Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Neurocomputing 71 (2008) 1274–1282
NEUROCOMPUTING
www.elsevier.com/locate/neucom

# Mixtures of robust probabilistic principal component analyzers

Cédric Archambeau[a,*], Nicolas Delannay[b,1], Michel Verleysen[b]

[a]Centre for Computational Statistics and Machine Learning, University College London, Gower Street, London WC1E 6BT, UK
[b]Machine Learning Group, Université catholique de Louvain, 3 Place du Levant, B-1348 Louvain-la-Neuve, Belgium

Available online 2 February 2008

## Abstract

Mixtures of probabilistic principal component analyzers model high-dimensional nonlinear data by combining local linear models. Each mixture component is specifically designed to extract the local principal orientations in the data. An important issue with this generative model is its sensitivity to data lying off the low-dimensional manifold. In order to address this problem, the mixtures of robust probabilistic principal component analyzers are introduced. They take care of atypical points by means of a long tail distribution, the Student-*t*. It is shown that the resulting mixture model is an extension of the mixture of Gaussians, suitable for both robust clustering and dimensionality reduction. Finally, we briefly discuss how to construct a robust version of the closely related mixture of factor analyzers.

*Keywords:* Mixture model; Principal component analysis; Dimensionality reduction; Robustness to outliers; Non-Gaussianity; EM algorithm

## 1. Introduction

Extracting information from high-dimensional data is problematic due to the curse of dimensionality. It is of practical importance to discover an implicit, low-dimensional representation whenever the core of the data lies on one or several directed manifolds. Principal component analysis (PCA) is a well-known statistical technique for linear dimensionality reduction [12,14]. It projects high-dimensional data into a low-dimensional subspace by applying a linear transformation that minimizes the mean squared reconstruction error. PCA is used as a pre-processing step in many applications involving data compression or data visualization. The approach has, however, severe limitations. Since it minimizes a mean squared error, it is very sensitive to atypical observations, which in turn leads to identifying principal directions strongly biased toward them.

Recently, PCA was reformulated as a *robust* probabilistic latent variable model based on the Student-*t* density function [2]. Among others, the (univariate) Student-*t* density arises in the problem of estimating the mean of a Gaussian random variable when the variance is unknown (and the sample size is small) [9]. More generally, the multivariate Student-*t* is a heavy tailed generalization of the multivariate Gaussian. Hence, adjusting the thickness of the distribution tails reduces the sensitivity of its mean and covariance estimates to outliers.

The robust probabilistic reformulation of PCA generalizes standard PPCA [20,26]. Increasing the robustness by replacing Gaussian densities with Student-*t* densities was also proposed in the context of finite mixture modelling [19,1]. In contrast with previous robust approaches to PCA (see for example [28,13], and the references therein), the probabilistic formalism has a number of important advantages. First, it only requires to choose the dimension of the projection space, the other parameters being set automatically by maximum likelihood (ML). Previous attempts need in general to optimize several additional parameters. Second, the probabilistic approach provides a natural framework for constructing mixture models. This enables us to model low-dimensional nonlinear relationships in the data by aligning a collection of local linear generative models, instead of using neighborhood preserving dimensionality reduction techniques [23,25,21,5]. Third, a probabilistic model provides

*Corresponding author.
*E-mail addresses:* c.archambeau@cs.ucl.ac.uk (C. Archambeau), nicolas.delannay@uclouvain.be (N. Delannay), michel.verleysen@uclouvain.be (M. Verleysen).

[1] This author is a research fellow of the Belgian national fund for scientific research (FNRS).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.11.029

likelihood measures for the data, which can be used to compute posterior probabilities and eventually to construct a Bayes classifier [4].

This article introduces mixtures of robust probabilistic principal component analyses (PPCAs). It is based on some earlier work [3] presented at the 15th *European Symposium on Artificial Neural Networks*. The method generalizes mixtures of standard PPCAs [27]. An interesting feature of the approach is that it can be used for robust density estimation and robust clustering, even in high-dimensional spaces. The main advantage resides in the fact that the full-rank, possibly ill-conditioned covariance matrices are approximated by low-rank covariance matrices, where the correlation between the (local) principal directions need not be neglected to avoid numerical instabilities. The number of free parameters per component depends on the specific choice for the dimension of the latent subspace. This procedure is more appealing than constraining the covariance matrices of the mixture components to be diagonal as it is often done in practice. Diagonal covariance matrices lead to axis aligned components, which are in general suboptimal [1].

PCA and PPCA are closely related to factor analysis (FA) [8] and ML FA [22], which can also be combined to form mixtures [11]. The mixture of PPCAs and its robust version assume that the likelihood of the data given the low-dimensional representation is isotropic. When considering a diagonal heteroscedastic noise model instead of the homoscedastic (or isotropic) one, we obtain the mixture of probabilistic factor analyzers (PFAs) [10]. This model is useful when it is reasonable to assume that the noise in the features is independent and of different amplitude. As mixtures of PPCAs, mixtures of PFAs can be made robust to atypical observations by formulating the probabilistic model in terms of the Student-*t* distribution.

Hence, the aim of this work is to show that inherent robustness (with respect to atypical observations) can be achieved in the class of generative latent variable models that provide locally linear approximations to implicit low-dimensional data manifolds. Other important questions, not discussed in this work, are model selection (i.e., the optimal number of mixture components) and the automatic identification of the optimal dimensionality of these manifolds. One possibility is to use cross-validation or bootstrap techniques [7]. However, these approaches are computationally intensive and they are only feasible when the number of hyperparameters is relatively small. Alternatively, (hierarchical) Bayesian techniques can be envisioned [4,16].

This paper is organized as follows. In Section 2 robust PCA is introduced and in Section 3 the corresponding mixture model is derived. ML estimates of the parameters are computed by means of the expectation–maximization (EM) algorithm [6]. The approach is validated in Section 4. Note that the EM algorithm for mixtures of robust PFAs is discussed in Appendix D.

## 2. Robust PPCA

PCA seeks a linear projection which maps a set of observations $\{\mathbf{y}_n\}_{n=1}^N$ to a set of lower dimensional latent (unobserved) vectors $\{\mathbf{x}_n\}_{n=1}^N$ such that the variance in the projection space is maximized [14]. The latent variable model can be formalized as follows:

$$\mathbf{y}_n = \mathbf{W}\mathbf{x}_n + \boldsymbol{\mu} + \boldsymbol{\varepsilon}_n, \tag{1}$$

where $\mathbf{y}_n \in \mathbb{R}^D$ and $\mathbf{x}_n \in \mathbb{R}^d$, with $D > d$. The matrix $\mathbf{W} \in \mathbb{R}^{D \times d}$ is the (transposed) orthogonal projection matrix. The data offset and the projection errors are denoted by $\boldsymbol{\mu}$ and $\{\varepsilon_n\}_{n=1}^N$, respectively. In PPCA [20,26], it is further assumed that the error terms, as well as the prior uncertainty, are drawn from zero mean isotropic Gaussian[2] densities. Tipping and Bishop [26] showed that ML leads to a solution that is equivalent to PCA up to a rotation in the projection space. The columns of the ML estimate of $\mathbf{W}$ span the same subspace as the $d$ principal eigenvectors of the sample covariance matrix and the ML estimate of the noise variance $\tau^{-1}$ is equal to the average lost variance (eigenvalues) in the discarded directions. As discussed in Appendix A, the rotational ambiguity can be ignored in the context of mixture modelling, unless we are explicitly interested in local principal directions.

However, PPCA (as well as its non-probabilistic counterpart) suffers from the fact that it is based on Gaussian noise model. As a result, it is very sensitive to atypical observations such as outliers, and more generally, to situations where the data are not well confined on the low-dimensional clusters. Unfortunately, such cases occur quite often in practice which motivates the approach proposed in the following.

### 2.1. Latent variable view of the Student-*t* distribution

Compared to the Gaussian density, the Student-*t* density has an additional parameter, called the number of degrees of freedom $v$. It regulates the thickness of the distribution tails and therefore reduces the sensitivity to atypical observations. In this work, we do not restrict $v$ to be an integer value.

As noted in [15], the ML estimates of the parameters of the Student-*t* density can be computed by an EM algorithm by viewing the density as the following latent variable model:

$$\begin{aligned}\mathcal{S}(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Lambda},v) &= \int_0^\infty \mathcal{N}(\mathbf{y}|\boldsymbol{\mu},u\boldsymbol{\Lambda})\mathcal{G}\left(u\middle|\frac{v}{2},\frac{v}{2}\right)\mathrm{d}u \\ &= \langle \mathcal{N}(\mathbf{y}|\boldsymbol{\mu},u\boldsymbol{\Lambda})\rangle_{u|v}, \quad v > 0,\end{aligned} \tag{2}$$

where $\langle\cdot\rangle_u$ denotes the expectation with respect to the latent (or unobserved) scale variable $u$, over which we marginalize and on which a gamma[3] prior is imposed. Hence, the Student-*t* density can be reformulated as an infinite mixture

[2] The multivariate Gaussian density with mean $\boldsymbol{\mu}$ and inverse covariance matrix (or precision) $\boldsymbol{\Lambda}$ is defined as $\mathcal{N}(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Lambda}) \propto |\boldsymbol{\Lambda}|^{1/2}\,\mathrm{e}^{-(1/2)(\mathbf{y}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}(\mathbf{y}-\boldsymbol{\mu})}$.

[3] The gamma density is defined as $\mathcal{G}(u|\alpha,\beta) \propto u^{\alpha-1}\mathrm{e}^{-\beta u}$ with $\alpha > 0$ and $\beta > 0$.

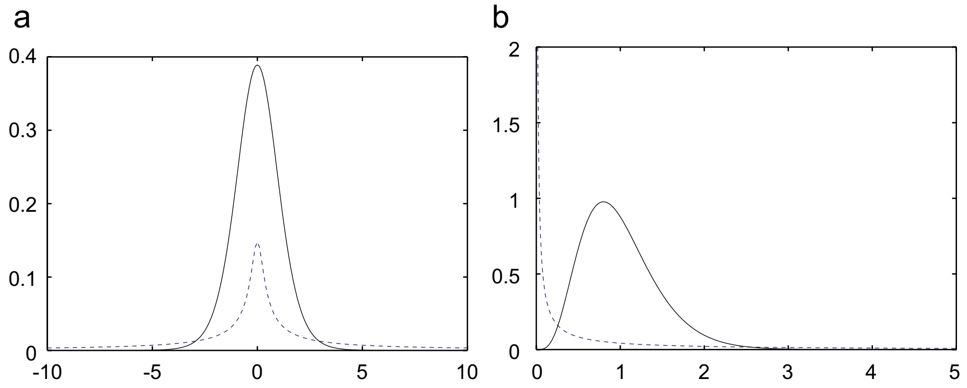


Fig. 1. (a) Univariate Student-$t$ density for $\nu = 10$ (solid) and $\nu = 0.1$ (dashed) and (b) the corresponding gamma prior on the latent scale variable.

of Gaussian densities with the same mean and where the prior on $u$ is a gamma density with parameters depending only on $\nu$. Examples of zero mean univariate Student-$t$ densities with unit variance and the corresponding gamma prior are shown in Fig. 1.

### 2.2. Robust reformulation of PPCA

As shown in [2], PPCA can be made robust by using a Student-$t$ model for the prior and the likelihood instead of a Gaussian one:

$$p(\mathbf{x}_n) = \langle \mathcal{N}(\mathbf{x}_n|\mathbf{0}, u_n\mathbf{I}_d)\rangle_{u_n|\nu} = \mathcal{S}(\mathbf{x}_n|\mathbf{0}, \mathbf{I}_d, \nu), \tag{3}$$

$$\begin{aligned} p(\mathbf{y}_n|\mathbf{x}_n) &= \langle \mathcal{N}(\mathbf{y}|\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu}, u_n\tau\mathbf{I}_D)\rangle_{u_n|\nu} \\ &= \mathcal{S}(\mathbf{y}_n|\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu}, \tau\mathbf{I}_D, \nu), \end{aligned} \tag{4}$$

where $\tau$ is the inverse residual variance, that is $\tau^{-1}$ accounts for the variance not captured by the low-dimensional latent vectors. Note that the scaled (inverse) covariance is data dependent; a different scale variable $u_n$ is assigned to each data point $\mathbf{y}_n$. Furthermore, the gamma prior on the scale variables is shared by the latent vectors and the observations, such that the robustness in the latent space and the measurement space is determined by a single $\nu$. Thus, when a data point is considered to be an outlier in the high-dimensional space, it does not contribute to the identification of the principal subspace as it is also considered to be an outlier in the projection space.

## 3. Mixtures of robust PPCAs

A mixture of $M$ robust probabilistic principal component analyzers is defined as follows:

$$p(\mathbf{y}) = \sum_{m=1}^{M} \pi_m p(\mathbf{y}|\boldsymbol{\theta}_m), \tag{5}$$

where $\{\boldsymbol{\theta}_m\}_{m=1}^M$ is the set of component parameters and $\{\pi_m\}_{m=1}^M$ is the set of mixture proportions, with $\sum_m \pi_m = 1$ and $\pi_m \geqslant 0$ for all $m$. The marginal likelihood $p(\mathbf{y}_n|\boldsymbol{\theta}_m)$ associated to the observation $\mathbf{y}_n$ is defined as a single robust PPCA according to (3) and (4):

$$\begin{aligned} p(\mathbf{y}_n|\boldsymbol{\theta}_m) &= \int_0^{+\infty}\int_{-\infty}^{+\infty} \mathcal{N}(\mathbf{y}_n|\mathbf{W}_m\mathbf{x}_{nm} + \boldsymbol{\mu}_m, u_n\tau_m\mathbf{I}_D) \\ &\qquad \times \mathcal{N}(\mathbf{x}_{nm}|\mathbf{0}, u_n\mathbf{I}_d)\mathcal{G}\left(u_n\middle|\frac{\nu_m}{2}, \frac{\nu_m}{2}\right) \mathrm{d}\mathbf{x}_{nm}\,\mathrm{d}u_n \\ &= \mathcal{S}(\mathbf{y}_n|\boldsymbol{\mu}_m, \mathbf{A}_m, \nu_m), \end{aligned} \tag{6}$$

where $\mathbf{A}_m^{-1} \equiv \mathbf{W}_m\mathbf{W}_m^\top + \tau_m^{-1}\mathbf{I}_D$. A set of low-dimensional latent variables $\{\mathbf{x}_{nm}\}_{n=1}^N$ and a set of latent scale variables $\{u_{nm}\}_{n=1}^N$ are associated to the $m$th robust PPCA model. For each observation $\mathbf{y}_n$, we also introduce the binary latent variable $\mathbf{z}_n$ indicating by which component $\mathbf{y}_n$ was generated. The resulting complete probabilistic model is defined as follows:

$$P(\mathbf{z}_n) = \prod_m \pi_m^{z_{nm}}, \tag{7}$$

$$p(\mathbf{u}_n|\mathbf{z}_n) = \prod_m \mathcal{G}\left(u_{nm}\middle|\frac{\nu_m}{2}, \frac{\nu_m}{2}\right)^{z_{nm}}, \tag{8}$$

$$p(\boldsymbol{\chi}_n|\mathbf{u}_n, \mathbf{z}_n) = \prod_m \mathcal{N}(\mathbf{x}_{nm}|\mathbf{0}, u_{nm}\mathbf{I}_d)^{z_{nm}}, \tag{9}$$

$$p(\mathbf{y}_n|\boldsymbol{\chi}_n, \mathbf{u}_n, \mathbf{z}_n) = \prod_m \mathcal{N}(\mathbf{y}_n|\mathbf{W}_m\mathbf{x}_{nm} + \boldsymbol{\mu}_m, u_{nm}\tau_m\mathbf{I}_D)^{z_{nm}}, \tag{10}$$

where $\mathbf{z}_n = (z_{n1}, \ldots, z_{nM})^\top$, $\mathbf{u}_n = (u_{n1}, \ldots, u_{nM})^\top$ and $\boldsymbol{\chi}_n = (\mathbf{x}_{n1}, \ldots, \mathbf{x}_{nM})^\top$. This probabilistic model can be represented by the graphical model shown in Fig. 2. Latent variables, indicated by unshaded nodes, are integrated out. Mathematically, this leads to (5) as desired.

### 3.1. Training algorithm

We seek ML estimates for the parameters $\boldsymbol{\theta} = \{\pi_m, \boldsymbol{\theta}_m\}_{m=1}^M$, with $\boldsymbol{\theta}_m \equiv \{\boldsymbol{\mu}_m, \mathbf{W}_m, \tau_m, \nu_m\}$ for all $m$. Unfortunately, the probabilistic formulation (7)–(10) of a mixture of robust principal component analyzers does not permit a direct maximization of the log-likelihood function $\ln\mathcal{L} = \sum_n \ln p(\mathbf{y}_n|\boldsymbol{\theta})$ as this quantity is intractable. Therefore, we adopt an EM approach [6], which finds ML parameters

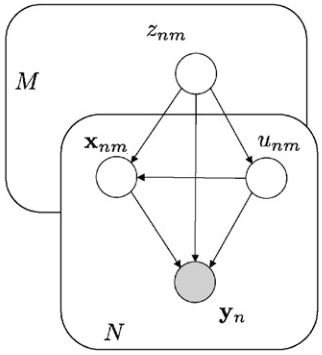


Fig. 2. Graphical model for the robust mixture of probabilistic principal component analyzers. A shaded node indicates that the random variable is observed, an arrow denotes a conditional dependency between the variables and a plate corresponds to a repetition.

estimates iteratively. First (E step), the posterior distribution of the latent variables is estimated for fixed parameters. Second (M step), the following quantity is maximized with respect to the parameters (see Appendix B for the explicit form):

$$\sum_n \langle \ln p(\mathbf{y}_n, \boldsymbol{\chi}_n, \mathbf{u}_n, \mathbf{z}_n|\boldsymbol{\theta})\rangle_{\boldsymbol{\chi}_n,\mathbf{u}_n,\mathbf{z}_n}. \tag{11}$$

The expectation is taken with respect to the joint posterior distribution of the latent variables, which is tractable (see Appendix C). The E step boils down to computing the expectations required in the M step:

$$\bar{\rho}_{nm} \equiv \langle z_{nm}\rangle = \frac{\pi_m \mathscr{S}(\mathbf{y}_n|\boldsymbol{\mu}_m, \mathbf{A}_m, v_m)}{\sum_k \pi_k \mathscr{S}(\mathbf{y}_n|\boldsymbol{\mu}_k, \mathbf{A}_k, v_k)}, \tag{12}$$

$$\bar{u}_{nm} \equiv \langle u_{nm}\rangle = \frac{D + v_m}{(\mathbf{y}_n - \boldsymbol{\mu}_n)^\top \mathbf{A}_m (\mathbf{y}_n - \boldsymbol{\mu}_m) + v_m}, \tag{13}$$

$$\ln \tilde{u}_{nm} \equiv \langle \ln u_{nm}\rangle = \psi\left(\frac{D+v_m}{2}\right) - \ln\left(\frac{(\mathbf{y}_n - \boldsymbol{\mu}_m)^\top \mathbf{A}_m (\mathbf{y}_n - \boldsymbol{\mu}_m) + v_m}{2}\right), \tag{14}$$

$$\bar{\mathbf{x}}_{nm} \equiv \langle \mathbf{x}_{nm}\rangle = \tau_m \mathbf{B}_m^{-1} \mathbf{W}_m^\top (\mathbf{y}_n - \boldsymbol{\mu}_m), \tag{15}$$

$$\bar{\mathbf{S}}_{nm} \equiv \langle z_{nm} u_{nm} \mathbf{x}_{nm} \mathbf{x}_{nm}^\top\rangle = \bar{\rho}_{nm} \mathbf{B}_m^{-1} + \bar{\omega}_{nm} \bar{\mathbf{x}}_{nm} \bar{\mathbf{x}}_{nm}^\top, \tag{16}$$

where $\bar{\omega}_{nm} \equiv \bar{\rho}_{nm}\bar{u}_{nm}$, $\mathbf{B}_m \equiv \tau_m \mathbf{W}_m^\top \mathbf{W}_m + \mathbf{I}_d$ and $\psi(\cdot) \equiv \Gamma'(\cdot)/\Gamma(\cdot)$ is the digamma function.

Maximizing (11) leads then to the following M step for the parameters:

$$\pi_m \leftarrow \frac{1}{N}\sum_n \bar{\rho}_{nm}, \tag{17}$$

$$\boldsymbol{\mu}_m \leftarrow \frac{\sum_n \bar{\omega}_{nm}(\mathbf{y}_n - \mathbf{W}_m \bar{\mathbf{x}}_{nm})}{\sum_n \bar{\omega}_{nm}}, \tag{18}$$

$$\mathbf{W}_m \leftarrow \left(\sum_n \bar{\omega}_{nm}(\mathbf{y}_n - \boldsymbol{\mu}_m)\bar{\mathbf{x}}_{nm}^\top\right)\left(\sum_n \bar{\mathbf{S}}_{nm}\right)^{-1}, \tag{19}$$

$$\tau_m^{-1} \leftarrow \sum_n \frac{\bar{\omega}_{nm}}{DN\pi_m}(\|\mathbf{y}_n - \boldsymbol{\mu}_m\|^2 - 2(\mathbf{y}_n - \boldsymbol{\mu}_m)^\top \mathbf{W}_m \bar{\mathbf{x}}_{nm}) + \frac{1}{DN\pi_m}\sum_n \mathrm{tr}\{\mathbf{W}_m \bar{\mathbf{S}}_{nm} \mathbf{W}_m^\top\}, \tag{20}$$

for all $m$. The contribution of each data point is weighted according to $\bar{\omega}_{nm}$, which accounts for both the effect of the responsibilities $\bar{\rho}_{nm}$ and the expected latent scale variables $\bar{u}_{nm}$. The latter ensures robustness as its value is small for $\mathbf{y}_n$ lying far from $\boldsymbol{\mu}_m$.

For the parameters $\{v_m\}_{m=1}^M$ there is no closed form ML estimate. Nevertheless, an ML solution can be computed at each EM iteration by solving the following expression by a line search algorithm (see for example [18]):

$$1 + \ln\left(\frac{v_m}{2}\right) - \psi\left(\frac{v_m}{2}\right) + \frac{1}{N\pi_m}\sum_n \bar{\rho}_{nm}\{\ln \tilde{u}_{nm} - \bar{u}_{nm}\} = 0, \tag{21}$$

for all $m$. Alternatively, a heuristic was proposed by Shoham [24] in the context of mixture modelling. This heuristic is also applicable here.

Since a line search is computationally inexpensive, the computational complexity of each EM step is $\mathcal{O}(MNDd)$. Hence, the overall complexity for mixtures of robust PPCAs is the same as for mixtures of ordinary PPCAs [27].

### 3.2. Low-rank approximation of the component covariances

Using a (latent) low-dimensional representation of the data has a clear advantage over a standard mixture of Gaussians (or Student-$t$'s) when considering a clustering problem or when estimating a density. Indeed, the number of parameters to estimate the covariance of each component is equal to $Dd_m + 1 - d_m(d_m - 1)/2$ (where the last term takes the rotational invariance into account) in the case of robust PPCAs and it is equal to $D(D+1)/2$ in the case of a standard mixture. The interesting feature of our approach is that the correlations between the principal directions are not neglected since $\mathbf{W}_m$ contains the local $d_m$ principal directions in the data. By contrast, it is common practice to force the covariance matrices to be diagonal (and thus axis aligned) in order to avoid numerical instabilities. This heuristic is clearly suboptimal.

### 3.3. Mixtures of robust PFAs

As mentioned earlier, PPCA is closely related to PFA [22]. If we assume in (10) that the covariance of the noise model is a diagonal matrix, we obtain a mixture of robust PFAs. The columns of the matrix $\mathbf{W}_m$ are called the *local factor loadings*, and the components $p(\mathbf{y}_n|\boldsymbol{\theta}_m)$ are now given by

$$p(\mathbf{y}_n|\boldsymbol{\theta}_m) = \mathscr{S}(\mathbf{y}_n|\boldsymbol{\mu}_m, \mathbf{A}_m, v_m), \tag{22}$$

where $\mathbf{A}_m^{-1} \equiv \mathbf{W}_m\mathbf{W}_m^\top + \boldsymbol{\Psi}_m^{-1}$. The diagonal matrix $\boldsymbol{\Psi}_m$ contains the inverse variances (or inverse *uniquenesses*) of the (local) factors and it corresponds to the precision of the conditional marginal $\mathscr{S}(\mathbf{y}_n|\mathbf{W}_m\mathbf{x}_{nm}+\boldsymbol{\mu}_m,\boldsymbol{\Psi}_m,v_m)$. The factors are thus independent given the latent variables. The EM algorithm for ML training is discussed in Appendix D.

## 4. Experiments

In this section, two types of experiments are considered. First, we illustrate how a low-dimensional nonlinear manifold spoiled by noisy data can still be recovered when using a robust approach. Second, the robust mixture modelling of high-dimensional data is demonstrated on two different data sets.

### 4.1. Robust reconstruction of low-dimensional manifolds

The following three-dimensional data set is considered:

$$y_{3n} = y_{1n}^2 + y_{2n}^2 - 1 + \varepsilon_n. \tag{23}$$

The data $\{y_{in} : i \in \{1,2\}\}_{n=1}^N$ are drawn from a uniform distribution in the $[-1,1]$ interval and the error terms $\{\varepsilon_n\}_{n=1}^N$ are distributed according to $\mathcal{N}(\varepsilon_n|0,\tau_\varepsilon)$, with $\tau_\varepsilon^{-1} = 0.01$. The data are located along a two-dimensional paraboloid; 500 training data were generated. The number of mixture components was fixed to 5 and $d_m$ was set to 2 (the true dimension of the manifold) for all $m$. Fig. 3 shows the results for a mixture of standard PPCAs and robust PPCAs in presence of 10% of outliers. These are drawn from a uniform distribution on the interval $[-1,1]$ in each direction. The shaded surfaces at the bottom of each plot indicate the regions associated to each component (or local linear model) after projection onto this two-dimensional surface. Each shaded region corresponds to a different component. The regions (data) are assigned to the component with highest responsibility. When the local models are nicely aligned with the manifold, the two-dimensional regions do not split. However, as shown in Fig. 3, only the mixture of robust PPCAs provides a satisfactory solution. Indeed, one of the components of the mixture of standard PPCAs is "lost" as it is used to model the outliers (and thus crosses the paraboloid).

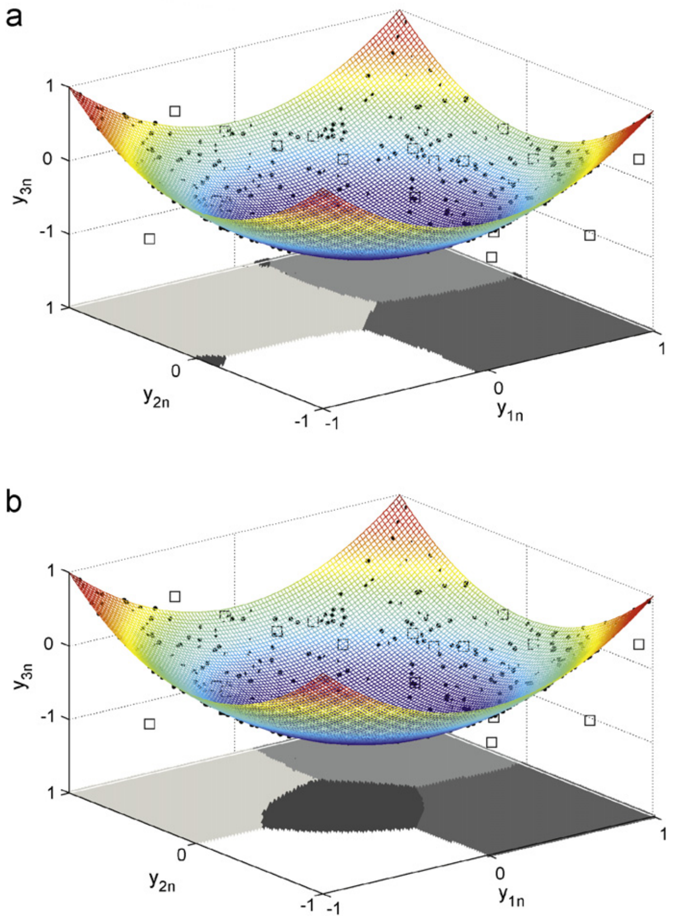


Fig. 3. Noisy paraboloid data set (black dots). Each two-dimensional shaded region is associated with a different local linear model (or component). They represent the dominant component membership of the points lying above it. The outliers are indicated by squares. (a) Projection by mixture of PPCAs. (b) Projection by mixture of robust PPCAs.

### 4.2. Analysis of high-dimensional clustering

First, we consider a three-dimensional synthetic example. Next, we discuss results on the high-dimensional USPS handwritten digit database.[4]

#### 4.2.1. Toy example

The data are grouped into three clusters (see Fig. 4). Each cluster corresponds to a three-dimensional Gaussian component with a diagonal covariance matrix equal to diag$\{5, 1, 0.2\}$ before rotation around the second coordinate axis. The two outer clusters are arranged at an angle of ±30° with respect to the middle one, which is horizontally aligned, and are, respectively, shifted by ±5 units along the axis of rotation. Hence, the data clusters lie essentially on a distorted two-dimensional plane.

The training data consist of 300 data points, of which 100 are drawn from each Gaussian component, and 5% of outliers. These are drawn from the uniform distribution in the hypercube centered on $\mathbf{0}$ and of side length equal to 20. The validation data consist of 900 data points (300 per cluster). The performance measure that we use to assess the mixture models (the standard mixture of Gaussians, the mixture of PPCAs and the mixture of robust PPCAs) is the log-likelihood of the validation data. Table 1 shows the results for 30 models trained on different training sets for $M = 3$ and 4. The dimension of the latent vectors is set to the same value for all components.

The overall best generalization on unseen data is obtained for the mixture of three robust PPCAs, with $d_m = 2$ for all $m$. The average validation log-likelihood is the highest and the standard deviation the smallest. Mixtures of standard PPCAs always perform poorer than their

[4] The USPS data were gathered at the Center of Excellence in Document Analysis and Recognition (CEDAR) at SUNY Buffalo during a project sponsored by the US Postal Service.

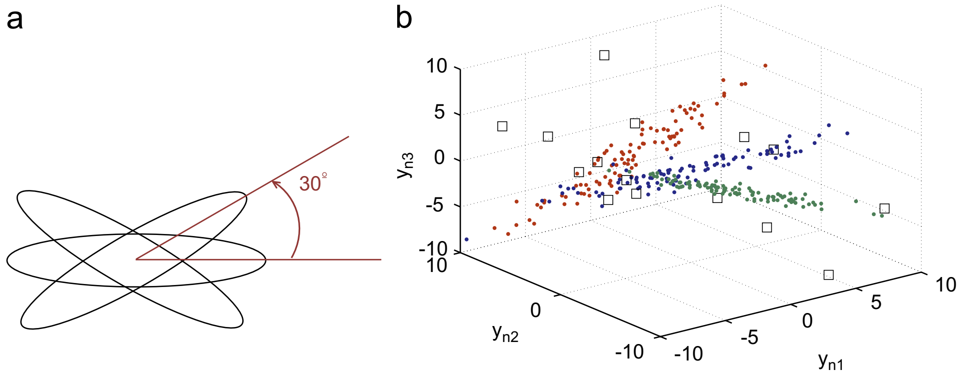


Fig. 4. Synthetic example for clustering with robust linear subspace models. (a) Vertical view of the mixture of Gaussians. (b) Example of a single training data set, with the outliers indicated by squares.

Table 1
Average log-likelihood of the validation data and the corresponding standard deviation

| | | $M=3$ | $M=4$ |
|---|---|---|---|
| $d=1$ | Mixt. of PPCAs | $-11.96\pm2.01$ | $-10.59\pm1.53$ |
| | Mixt. of robust PPCAs | $-10.26\pm0.68$ | $-10.42\pm0.41$ |
| $d=2$ | Mixt. of PPCAs | $-13.02\pm1.61$ | $-10.32\pm1.62$ |
| | Mixt. of robust PPCAs | $\mathbf{-9.34\pm0.26}$ | $-9.53\pm0.32$ |
| $d=3$ | Mixt. of Gaussians | $-13.27\pm1.86$ | $-10.33\pm1.88$ |

All numbers need to be multiplied by 1000. The training/validation process is repeated 30 times with different data sets.

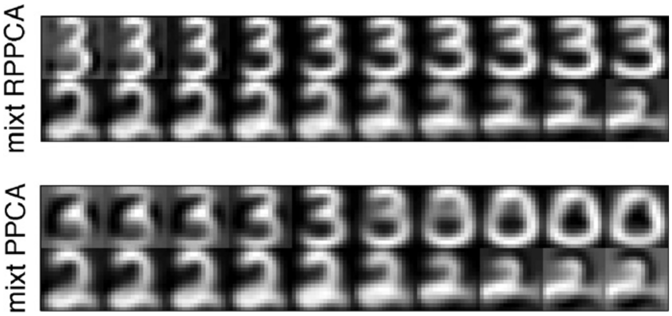


Fig. 5. Mixture of two component PPCAs with one-dimensional latent space to cluster USPS digit 2 and 3, and outliers digit 0. Top: robust PPCA. Bottom: standard PPCA.

robust counterpart. However, the former favor a one-dimensional latent space when $M=3$, while the latter favors a two-dimensional one. Interestingly, unconstrained mixtures of Gaussians perform well when the number of components is equal to 4. The reason for this is that the 4th component accounts for the outliers. This was also observed in [19]. Still, mixtures of robust PPCAs perform better. In fact, the two outer components are approximately Gaussian as $\nu$ is in the range of 10 for both of them. The remaining two components are centered on the origin, one being approximately Gaussian ($\nu\approx20$) and the other being heavy tailed ($\nu\approx1$). By contrast, when the number of components is set to 3, the middle component is approximately Gaussian ($\nu\approx10$) and the two outer components are heavy tailed ($\nu\approx2$).

Finally, it should be noted that the quality of the model provided by the mixtures of robust PPCAs is not affected by asymmetric noise. For example, when considering outliers only in the hypercube of side 10 and centered on $(2.5,2.5,2.5)^\top$, we obtain an average validation log-likelihood which is not significantly different ($-9.14\times10^3\pm0.24\times10^3$).

### 4.2.2. USPS handwritten digit data

In this experiment, the well-known USPS handwritten digit data are considered. The data are grayscale $16\times16$-pixels images of digits (0–9). To simplify the illustration, we kept only the images of digit 2 and digit 3 (respectively, 731 and 658 images), as well as 100 (randomly chosen) images of digit 0. In this setting, these are outliers as they stand outside the two main clusters of digit 2 and 3. We compared the mixture of PPCAs and the mixture of robust PPCAs in their ability to find the two main clusters assuming a one-dimensional latent space. The result is shown in Fig. 5. Each row represents images generated along the principal directions. The mixture of robust PPCAs completely ignores the outliers. The first component concentrates on the digits 3 and the second on the digits 2. Interestingly, the model is able to discover that the main variability of digits 3 is along their width, while the main variability of digits 2 is along their height. On the other hand, the mixture of PPCAs is very sensitive to the outliers as its first component makes the transition between digits 3 and outliers digits 0. This is undesirable in general as we prefer each component to stick to a single cluster. Of course, one could argue that three components would be a better choice in this case. However, we think that this example exploits a very common property of high-dimensional data, namely that the major mass of the density is confined in a low-dimensional subspace (or clusters of them), but not entirely. This experiment shows that the mixture of robust PPCAs is able to model such noisy manifolds, which are common in practice.

## 5. Conclusion

PCA and factor analysis are elementary and fundamental tools for exploratory data mining and data visualization. When tackling real-life problems, such as digit recognition, it is essential to take a robust approach. Here, the term "robust" is used to indicate that the performance of the methods is not spoiled by non-Gaussian noise (e.g., outliers). This property is obtained by exploiting the adaptive distribution tails of the Student-*t*. In this paper, mixtures of robust probabilistic principal component/factor analyzers were introduced. They provide a practical approach for discovering nonlinear relationships in the data by combining robust local linear models. More generally, they are suitable for robust clustering, while performing dimensionality reduction at the same time, and for the visualization of noisy high-dimensional data.

## Appendix A. On the rotational ambiguity of the projection matrix

The ML estimate of the PPCA projection matrix has the following form [26]:

$$\mathbf{W}_{\mathrm{ML}} = \mathbf{U}_d(\mathbf{P}_d - \tau^{-1}\mathbf{I}_d)^{1/2}\mathbf{R}, \tag{A.1}$$

where the columns of $\mathbf{U}_d \in \mathbb{R}^{D\times d}$ are the eigenvectors of the sample covariance matrix corresponding to the $d$ largest eigenvectors, $\mathbf{P}_d \in \mathbb{R}^{d\times d}$ is the diagonal matrix of these eigenvectors and $\mathbf{R} \in \mathbb{R}^{d\times d}$ is an orthogonal matrix.

From (A.1) it is clear that $\mathbf{W}_{\mathrm{ML}} \in \mathbb{R}^{D\times d}$ spans the same subspace as PCA and that the (scaled) principal directions are found up to a rotation $\mathbf{R}$. As noted in [26], this rotational ambiguity can easily be removed by a post-processing step. However, this step can be ignored in the context of mixture modelling since the mixture components are the marginals $\{p(\mathbf{y}|\boldsymbol{\theta}_m)\}_{m=1}^M$. These densities depend on the inverse covariance matrices $\{\mathbf{A}_m\}_{m=1}^M$ (see (6)), which are independent of $\{\mathbf{R}_m\}_{m=1}^M$ since $\mathbf{R}_m\mathbf{R}_m^\top = \mathbf{I}_d$ for all $m$.

## Appendix B. Variational lower bound to the log-likelihood

When computing ML estimates of the parameters by the EM algorithm, the log-likelihood is maximized iteratively by maximizing a lower bound, which is the variational negative free energy [17]:

$$-\mathscr{F}(q,\boldsymbol{\theta}) = \sum_n \langle \ln p(\mathbf{y}_n, \boldsymbol{\chi}_n, \mathbf{u}_n, \mathbf{z}_n|\boldsymbol{\theta})\rangle + \mathrm{H}[q], \tag{B.1}$$

with

$$\ln \mathscr{L} = -\mathscr{F}(q,\boldsymbol{\theta}) + \mathrm{KL}[q\|p] \geqslant -\mathscr{F}(q,\theta). \tag{B.2}$$

The variational distribution $q$ approximates the posterior of the latent variables given the parameters $\boldsymbol{\theta}$. At each iteration, the bound decreases monotonically, which provides a sanity check for the training algorithm.

For model (7)–(10) and the exact posteriors (see Appendix C), the negative free energy is given by

$$\begin{aligned}
-\mathscr{F}(q,\boldsymbol{\theta}) = {} & \sum_n\sum_m \bar{\rho}_{nm}\Big\{\ln\pi_m + \frac{v_m}{2}\ln\frac{v_m}{2} \\
& - \ln\Gamma\Big(\frac{v_m}{2}\Big) + \Big(\frac{v_m}{2} - 1\Big)\ln\tilde{u}_{nm}\Big\} \\
& - \sum_n\sum_m \frac{\bar{\omega}_{nm}v_m}{2} - \frac{NMD}{2}\ln 2\pi \\
& + \sum_n\sum_m \frac{\bar{\rho}_{nm}D}{2}\ln\tilde{u}_{nm} - \sum_n\sum_m \frac{1}{2}\,\mathrm{tr}\{\bar{\mathbf{S}}_{nm}\} \\
& + \sum_n\sum_m \frac{\bar{\rho}_{nm}D}{2}\ln\tau_m \\
& - \sum_n\sum_m \frac{\bar{\omega}_{nm}\tau_m}{2}\|\mathbf{y}_n - \boldsymbol{\mu}_m\|^2 \\
& + \sum_n\sum_m \bar{\omega}_{nm}\tau_m(\mathbf{y}_n - \boldsymbol{\mu}_m)^\top\mathbf{W}_m\bar{\mathbf{x}}_{nm} \\
& - \sum_n\sum_m \tau_m\,\mathrm{tr}\{\mathbf{W}_m\bar{\mathbf{S}}_{nm}\mathbf{W}_m^\top\} \\
& - \sum_n\sum_m \bar{\rho}_{nm}\{\ln\bar{\rho}_{nm} + \alpha_m\ln\beta_{nm} - \ln\Gamma(\alpha_m) \\
& + (\alpha_m - 1)\ln\tilde{u}_{nm}\} + \sum_n\sum_m \bar{\omega}_{nm}\beta_{nm} \\
& + \sum_m \frac{Nd_m}{2} - \sum_n\sum_m \frac{\bar{\rho}_{nm}}{2}\ln|\mathbf{B}_m|,
\end{aligned} \tag{B.3}$$

where $\alpha_m \equiv (D + v_m)/2$, $\beta_{nm} \equiv ((\mathbf{y}_n - \boldsymbol{\mu}_m)^\top\mathbf{A}_m(\mathbf{y}_n - \boldsymbol{\mu}_m) + v_m)/2$ and $\bar{\omega}_{nm} = \bar{\rho}_{nm}\bar{u}_{nm}$. The special quantities $\bar{\rho}_{nm}$, $\bar{u}_{nm}$, $\tilde{u}_{nm}$, $\bar{\mathbf{x}}_{nm}$ and $\bar{\mathbf{S}}_{nm}$ are, respectively, defined in (12)–(16).

## Appendix C. Posterior distributions of the latent variables

The posterior distributions of the latent variables are computed by applying the Bayes rule. Here, they are all tractable.

The posterior probabilities of the indicator variables are given by

$$P(z_{nm} = 1|\mathbf{y}_n) = \frac{\pi_m\mathscr{S}(\mathbf{y}_n|\boldsymbol{\mu}_m, \mathbf{A}_m, v_m)}{\sum_n \pi_m\mathscr{S}(\mathbf{y}_n|\boldsymbol{\mu}_m, \mathbf{A}_m, v_m)}, \tag{C.1}$$

for all $n$ and $m$. This quantity is called the *responsibility*. It is the posterior probability that the observation $\mathbf{y}_n$ was generated by the component $m$.

The gamma distribution is conjugate to the Gaussian distribution. Therefore, the posteriors of the scale variables are again gamma distributions:

$$\begin{aligned}
& p(u_{nm}|\mathbf{y}_n, z_{nm} = 1) \\
& \quad \propto \mathscr{N}(\mathbf{y}_n|\boldsymbol{\mu}_m, u_{nm}\mathbf{A}_m)\mathscr{G}\Big(u_{nm}\Big|\frac{v_m}{2}, \frac{v_m}{2}\Big) \\
& \quad = \mathscr{G}\Big(u_{nm}\Big|\frac{D + v_m}{2}, \frac{(\mathbf{y}_n - \boldsymbol{\mu}_m)^\top\mathbf{A}_m(\mathbf{y}_n - \boldsymbol{\mu}_m) + v_m}{2}\Big),
\end{aligned} \tag{C.2}$$

for all $n$ and $m$.

The posterior distributions of the low-dimensional latent vectors are given by

$$\begin{aligned} p(\mathbf{x}_{nm}|\mathbf{y}_n, u_{nm}, z_{nm} = 1) \\ &\propto \mathcal{N}(\mathbf{y}_n|\mathbf{W}_m\mathbf{x}_{nm} + \boldsymbol{\mu}_m, u_{nm}\tau_m\mathbf{I}_D)\mathcal{N}(\mathbf{x}_{nm}|\mathbf{0}, u_{nm}\mathbf{I}_d) \\ &= \mathcal{N}(\mathbf{x}_{nm}|\tau_m\mathbf{B}_m^{-1}\mathbf{W}_m^\top(\mathbf{y}_n - \boldsymbol{\mu}_m), u_{nm}\mathbf{B}_m), \end{aligned} \tag{C.3}$$

for all $n$ and all $m$. The inverse covariance is defined as $\mathbf{B}_m = \tau_m\mathbf{W}_m^\top\mathbf{W}_m + \mathbf{I}_d$.

Finally, the joint posterior of the latent variables is given by

$$\begin{aligned} \prod_n p(\boldsymbol{\chi}_n, \mathbf{u}_n, \mathbf{z}_n|\mathbf{y}_n) = \prod_n \prod_m p(\mathbf{x}_{nm}|u_{nm}, z_{nm}, \mathbf{y}_n) \\ \times p(u_{nm}|z_{nm}, \mathbf{y}_n)p(z_{nm}|\mathbf{y}_n). \end{aligned} \tag{C.4}$$

## Appendix D. EM algorithm for robust mixtures of factor analyzers

ML estimates for the parameters of mixtures of PFAs can be computed by the EM algorithm. For the E step, (12)–(14), (16) still hold, but the updates for $\mathbf{x}_{nm}$ and $\mathbf{B}_m$ are given by

$$\bar{\mathbf{x}}_{nm} = \mathbf{B}_m^{-1}\mathbf{W}_m^\top\boldsymbol{\Psi}_m(\mathbf{y}_n - \boldsymbol{\mu}_m), \tag{D.1}$$

$$\mathbf{B}_m = \mathbf{W}_m^\top\boldsymbol{\Psi}_m\mathbf{W}_m + \mathbf{I}_d, \tag{D.2}$$

for all $n$ and $m$.

The M step is identical to robust PPCAs, except for the update of the diagonal precisions (inverse uniquenesses):

$$\begin{aligned} \boldsymbol{\Psi}_m^{-1} \leftarrow \text{diag}\Bigg\{\frac{1}{N\pi_m}\sum_n (\bar{\varpi}_{nm}(\mathbf{y}_n - \boldsymbol{\mu}_m)(\mathbf{y}_n - \boldsymbol{\mu}_m)^\top \\ - \mathbf{W}_m\bar{\mathbf{S}}_{nm}\mathbf{W}_m^\top)\Bigg\}, \end{aligned} \tag{D.3}$$

where diag$\{\cdot\}$ sets all the off-diagonal elements to zero.

## References

[1] C. Archambeau, Probabilistic models in noisy environments—and their application to a visual prosthesis for the blind, Ph.D. Thesis, Université catholique de Louvain, Louvain-la-Neuve, Belgium, 2005.

[2] C. Archambeau, N. Delannay, M. Verleysen, Robust probabilistic projections, in: W.W. Cohen, A. Moore (Eds.), 23rd International Conference on Machine Learning (ICML), ACM, New York, 2006, pp. 33–40.

[3] C. Archambeau, N. Delannay, M. Verleysen, Mixtures of robust probabilistic principal component analyzers, in: 15th European Symposium on Artificial Neural Networks (ESANN), 2007, pp. 229–234.

[4] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.

[5] T. Cox, M. Cox, Multidimensional Scaling, Chapman & Hall, London, 2001.

[6] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via EM algorithm, J. R. Stat. Soc. B 39 (1) (1977) 1–38.

[7] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, London, 1993.

[8] B.S. Everitt, An Introduction to Latent Variable Models, Chapman & Hall, London, 1983.

[9] R.A. Fisher, Applications of "Student's" distribution, Metron 5 (1925) 90–104.

[10] Z. Ghahramani, G.E. Hinton, The EM algorithm for mixtures of factor analyzers, Technical Report CRG-TR-96-1, Department of Computer Science, University of Toronto, 1996.

[11] G.E. Hinton, P. Dayan, M. Revow, Modeling the manifolds of images of handwritten digits, IEEE Trans. Neural Networks 8 (1997) 65–74.

[12] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417–441.

[13] K. Huang, Y. Ma, R. Vidal, Minimum effective dimension for mixtures of subspaces: a robust GPCA algorithm and its applications, in: 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2004, pp. 631–638.

[14] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 1986.

[15] C. Liu, D.B. Rubin, ML estimation of the *t* distribution using EM and its extensions, ECM and ECME, Stat. Sinica 5 (1995) 19–39.

[16] T.P. Minka, Automatic choice of dimensionality for PCA, in: T.K. Leen, T.G. Dietterich, V. Tresp (Eds.), Advances in Neural Information Processing Systems (NIPS), vol. 13, MIT Press, Cambridge, MA, 2001, pp. 598–604.

[17] R.M. Neal, G.E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in: M.I. Jordan (Ed.), Learning in Graphical Models, MIT Press, Cambridge, MA, 1998, pp. 355–368.

[18] J. Nocedal, S.J. Wright, Numerical Optimization, Springer, Berlin, 2000.

[19] D. Peel, G.J. McLachlan, Robust mixture modelling using the *t* distribution, Stat. Comput. 10 (2000) 339–348.

[20] S.T. Roweis, EM algorithms for PCA and SPCA, in: M.I. Jordan, M.J. Kearns, S.A. Solla (Eds.), Advances in Neural Information Processing Systems (NIPS), vol. 10, MIT Press, Cambridge, MA, 1998.

[21] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.

[22] D.B. Rubin, D.T. Thayer, EM algorithms for ML factor analysis, Psychometrika 47 (1) (1982) 69–76.

[23] J.W. Sammon Jr., A nonlinear mapping for data structure analysis, IEEE Trans. Comput. 18 (1969) 401–409.

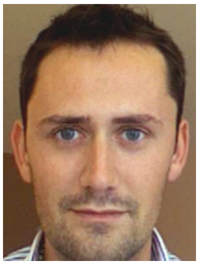
[24] S. Shoham, Robust clustering by deterministic agglomeration EM of mixtures of multivariate *t*-distributions, Pattern Recognition 35 (5) (2002) 1127–1142.

[25] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.

[26] M.E. Tipping, C.M. Bishop, Probabilistic principal component analysis, J. R. Stat. Soc. B 61 (1999) 611–622.

[27] M.E. Tipping, C.M. Bishop, Mixtures of probabilistic principal component analyzers, Neural Comput. 11 (2) (1999) 443–482.

[28] L. Xu, A.L. Yuille, Robust principal component analysis by self-organizing rules based on statistical physics approach, IEEE Trans. Neural Networks 6 (1) (1995) 131–143.



**Cédric Archambeau** received the Electrical Engineering degree and the Ph.D. in Applied Sciences from the Université catholique de Louvain, respectively, in 2001 and 2005. Until mid-2005, he was involved in an European biomedical research project, which aimed at developing a visual prosthesis for blind people. Since April 2006, he is a research associate at the University College London in the Centre for Computational Statistics and Machine Learning. His current research interests include approximate Bayesian inference, stochastic processes and dynamical systems, as well as clustering, classification and regression in very noisy environments.

**Nicolas Delannay** was born in 1980, Ottignies, Belgium. He graduated Master of electromechanical engineering in 2003 from Université catholique de Louvain (UCL), Belgium. He then received a research fellow grant from the FRS-FNRS to work on a Ph.D. thesis. The research took place from October 2003 to October 2007 within the Machine Learning Group at UCL (www.ucl.ac.be/mlg). The main topics covered by his thesis are: machine learning, data mining, Bayesian modelling, regression and collaborative filtering.

**Michel Verleysen** was born in 1965 in Belgium. He received the M.S. and Ph.D. degrees in electrical engineering from the Université catholique de Louvain (Belgium) in 1987 and 1992, respectively. He was an invited professor at the Swiss E.P.F.L. (Ecole Polytechnique Fédérale de Lausanne, Switzerland) in 1992, at the Université d'Evry Val d'Essonne (France) in 2001, and at the Université ParisI-Panthéon- Sorbonne from 2002 to 2007, respectively. He is now a professor at the Université catholique de Louvain, and Honorary Research Director of the Belgian F.N.R.S. (National Fund for Scientific Research). He is editor-in-chief of the Neural Processing Letters journal, chairman of the annual ESANN conference (European Symposium on Artificial Neural Networks), associate editor of the IEEE Trans. on Neural Networks journal, and member of the editorial board and program committee of several journals and conferences on neural networks and learning. He is author or co-author of more than 200 scientific papers in international journals and books or communications to conferences with reviewing committee. He is the co-author of the scientific popularization book on artificial neural networks in the series ''Que Sais-Je?,'' in French, and of the ''Nonlinear Dimensionality Reduction'' book published by Springer in 2007. His research interests include machine learning, artificial neural networks, self-organization, time-series forecasting, nonlinear statistics, adaptive signal processing, and high-dimensional data analysis.