Data-Pattern Discovery Methods for Detection in Nongaussian High-dimensional Data Sets

Cécile Levasseur[1], K. Kreutz-Delgado[1], U. Mayer[2] and G. Gancarz[2]
[1]University of California, San Diego, Jacobs School of Engineering & [2]Fair Isaac Corporation

Abstract

Many important expert system applications depend on the ability to accurately detect or predict the occurrence of key events given a data set of observations. We concentrate on multidimensional data that are highly nongaussian (continuous and/or discrete), noisy and nonlinearly related. We investigate the feasibility of data-pattern discovery and event detection by applying generalized principal component analysis (GPCA) techniques for pattern extraction based on an exponential family probability distribution assumption. We develop theoretical extensions of the GPCA model by exploiting results from the theory of generalized linear models and nonparametric mixture density estimation.

1 Introduction

We are interested in the unsupervised learning of statistical models and the development of algorithms which will enable the detection of fraudulent or anomalous events.

In many important risk assessment problems (medical, financial, economic, and others), the properties of the measured objects are as follows:

- *nongaussian*,
- *very noisy*,
- *highly nonlinearly related*,
- *high-dimensional*,
- *hybrid*: comprised of categorical, count, and continuous measurements.

As a consequence, the resulting statistical modeling and detection problem is very difficult.

In order to solve this problem, we seek valuable insight into the underlying probabilistic structure of the data. We aim at enabling low-dimensional parameterizations as well as associated low-dimensional approximate statistics which can be used for fraud or anomaly detection.

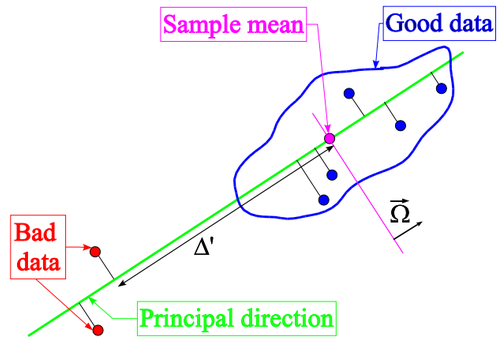
We consider a unifying approach called generalized linear statistics (GLS). This approach assembles generalized latent variable models (GLV's) and random-effects generalized linear models (RE-GLM's).
The RE-GLM theory is a generalization of the classical theory of linear regression, while the GLV theory is a generalization of the classical theory of factor analysis (FA) and principal component analysis (PCA). In both cases, the generalization is based on a move from the data/description space containing the measurement vectors to the parameter space via a generally nonlinear transformation known as a link function.

In the latter space, a linear relationship is assumed to hold.

A variety of well-known techniques fall into the GLS approach. Here, we present a single application of such generalized techniques within the framework of fraud detection. We show that principal component analysis (PCA) can be outperformed by its generalization to exponential family distributions (GPCA) when the data is not Gaussian distributed.

2 Principal Component Analysis (PCA)

Principal component analysis (PCA) seeks a projection that best represents the data in a least squares sense. Equivalently, the high-dimensional data is projected onto a lower dimensional subspace defined as the direction of maximum variance.



Principal component analysis (PCA): projecting along the direction of maximum variance

A detection technique based on PCA works well when the data is *linearly separable* along a principal direction.

Probabilistic interpretation of PCA: The high-dimensional data points are considered to be noise-corrupted versions of some true points which lie in a lower-dimensional subspace; the goal is to find these true points, and the main assumption is that the noise is Gaussian distributed.

3 Generalized PCA

Generalized PCA is a generalization of principal component analysis to the exponential family recently proposed by Collins, Dasgupta and Shapire. The noise distribution assumption is extended to the rest of the exponential family distributions.
This technique takes advantage of exponential family properties, providing a generalization of generalized linear models (GLM's) as follows.

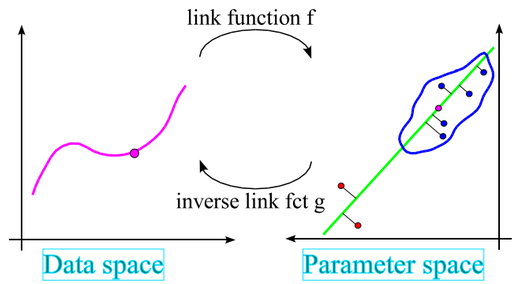
- The standard Gaussian linear model

$$\boldsymbol{\mu} \triangleq E[\mathbf{x}|\boldsymbol{\theta}] = \boldsymbol{\theta} = \boldsymbol{\beta} + \mathbf{V} \cdot \mathbf{a}$$

  - $p(\mathbf{x}|\boldsymbol{\theta})$ is a Gaussian distribution.
  - $\mathbf{V}$ is *deterministic & known*
  - $\boldsymbol{\beta}$ and $\mathbf{a}$ are *deterministic & unknown*
- The standard generalized linear model (GLM)

$$\boldsymbol{\mu} \triangleq E[\mathbf{x}|\boldsymbol{\theta}] \qquad f(\boldsymbol{\mu}) = \boldsymbol{\theta} = \boldsymbol{\beta} + \mathbf{V} \cdot \mathbf{a}$$

  - $p(\mathbf{x}|\boldsymbol{\theta})$ is a member of the exponential family
  - $\mathbf{V}$ is *deterministic & known*
  - $\boldsymbol{\beta}$ and $\mathbf{a}$ are *deterministic & unknown*
  - $f(\cdot)$ is called the *canonical link* function.
- The random effect generalized linear model (RE-GLM)
  - $\mathbf{a}$ is *random & unknown*
- The blind random effect generalized linear model (B-RE-GLM)
  - $\mathbf{V}$ is *deterministic & unknown*



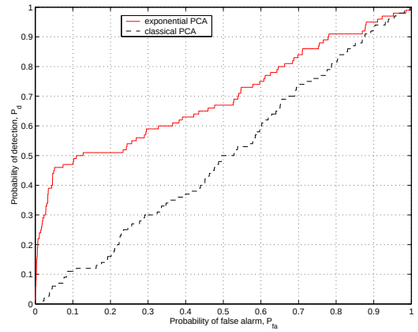
Exponential PCA in data space ⇔ PCA in parameter space

Using the generic discrepancy between the data space and the parameter space for exponential family distributions, the generalized PCA performs a PCA-like linear regression and feature extraction in the parameter space, which corresponds to a nonlinear regression and feature extraction in the data space.

4 Detection Performance

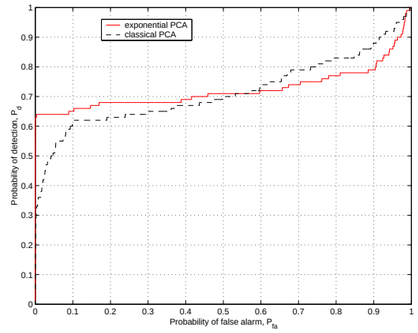
To assess the performance of the detection techniques, we provide ROC curves (probability of detection P_D versus probability of false alarm P_{FA}).
The synthetic data are independently drawn from an exponential distribution.

We use a 1-dimensional projection for both the PCA and the generalized PCA techniques. Following the synthetic data structure, generalized PCA is applied with the assumption of exponentially distributed data.



Synthetic data - ROC curves: Comparing the performance of the detection using 1-dim. PCA vs. 1-dim. generalized PCA

We now use a 2-dimensional projection for both the PCA and the generalized PCA techniques.



Synthetic data - ROC curves: Comparing the performance of the detection using 2-dim. PCA vs. 2-dim. generalized PCA

We intend to apply those techniques on UCI repository data.

5 Conclusion

Generalized linear statistics (GLS) techniques provide a valuable insight into the underlying structure of the data, enabling good fraud detection performance. We presented a synthetic data example for which generalized PCA performs significantly better than classical PCA in terms of fraud detection.
In the future, we will extend these results to hybrid multidimensional projection.