distributions, including Gaussian, exponential, Bernoulli, binomial and Poisson.

3.1. Positivity constraints

For the exponential and inverse Gaussian distributions, an additional positivity constraint on the natural parameter values has to be taken into account in order to fully comply with their definition. Three alternative ways to deal with the positivity constraint are: (1) the use of Lagrange multipliers and Kuhn-Tucker theory; (2) the use of penalty functions; and (3) the use of a non-canonical link function which enables one to

work in an unconstrained parameter space. The latter option is investigated. Here, the non-canonical link function is chosen to be the composition of the canonical link function with the absolute value function, and the loss function becomes:

$$\widetilde{L}(\mathbf{V},\mathbf{A}) = \sum_{k=1}^{n} \left\{ G(-|\underline{\mathbf{a}}_k \mathbf{V}|) + |\underline{\mathbf{a}}_k \mathbf{V}|\, \mathbf{x}[k]^T \right\}.$$

The iterative minimization algorithm based on IRLS is then used on $\widetilde{L}(\mathbf{V},\mathbf{A})$ as done previously on $L(\mathbf{V},\mathbf{A})$.

3.2. Mixed data

The case of hybrid or mixed data occurs for a problem in which different types of distributions can be used for different descriptors. For simplicity of presentation, two types of exponential family distribution, $p^{(\mathbf{1})}$ and $p^{(\mathbf{2})}$, are discussed here. The matrix of observations becomes $\mathbf{X} = (\mathbf{X}^{(\mathbf{1})}|\mathbf{X}^{(\mathbf{2})})$, the parameters matrix $\mathbf{\Theta} = \mathbf{AV} = (\mathbf{\Theta}^{(\mathbf{1})}|\mathbf{\Theta}^{(\mathbf{2})})$, the lower dimensional subspace basis matrix $\mathbf{V} = (\mathbf{V}^{(\mathbf{1})}|\mathbf{V}^{(\mathbf{2})})$. However, the matrix of principal components $\mathbf{A}$ remains common to both $\mathbf{\Theta}^{(\mathbf{1})}$ and $\mathbf{\Theta}^{(\mathbf{2})}$. Then, the loss function (10) takes the following form:

$$\begin{aligned} L(\mathbf{V},\mathbf{A}) &= L^{(\mathbf{1})}(\mathbf{V}^{(\mathbf{1})},\mathbf{A}) + L^{(\mathbf{2})}(\mathbf{V}^{(\mathbf{2})},\mathbf{A}) \\ &= \sum_{k=1}^{n} \left\{ G^{(\mathbf{1})}\big(\underline{\mathbf{a}}_k \mathbf{V}^{(\mathbf{1})}\big) - \big(\underline{\mathbf{a}}_k \mathbf{V}^{(\mathbf{1})}\big)\mathbf{x}^{(\mathbf{1})}[k]^T \right\} \\ &\quad + \sum_{k=1}^{n} \left\{ G^{(\mathbf{2})}\big(\underline{\mathbf{a}}_k \mathbf{V}^{(\mathbf{2})}\big) - \big(\underline{\mathbf{a}}_k \mathbf{V}^{(\mathbf{2})}\big)\mathbf{x}^{(\mathbf{2})}[k]^T \right\} \end{aligned}. \tag{13}$$

As done previously the loss (13) is minimized using the Newton-Raphson approach. In order to avoid confusion, the step superscripts $^{(l)}$ and $^{(l+1)}$ are not bold whereas the mixture superscripts $^{(\mathbf{1})}$ and $^{(\mathbf{2})}$ are. For the first step, the update equations for $k = 1, \ldots, n$ are:

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

For the second step, the two sets of row vectors $\left\{\mathbf{v}_r^{(\mathbf{1})}\right\}_{r=1}^{q}$ and $\left\{\mathbf{v}_r^{(\mathbf{2})}\right\}_{r=1}^{q}$ are updated separately. For the sake of simplicity, the following derivations are made for the set $\left\{\mathbf{v}_r\right\}_{r=1}^{q}$ indistinctively of the mixed data superscript. The update equations can then be used for $\left\{\mathbf{v}_r^{(\mathbf{1})}\right\}_{r=1}^{q}$ and $\left\{\mathbf{v}_r^{(\mathbf{2})}\right\}_{r=1}^{q}$ by changing $\mathbf{v}_r$ to $\mathbf{v}_r^{(\mathbf{1})}$, respectively to $\mathbf{v}_r^{(\mathbf{2})}$, $G(\cdot), G'(\cdot)$, 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 $G^{(\mathbf{2})''}(\cdot)$. Then, the update equations are as follows for $r = 1, \ldots, q$:

$$\begin{aligned} \mathbf{v}_r^{(l+1),T} &= \mathbf{v}_r^{(l),T} \\ &\quad - \alpha_{\mathbf{v}}^{(l+1)} \left(\sum_{k=1}^{n} \big(\underline{a}_{k,r}^{(l+1)}\big)^2 G''\big(\underline{\mathbf{a}}_k^{(l+1)} \mathbf{V}^{(l)}\big) \right)^{-1} \\ &\quad \cdot \left(\sum_{k=1}^{n} \underline{a}_{k,r}^{(l+1)} \left\{ G'\big(\underline{\mathbf{a}}_k^{(l+1)} \mathbf{V}^{(l)}\big) - \mathbf{x}[k]^T \right\} \right). \end{aligned}$$

This approach can be naturally generalized to any number s of exponential family distributions, resulting in a single update equation for $\mathbf{A}$ and s independent update equations for $\mathbf{V}$.

3.3. Minority class detection algorithm

The minority class detection technique using the IRLS-based learning algorithm in the parameter space works as follows: first, given the training set $\{\mathbf{x}[k]\}_{k=1}^{n}$, we learn the direction of projection in the parameter space, namely $\mathbf{V}$, by using the IRLS-based iterative algorithm, and compute the training set mean-image in the parameter space, namely $\frac{1}{n}\sum_{k=1}^{n} \underline{\mathbf{a}}_k$. Then, the new data point is moved from the data space to the parameter space using the link function. We project the obtained point onto the learned direction of projection and compute its distance to the training set mean-image. Finally, we compare the obtained distance to a given threshold to make a decision. If the distance is greater than the threshold, then the new point is declared to belong to the minority class, otherwise it is declared to belong to the majority class.

4. SIMULATION RESULTS

The IRLS-based learning algorithm has been implemented for a dictionary of exponential family distributions: Gaussian, exponential distribution for continuous data, Bernoulli, binomial and Poisson distribution for discrete/count data.

Fig. 2 below 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 Gaussian component. The data are generated by two different classes, a minority 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 detection performance, we consider a situation where the minority class is a rare occurrence (1 percent of 10,000 training points), and we perform the detection algorithm described above. The proposed technique is compared to classical PCA used in the data space with a threshold test performed on new data projected along the first principal axis, as well as to a supervised Bayes (minimum rate) detector for the sake of an optimal benchmark.

Fig. 3 shows a comparison between the supervised Bayes detector, the minority class detector based on the utilization of the proposed algorithm to perform detection in the parameter space, and the minority class detector based on classical PCA. This illuminating example shows that there are domains for which classical PCA performs far from optimal.

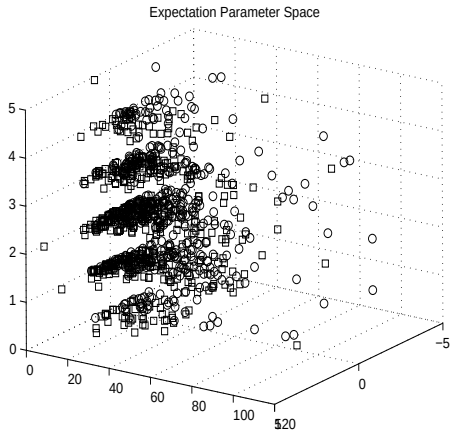

Fig. 2. Data samples of a 3-dimensional data having binomial, exponential and Gaussian components (circles for one class and squares for the other class).

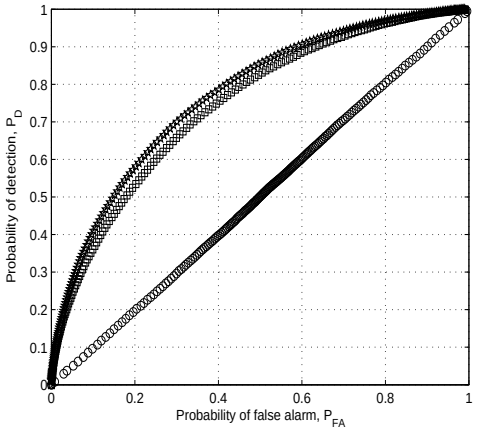

Fig. 3. Comparison of supervised Bayes optimal (top with pentagrams), proposed GLS technique (middle with squares) and classical PCA (bottom with circles) ROC curves.

5. CONCLUSION

A graphical model approach for minority class detection in an unsupervised learning context for data of mixed type was proposed and referred to as Generalized Linear Statistics (GLS). An Iteratively Reweighted Least Squares algorithm was presented for learning distribution parameters of observed nodes, as well as a nonparametric density estimation method for hidden nodes. Detection was then performed using discriminant thresholding in the parameter space, instead of the data space as in traditional methods. In contrast to classical methods, the proposed method allows for each data component to have its own parametric form and enables unsupervised minority class detection in the case of a rare occurrence of minority class objects. Initial results on synthetic data are encouraging, and they allow for the prediction of quality results on financial data sets which is now underway.

Furthermore, the possibility of utilizing novel hybrid detection techniques that work partially in data space and partially in parameter space is being investigated.

6. REFERENCES

[1] M. I. Jordan and T. J. Sejnowski, *Graphical Models: Foundations of Neural Computation*, (MIT Press, 2001).

[2] C. Levasseur, K. Kreutz-Delgado, U. Mayer and G. Gancarz, Data-pattern discovery methods for detection in nongaussian high-dimensional data sets, *Asilomar Conference on Signals, Systems and Computers*, 2005.

[3] C.E. McCulloch, *Generalized Linear Mixed Models*, (Institute of Mathematical Statistics, 2003).

[4] M. Collins, S. Dasgupta and R. Shapire, A generalization of principal component analysis to the exponential family, *Neural Information Processing Systems*, 2001.

[5] Sajama and A. Orlitsky, Semi-parametric exponential family PCA, *Neural Information Processing Systems*, 2004.

[6] D. J. Bartholomew and M. Knott, *Latent Variable Models and Factor Analysis*, (Oxford University Press, 2nd edition, 1999).

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

[8] B. G. Lindsay, *Mixture Models: Theory, Geometry, and Applications*, (Institute of Mathematical Statistics, 1995).

[9] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, (John Wiley, 2nd edition, 2001).

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

[11] L. Fahrmeir and G. Tutz, *Multivariate Statistical Modelling Based on Generalized Linear Models*, (Springer-Verlag, 2nd edition, 2001).