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

Cécile Levasseur and Kenneth Kreutz-Delgado
Jacobs School of Engineering
University of California, San Diego
La Jolla, CA 92093-0407
clevasse@ucsd.edu, kreutz@ece.ucsd.edu

Uwe Mayer and Gregory Gancarz
Fair Isaac Corporation
San Diego, CA 92130
{UweMayer, GregoryGancarz}@fairisaac.com

***Abstract*— Many important analytic 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 in such domains 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.**

I. Introduction

Many important risk assessment system applications depend on the ability to accurately predict the probabilities of key events given a large data set of observations. For example this problem arises in medicine ("Do the epidemiological data suggest that the trace elements in the local water supply cause cancer?"); health care ("Do the descriptors associated with the professional behavior of a medical doctor suggest that he/she is an outlier in the category he/she was assigned to?"); and drug discovery ("Do the molecular descriptors associated with known drugs suggest that a new, candidate drug will have low toxicity and high effectiveness?"). In many of these domains, there is little or no *a priori* knowledge regarding the true sources of any causal relationships which may exist between variables of interest. In these situations, meaningful information regarding the occurrences of key events must be extracted from the data itself, a problem which can be viewed as an important application of data-driven pattern recognition or prediction. The problem of *unsupervised* data-driven detection or prediction is one of relating descriptors of a large unlabeled database of "objects" to measured properties of these objects, and then using these empirically determined relationships to infer or detect the properties of new objects. This work considers measured object properties that are nongaussian (and comprised of continuous and discrete data), very noisy, and highly nonlinearly related. Data comprised of measurements of such disparate properties are said to be hybrid or of mixed type. As a consequence, the resulting detection problem is very difficult. The difficulties are further compounded because the descriptor space of objects is of high dimension.

This work is specifically concerned with efficient statistical modeling of unlabeled, nongaussian, high-dimensional, mixed continuous and discrete data for the practical purposes of: 1) creating statistically faithful synthetic data for statistical testing of proposed anomaly detection algorithms; and 2) developing unsupervised learning-based anomaly detection algorithms. This paper describes a first examination of the theoretical development of a promising generalized principal component analysis (GPCA) technique recently presented in [6]. This approach is based on the use of exponential families of distributions to model the various types (continuous and discrete) of data measurements which consist of the components of a single vector observation, $\mathbf{x}$, of relevant variables, x_i, encompassing measured properties of an "object" of interest. For this model, we need to determine both the type of distribution used for each data feature, x_i, and the natural parameter θ_i appropriate for this distribution. The constraint is then imposed that the vector of natural parameters $\boldsymbol{\theta}$ lies in a lower dimensional subspace.

This approach exploits the well-known distinction which exists between the data space and the parameter space for exponential family distributions [9]. When a new data vector is observed, the data is first transformed to the parameter space using a *link function* which is obtained from the model, and then its image in the parameter space is projected into a lower dimensional subspace. This process is a generalization

of the classical principal component analysis (PCA) and is effectively a way of projecting the data onto *principal components in the natural parameter space*. These principal components do not lie in the data space as in conventional projection, but instead lie in a hyperplane of the parameter space, and provide a novel way to extract features from data of mixed type.

For anomaly detection, the Euclidean distance of the new projected point is compared to the sample mean of the projected points obtained from the training set. The receiver operating characteristics (ROC) curve of the detector shows its performance as a trade off between selectivity and sensitivity. The curve presents the probability of detection as a function of the probability of false alarm and is obtained by varying the sensitivity or threshold parameter. A noteworthy additional benefit of this approach is that having fit an exponential family model to the data will allow to generate synthetic data Monte Carlo-based assessments of the model and the performance of proposed detection algorithms.

Theoretical extensions of the model are developed by exploiting results from the theory of generalized linear models [9] and nonparametric mixture density estimation [1].

II. Bregman divergence and exponential families

A distribution is said to be a member of the exponential family if it has a density function of the form

$$p(x;\theta) \quad = \quad \exp(x\theta - G(\theta))p_0(x),$$

where $p_0(x)$ represents any factor of the density which does not depend on θ. Equivalently, one can write

$$\log p(x;\theta) \quad = \quad \log p_0(x) + x\theta - G(\theta).$$

Given a function $G(\theta)$ and its gradient $g(\theta)$, the "dual" function $F(x)$ and its gradient $f(x)$ are given by

$$\begin{aligned} F(g(\theta)) + G(\theta) &= g(\theta)\cdot\theta, \\ f(x) = F'(x) &= g^{-1}(x). \end{aligned}$$

Let $F : \Delta \to \mathbb{R}$ be a differentiable and strictly convex function defined on a closed, convex set $\Delta \subseteq \mathbb{R}$. The *Bregman divergence* [3] associated with F is defined for φ, $\psi \in \Delta$ to be

$$B_F(\varphi\|\psi) \quad \triangleq \quad F(\varphi) - F(\psi) - f(\psi)(\varphi - \psi),$$

where $f(x) = F'(x)$. The negative log-likelihood function of a scalar exponentially distributed random variable x can be expressed in terms of the Bregman divergence as

$$-\log p(x;\theta) \quad = \quad -\log p_0(x) - F(x) + B_F(x\|g(\theta)).$$

The Bregman divergence can be defined between vectors or matrices as well.

III. The Generalized Linear Model

Here the standard linear model is first generalized to accommodate nongaussian outcome variables [9]. An alternate theoretical derivation for the generalized PCA method given in [6] is then developed.

A. The standard Gaussian linear model

Consider the probability density function $p(\mathbf{x}|\boldsymbol{\theta})$ to be a Gaussian distribution with mean $\boldsymbol{\mu}$ and known covariance matrix. Note that, in the special case of a Gaussian assumption, the mean vector $\boldsymbol{\mu}$ is equal to the parameter $\boldsymbol{\theta}$ of the distribution. The standard Gaussian linear model expresses the mean vector $\boldsymbol{\mu}$ in a linear structure as follows:

$$\boldsymbol{\mu} \triangleq E[\mathbf{x}|\boldsymbol{\theta}] = \boldsymbol{\theta} = \mathbf{b} + \mathbf{V}\,\mathbf{a}. \tag{1}$$

The matrix $\mathbf{V}$ is assumed to be *deterministic and known*, and the vectors $\mathbf{a}$ and $\mathbf{b}$ *deterministic and unknown*.

B. The generalized linear model (GLM)

The probability density function $p(\mathbf{x}|\boldsymbol{\theta})$ is now assumed to be a member of the exponential family of distributions, with parameter vector $\boldsymbol{\theta}$. Since the Gaussian distribution belongs to the exponential family, allowing $p(\mathbf{x}|\boldsymbol{\theta})$ to be any member of the family means generalizing the standard Gaussian linear model. The mean vector $\boldsymbol{\mu}$ is *linked* to the natural parameter $\boldsymbol{\theta}$ by using the so called "canonical link" function $f(\cdot)$. Then, as done previously for the standard Gaussian linear model, the parameter $\boldsymbol{\theta}$ is expressed in a linear structure as follows:

$$\boldsymbol{\mu} \triangleq E[\mathbf{x}|\boldsymbol{\theta}] \quad \text{and} \quad f(\boldsymbol{\mu}) = \boldsymbol{\theta} = \mathbf{b} + \mathbf{V}\,\mathbf{a}. \tag{2}$$

Again, the matrix $\mathbf{V}$ is assumed to be *deterministic and known*, and the vectors $\mathbf{b}$ and $\mathbf{a}$ *deterministic and unknown*. The link function provides a bijective relationship between the data space and the parameter space.

C. The random effect generalized linear model (RE-GLM)

The model is the same as that described for the generalized linear model, i.e.,

$$f(\boldsymbol{\mu}) = \boldsymbol{\theta} = \mathbf{b} + \mathbf{V}\,\mathbf{a}, \tag{3}$$

except that now the vector $\mathbf{a}$ is assumed to be *random and unknown*.

D. The blind random effect generalized linear model (BRE-GLM)

The *blind* random effect generalized linear model differs from the RE-GLM in that the matrix $\mathbf{V}$ is additionally assumed to be deterministic and *unknown*. The generalized PCA method described in [6] belongs to the large class of BRE-GLM's, and this is the model considered in this paper.

IV. Underlying Statistical Structure Discovery Method

A. Theoretical framework

A particular "object" of interest can be associated with a variety of descriptor random variables. These descriptors can be viewed as comprising the components of a random vector $\mathbf{x} = (x_1, \cdots, x_d)^T \in \mathbb{R}^d$, where the dimension d is equal to the number of descriptors. A given set of n observed d-dimensional points $\{\mathbf{x}[k]\}_{k=1}^n = \{(x_1[k], \cdots, x_i[k], \cdots, x_d[k])^T\}$ is considered.

The following assumptions are made:

- the samples $\mathbf{x}[k]$, k = 1,$\cdots, n$ are drawn independently;
- the components x_i, i = 1,$\cdots, d$ are independent when conditioned on the random parameter vector $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d)^T \in \mathbb{R}^d$, i.e., $p(\mathbf{x}|\boldsymbol{\theta}) = p_1(x_1|\theta_1) \cdots p_d(x_d|\theta_d)$;
- $p_i(x_i|\theta_i)$ is any one-parameter exponential family distribution with θ_i taken to be the natural parameter of the exponential family density p_i.

The marginal densities $p_i(\cdot|\cdot)$ can all be different, allowing for the possibility of $\mathbf{x}$ containing continuous and discrete valued components.

Let $\mathbf{X}$ be the $n \times d$ matrix of observations or descriptors whose k^{th} row is $\mathbf{x}[k]$. Let $\boldsymbol{\Theta}$ be a $n \times d$ matrix of corresponding parameters whose k^{th} row is $\boldsymbol{\theta}[k]$. Following the probabilistic generalized latent variable formalism described in [4], collected data points are assumed to have been generated from populations having class-conditional probability density functions[1] satisfying the previously stated assumptions, and the corresponding log-likelihood function takes the following form:

$$p(\mathbf{X}|\boldsymbol{\Theta}) = \prod_{k=1}^{n} \prod_{i=1}^{d} p_i(x_i[k]|\theta_i[k]). \quad (4)$$

Following the BRE-GLM model described in the previous section,

$$\boldsymbol{\theta}[k] = \mathbf{b} + \mathbf{V}\,\mathbf{a}[k] \quad (5)$$

with $\mathbf{V} \in \mathbb{R}^{d \times q}$ and $\mathbf{b} \in \mathbb{R}^d$ deterministic, $\mathbf{a}[k] \in \mathbb{R}^q$ random where $q < d$ (and ideally $q \ll d$). Following the BRE-GLM model, *all* of the quantities $\mathbf{V}$, $\mathbf{b}$, and $\mathbf{a}$ are assumed to be unknown, and hence need to be identified[2]. This estimation is performed by maximizing the log-likelihood function (4), i.e., by minimizing the negative log-likelihood function. As explained in Section II, this is identical to minimizing the corresponding Bregman divergence. Hence, [6] exploits the properties of a Bregman divergence to create an iterative minimization algorithm to solve the estimation problem. Learning the matrix $\mathbf{V}$ and the vector $\mathbf{b}$ implies identifying a lower dimensional subspace in the parameter space.

[1]Delta-functions are admitted so that densities are well-defined for discrete, continuous, and mixed random variables.

[2]Resulting in a so-called "blind" estimation problem.

B. Nonparametric maximum likelihood estimation

Given the n iid random variables $\mathbf{X} = [\mathbf{x}_1; \cdots; \mathbf{x}_n]$, the (nonconditional) density $p(\mathbf{X})$ requires a generally difficult integration over the parameters:

$$\begin{aligned} p(\mathbf{X}) &= \prod_{k=1}^{n} \int p(\mathbf{x}[k]|\boldsymbol{\theta}[k])\pi(\boldsymbol{\theta}[k])d\boldsymbol{\theta}[k] \quad (6) \\ &= \prod_{k=1}^{n} \int \prod_{i=1}^{d} p_i(x_i[k]|\theta_i[k])\pi(\boldsymbol{\theta}[k])d\boldsymbol{\theta}[k] \end{aligned}$$

where $\pi(\boldsymbol{\theta}[k])$ is the probability density function of $\boldsymbol{\theta}[k] = \mathbf{b} + \mathbf{V}\mathbf{a}[k]$. For specified exponential family densities $p_i(\cdot|\cdot)$, $i = 1, \cdots, d$, maximum likelihood identification of the model (6) corresponds to identifying $\pi(\boldsymbol{\theta})$, which, under the condition $\boldsymbol{\theta} = \mathbf{b} + \mathbf{V}\,\mathbf{a}$, corresponds to identifying the matrix $\mathbf{V}$, the vector $\mathbf{b}$, and a density function, $\mu(\mathbf{a})$, on the vector $\mathbf{a}$ via a maximization of the likelihood function $p(\mathbf{X})$ with respect to $\mathbf{V}$, $\mathbf{b}$, and $\mu(\mathbf{a})$. This is generally a quite difficult problem [9] and is usually solved using approximation methods which correspond to replacing the integral in (6) by a sum [2].

$$\begin{aligned} p(\mathbf{X}) &= \prod_{k=1}^{n} \sum_{j=1}^{m} p(\mathbf{x}[k]|\boldsymbol{\theta}_j)\pi_{j,k} \quad (7) \\ &= \prod_{k=1}^{n} \sum_{j=1}^{m} \prod_{i=1}^{d} p_i(x_i[k]|\theta_{i,j})\pi_{j,k} \end{aligned}$$

over a finite number of support points $\boldsymbol{\theta}_j$ (equivalently, $\mathbf{a}_j$) for $j = 1, \cdots, m$ with point mass probabilities[3]

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

$$\pi_{j,k} \triangleq \pi(\boldsymbol{\theta}[k] = \boldsymbol{\theta}_j) = \pi(\mathbf{a}[k] = \mathbf{a}_j) = \pi_j.$$

In particular, $\pi_{j,k}$ is actually independent of k. As clearly described in [2], this approximation is justified either as a Gaussian quadrature approximation to the integral in (6) [9] or by appealing to the fact that the *nonparametric maximum likelihood estimate* (NMLE) of the mixture density $\pi(\boldsymbol{\theta})$ yields a solution which takes a finite number of points of support [8]. With

[3]Note that $\boldsymbol{\theta}$, $\mathbf{a}$, $\boldsymbol{\theta}[k]$, and $\mathbf{a}[k]$ are (discrete) *random variables* while $\boldsymbol{\theta}_j$ and $\mathbf{a}_j$, $j = 1, \cdots, m$ are the m *nonrandom* support point values, i.e., the values of the random variables having nonzero probabilities. Also note that taking $\pi(\boldsymbol{\theta}_j) = \pi(\mathbf{a}_j)$ for $\boldsymbol{\theta}_j = \mathbf{V}\mathbf{a}_j + \mathbf{b}$ means that we are assuming that the relationship between the discrete values $\boldsymbol{\theta}_j$ and $\mathbf{a}_j$ is one-to-one.

$\boldsymbol{\theta} = \mathbf{V}\mathbf{a} + \mathbf{b}$, with $\mathbf{V}$, $\mathbf{b}$ fixed and $\mathbf{a}$ random, the likelihood (7) is equal to

$$\begin{aligned} p(\mathbf{X}) &= \prod_{k=1}^{n}\sum_{j=1}^{m} p(\mathbf{x}[k]|\boldsymbol{\theta}_j)\pi_j \\ &= \prod_{k=1}^{n}\sum_{j=1}^{m} p(\mathbf{x}[k]|\mathbf{V}\mathbf{a}_j + \mathbf{b})\pi_j. \end{aligned} \quad (8)$$

The generalized PCA method described in [6] does not perform the estimation of the point-mass probability estimates and instead assumes that there are $m = n$ points of supports $\boldsymbol{\theta}_j$ with $\pi_j = 1/n$ for all $j = 1, \cdots, m$.

V. Synthetic data experiments

This paper presents anomaly detection simulation results obtained by using synthetic data generated from the BRE-GLM model (4)-(5), comparing the detection performance of the generalized PCA to the detection performance of the classical PCA. It is shown that in certain environments the classical PCA provides poor detection performance while the generalized PCA performs well.

The data vectors lie in a 3-dimensional space and are either drawn from a "bad" class or a "good" class, each class generated according to a density of the form (4). For each "bad" data vector sample, there are 100 "good" data vectors, and a total of 10,000 records were generated. The data are equally divided into a training data set and a test data set. Since the "bad" data is a small proportion of the data set, by learning the underlying structure of the whole training data the underlying structure of the "good" data is approximately learned.

Once the classical PCA and the generalized PCA have learned a feature space projection, for each algorithm the sample mean of the features obtained by projecting the training data is computed. The sample mean then is taken as an approximation to the cluster mean in feature space of the good data class. A new data vector, or test vector, is randomly generated from either class, then projected to a feature vector using the learned projection. The distance between this projected point and the previously computed sample mean is compared to a threshold value λ. The new point is declared to be "bad" (i.e., an outlier) if the distance is higher than λ, otherwise it is declared to be "good". This procedure is done for all of the test set data, and the detection performance is assessed by plotting ROC curves 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.

Data for which the classical PCA will fail to provide accurate detection are easily created, using the knowledge that the classical PCA defines the direction of projection as the direction of maximum variance in data space. The classical PCA will therefore give poor performance on data for which the direction of maximum variance is inappropriate for separating "bad" from "good" data. The exponential distribution $p(x;\theta) = \lambda\exp(-\lambda x), \theta = -\lambda$ is used as an example in the simulations. Because 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 the classical PCA should perform poorly. To test this expectation, two different experiments are performed:

- data latent structure is 1-dimensional ($q = 1$)
- data latent structure is 2-dimensional ($q = 2$)

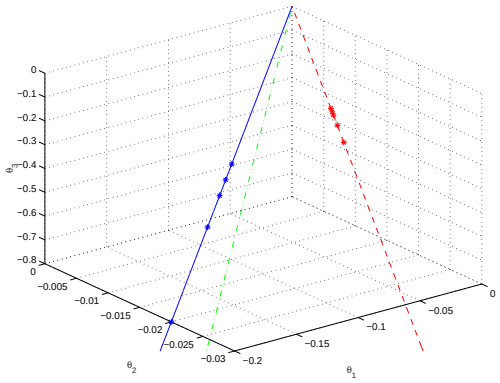

Fig. 1. Parameter Space Feature-subspaces: the 1-dimensional feature subspace spanned by the "good" data is given by the solid line on the left. The subspace spanned by the "bad" data is the dashed line on the right. The unlabeled training data feature subspace learned by the generalized PCA is the dash-point line in the middle.

For the first experiment, Figure (1) shows the two 1-dimensional "good" and "bad" data-generating subspaces, and the feature manifold learned by the generalized PCA algorithm when run on the mixed, unlabeled training data set. For the second experiment, Figure (2) shows the two 2-dimensional data generating subspaces, where each 2-dimensional space is represented by 2 independent vectors lying in the subspace. The 2-dimensional subspace learned by the generalized PCA algorithm when run on the mixed, unlabeled training data-set is close to the "good" subspace, and is not shown on the figure.

The relative detection performances of the generalized PCA and the classical PCA are compared for the two experiments on new data. Figure (3) shows the resulting ROC curves for the 1-dimensional case. Figure (4) presents the obtained ROC curves for the 2-dimensional case. In the 1-dimensional case, the generalized PCA performs uniformly better than the classical PCA. For the 2-dimensional case, the generalized PCA performs significantly better than the

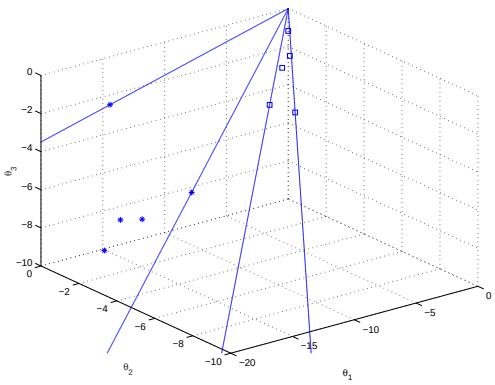

Fig. 2. Parameter Space Feature-subspaces: the 2-dimensional feature subspace spanned by the "good" data contains the two "star-pointed" heavy lines on the left. The subspace spanned by the "bad" data contains the two "square-pointed" thin lines on the right.

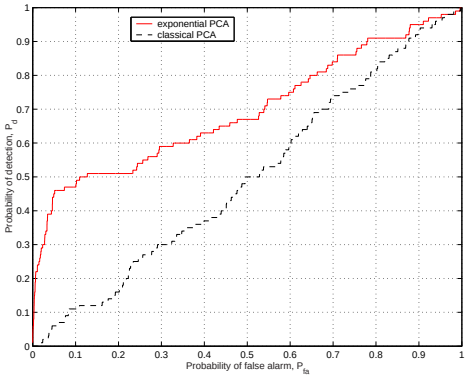

Fig. 3. ROC Curve: comparison of the performance of the classical PCA versus the generalized PCA for the 1-dimensional projection. Note the poor performance of the classical PCA algorithm.

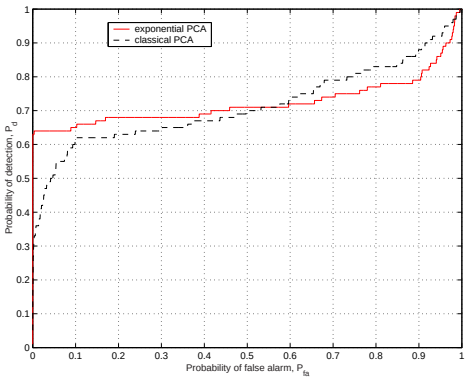

Fig. 4. ROC Curve: comparison of the performance of the classical PCA versus the generalized PCA for the 2-dimensional projection

classical PCA in the low probability of false alarm regime.

VI. Conclusion

This paper focused on the problem of anomaly detection in an unsupervised learning context, utilizing and extending an approach appropriate for exponential family distributions, which has been proposed in [6]. The use of exponential family distributions allows to work with hybrid, or mixed, data having continuous and discrete-valued attributes. Some initial comparisons of the detection performance of the classical principal component analysis (PCA) to the detection performance of the generalized PCA algorithm have been provided. In particular, the probabilistic understanding of the algorithms has been utilized to create a statistical environment for which the classical PCA yields inadequate behavior while the generalized PCA algorithm, specifically designed to fit nongaussian data, provided good detection performance.

Exploiting insights from the RE-GLM literature, [9], [1] the basic model proposed in [6] has been generalized to a nonparametric mixture-prior form of the type analyzed in [1]. The possibility of using this richer class of model for learning true class-prior and class-conditional probabilities is being investigated. If this can be done, then, at least in principle, true Bayes-optimal classifiers can be constructed, yielding superior performance to the anomaly detection described above [7]. Furthermore, the use of other algorithms to perform the required Bregman divergence minimizations is being examined, particularly along the lines proposed in [5].

References

[1] M. Aitkin, "A General Maximum Likelihood Analysis of Overdispersion in Generalized Linear Models," *Statistics and Computing*, **6**:251-62, 1999.

[2] M. Aitkin, "A Generalized Maximum Likelihood Analysis of Variance Components in Generalized Linear Models," *Biometrics*, **55**:117-28, 1999.

[3] K. Azoury & M. Warmuth, "Relative loss bounds for on-line density estimation with the exponential family of distributions," *Machine Learning*, **43**:211-46, 2001.

[4] D. Bartholomew & M. Knott, *Latent Variable Models and Factor Analysis*, 2nd Ed., Arnold Publishers, 1999.

[5] Y. Censor and S. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*, Oxford University Press, 1997.

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

[7] R. Duda, P. Hart & D. Stork, *Pattern Classification*, Wiley, 2001.

[8] Nan Laird, "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of the American Statistical Society*, **73**:805-11, 1978.

[9] C. McCulloch & S. Searle, *Generalized, Linear, and Mixed Models*, Wiley, 2001.

[10] G. McLachlan & D. Peel, *Finite Mixture Models*, Wiley, 2000.

[11] M. Tipping & C. Bishop, "Probabilistic principal component analysis," *J. Royal Stat. Soc., Ser. B*, **61**(3):611-22, 1999.