Improving the Robustness to Outliers of Mixtures of Probabilistic PCAs

Nicolas Delannay[1], Cédric Archambeau[2], and Michel Verleysen[1]

[1] Université catholique de Louvain, Machine Learning Group - DICE
3 place du Levant, B-1348 Louvain-la-Neuve, Belgium
michel.verleysen@uclouvain.be
[2] Centre for Comput. Statistics and Machine Learning, University College London
Gower Street, London WC1E 6BT, U.K.
c.archambeau@cs.ucl.ac.uk

Abstract. Principal Component Analysis, when formulated as a probabilistic model, can be made robust to outliers by using a Student-t assumption on the noise distribution instead of a Gaussian one. On the other hand, mixtures of PCA is a model aimed to discover nonlinear dependencies in data by finding clusters and identifying local linear submanifolds. This paper shows how mixtures of PCA can be made robust to outliers too. Using a hierarchical probabilistic model, parameters are set by likelihood maximization. The method is shown to be effectively robust to outliers, even in the context of high-dimensional data.

1 Introduction

Principal Component Analysis (PCA) is a well-known data analysis and visualization tool. It provides a simple, algebraic way to choose axes in the data space that most fit the data, i.e. that maximize the variance after projection on the subspace spanned by these axes, or alternatively that minimize the projection error. A lower-dimensional representation of data is obtained by selecting a restricted number of the principal axes. However, maximal variance and minimal projection error are quadratic measures: a few outliers may dramatically influence the direction of principal axes, especially in high-dimensional spaces.

Probabilistic PCA [10,13] is a way to formalize the PCA problem as a latent variable model into a probabilistic framework. One of the nice features of the probabilistic framework is that non-traditional assumptions can easily be added to the model, the only price to pay being that the optimization of the model may reveal more difficult. For example, the traditional Gaussian noise hypothesis leads to the above detailed quadratic measures of errors and variances; replacing this hypothesis by, for instance, a Student-t noise distribution leads to a robust version of PCA [2]. In contrast to other robust approaches to PCA which usually require to optimize several additional parameters, the probabilistic formalism only requires to choose the dimension of the projection space, the other parameters being set automatically by maximum likelihood (ML). Another advantage is that the probabilistic model provides likelihood measures, which can be used to compute posterior probabilities and eventually to construct a Bayes classifier.

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 527–535, 2008.

Mixtures of (local) PCA may be used to uncover nonlinear manifolds in data, and are also nicely formalized into a probabilistic framework [12]. The principle is to attribute each observed data to a specific (unknown) local model (or component), through an indicator variable, and then to mix the local models. An expectation-maximization algorithm can be used to set the parameters of the model, including these indicator variables. An advantage of mixtures of PCA, compared to other mixtures models (a.o. Gaussian mixtures), is that the full-rank, possibly ill-conditioned covariance matrices are approximated by low-rank covariance matrices, without having to neglect the correlations between the (local) principal directions to avoid numerical instabilities. The other way to avoid ill-conditioned covariance matrices is to constrain them to be diagonal, leading to suboptimal axis-aligned components [1]. Besides nonlinear manifold uncovering, mixtures models can be used in a straighforward way for clustering, and probability density estimation. In both cases the same limitations related to ill-conditioned covariance matrices apply though.

Mixtures of probabilistic PCA [3] can be made robust to atypical observations by using a Student-t noise distribution hypothesis. This paper shows the complete probabilistic learning procedure for this model. It is shown that all parameters (with the exception of the number of components and their dimensionality) may be easily optimized by an Expectation-Maximization procedure, without additional complexity with respect to the non-robust version.

The following of this paper is organized as follows. The next section first reminds the Probabilistic PCA model and its robust extension, and then introduces the Mixtures of Robust Probabilistic PCA model. Section 3 details how the parameters of the model may be optimized, and Section 4 illustrates the robustness of the model to atypical observations.

2 Robust Probabilistic PCA and Mixtures

PCA can be formulated as the search for an optimal linear projection minimizing a reconstruction error. The principal components are derived from the observations by projecting them on the principal directions. In the probabilistic formulation, the view is inverted in the sense that the observations $\{\mathbf{y}_n\}_{n=1}^{N}$ where $\mathbf{y}_n \in \mathbb{R}^D$, are assumed to be generated from a low dimension latent representation $\{\mathbf{x}_n\}_{n=1}^{N}$, where $\mathbf{x}_n \in \mathbb{R}^J$, $J < D$.

The principle of probabilistic modeling is to express the uncertainty about (some of) the parameters of the model by prior distributions. Probabilistic PCA (PPCA) was proposed in [10,13]; Gaussian priors are used in PPCA. Maximising the likelihood of the observations in PPCA leads to principal axes that are equivalent to the principal axes found by the standard PCA, up to a rotation and a scaling [13]; the same subspace is thus spanned.

PCA and PPCA are sensitive to atypical observations and observations not well confined in a low-dimensional subspace, because of their quadratic criterion and Gaussian noise model respectively. The robust probabilistic PCA [2] extends PPCA to make it applicable on datasets containing atypical samples. Instead

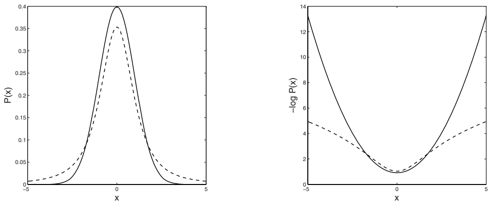


Fig. 1. Left: Probability density functions of a Gaussian $(-)$ and a Student-t with $\nu = 2$ $(--)$. Right: Negative log-likelihood of these same distributions.

of the Gaussian noise assumption, the randomness in observations is modeled by a Student-t distribution with an additional parameter ν (called the *number of degrees of freedom*), which regulates the thickness of the distribution's tail. Figure 1(left) shows unit-variance Gaussian and Student-t distributions ($\nu = 2$). Figure 1(right) shows the corresponding negative-log-likelihood which appears in the training criterion of probabilistic models. We see that when ν is small, the Student-t attributes a much smaller cost than the Gaussian to points lying far from the mean. The sensitivity to atypical observations is therefore reduced.

PPCA makes the assumption that atypical samples might come either from the generation of latent vectors $\mathbf{x}$ or from the noise contribution. This is expressed by Student-t distributions on the prior of the latent vectors and on the conditional distribution of observations: $P(\mathbf{x}) = \mathcal{S}t(\mathbf{x}|\mathbf{0}, \mathbf{I}_J, \nu)$, $P(\mathbf{y}|\mathbf{x}) = \mathcal{S}t(\mathbf{y}|\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \tau^{-1}\mathbf{I}_D, \nu)$. Note that in the traditional PPCA model, the Student-t distributions are replaced by Gaussian ones. To simplify the parameterization, both distributions are attributed the same degree of freedom ν. This choice will be commented below. The Student-t distribution can be reformulated as an infinite mixture of Gaussian distributions $\mathcal{S}t(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \int_0^\infty \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \frac{1}{u}\boldsymbol{\Sigma})\ \mathcal{G}a(u|\frac{\nu}{2}, \frac{\nu}{2})$ du, $\nu > 0$, where $\mathcal{G}a(u|\cdot, \cdot)$ is a Gamma distribution over the precision factor u. Making use of this factorization, the generative model can be represented with an additional level in the hierarchy where the latent precision u appears:

$$P(u) = \mathcal{G}a(u|\tfrac{\nu}{2}, \tfrac{\nu}{2}) \tag{1}$$

$$P(\mathbf{x}|u) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \frac{1}{u}\mathbf{I}_J) \tag{2}$$

$$P(\mathbf{y}|\mathbf{x}, u) = \mathcal{N}(\mathbf{y}|\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \frac{1}{u\tau}\mathbf{I}_D)\ . \tag{3}$$

From this generative formulation, we see that the uncertainty about the observation (i.e. expressed by the variance in (3)) can be amplified by a small latent precision variable u, shared by the $\mathbf{x}$ and $\mathbf{y}$ conditional distributions. According to intuition, this constraint implies that outliers $\mathbf{y}$ in the observation space are also considered as outliers $\mathbf{x}$ in the latent space so their contributions to the identification of the latent space are down-weighted.

For robust PPCA, the marginal distribution of the observations is tractable: $P(\mathbf{y}) = \int_0^\infty \int_{\mathcal{X}} P(\mathbf{y}|\mathbf{x}, u)\, P(\mathbf{x}|u)\, P(u)\, d\mathbf{x}\ = \mathcal{S}t(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ where $\boldsymbol{\Sigma} \equiv \mathbf{W}\mathbf{W}^\top +$

$\tau^{-1}\mathbf{I}_D$. The training procedure consists in maximizing this (marginal) likelihood with respect to $\boldsymbol{\theta} \equiv (\mathbf{W}, \boldsymbol{\mu}, \tau, \nu)$.

In contrast with previous robust approaches to the PCA (see for example [15] and [7], and the references therein), this probabilistic formalism only requires to select the dimension of the projection space (see Section 3), the other parameters being estimated by the maximum likelihood criterion.

Even in its robust and probabilistic versions, PCA is not adequate for representing clusters or nonlinear dependencies in the data. The mixture of PPCA [12] may solve this problem, but is again too sensitive to atypical samples limiting its use on many real world datasets. It is thus natural to look for a robust formulation of the mixture of PPCA.

The probability distribution of a sample generated from a mixture of K robust PPCA is defined as $P(\mathbf{y}) = \sum_k \pi_k P_k(\mathbf{y})$ where $\{\pi_k\}_{k=1}^K$ is the set of positive mixture proportions, with $\sum_k \pi_k = 1$; the $P_k(\mathbf{y})$ are defined as single robust PPCA components $P_k(\mathbf{y}) = \mathcal{S}t(\mathbf{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k)$ in which $\boldsymbol{\Sigma}_k \equiv \mathbf{W}_k\mathbf{W}_k^\top + \tau_k^{-1}\mathbf{I}_D$. The set of parameters of this model is $\boldsymbol{\theta} \equiv \{(\mathbf{W}_k, \boldsymbol{\mu}_k, \tau_k, \nu_k, \pi_k)\}_{k=1}^K$.

Using a latent indicator variable $\mathbf{z} = [z_1, \ldots, z_K]$ (with $z_k = 1$ if the kth component generated the observation $\mathbf{y}$, otherwise $z_k = 0$) simplifies the derivation of an EM algorithm. The factorized mixture of robust PPCA is then

$$P(\mathbf{z}) = \textstyle\prod_k \pi_k^{z_k} \quad , \tag{4}$$

$$P(\mathbf{u}|\mathbf{z}) = \textstyle\prod_k \mathcal{G}a(u_k|\frac{\nu_k}{2}, \frac{\nu_k}{2})^{z_k} \quad , \tag{5}$$

$$P(\boldsymbol{\chi}|\mathbf{u}, \mathbf{z}) = \textstyle\prod_k \mathcal{N}(\mathbf{x}_k|\mathbf{0}, \frac{1}{u_k}\mathbf{I}_J)^{z_k} \quad , \tag{6}$$

$$P(\mathbf{y}|\boldsymbol{\chi}, \mathbf{u}, \mathbf{z}) = \textstyle\prod_k \mathcal{N}(\mathbf{y}|\mathbf{W}_k\mathbf{x}_k + \boldsymbol{\mu}_k, \frac{1}{u_k\tau_k}\mathbf{I}_D)^{z_k} \quad , \tag{7}$$

where $\mathbf{u} = [u_1, \ldots, u_K]$ and $\boldsymbol{\chi} = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$; the different components could also have different latent dimensionalities $\{J_k\}_{k=1}^K$.

Increasing the robustness by replacing Gaussian densities with Student-t ones was also proposed for finite mixture models [8,1]. The main advantage of mixtures of PPCA resides in the fact that the full-rank, possibly ill-conditioned covariance matrices are approximated by constrained covariance matrices $\boldsymbol{\Sigma}_k$, strongly reducing the number of free parameters per component. By contrast, constraining the covariance to be diagonal leads to axis-aligned components which does not take the dominant correlations into account [1].

3 Learning Procedure

The factorization of the model (4)-(7) allows us to derive an exact Expectation-Maximization (EM) algorithm. Note that this algorithm encompasses the optimization of the (mixture of) probabilistic PCA: one only needs to add the constraint $\nu_k = \infty$ (for all k) such that the Student-ts are in fact Gaussian distributions.

We seek an optimum of the marginal distribution of the observations to estimate the parameters $\boldsymbol{\theta} \equiv \{(\mathbf{W}_k, \boldsymbol{\mu}_k, \tau_k, \nu_k, \pi_k)\}_{k=1}^K$. The simplest way to proceed is by deriving an EM algorithm [6] on the factorised distribution (4)-(7). The starting point of the algorithm is to bound the marginal likelihood (making use of the Jensen's inequality):

$$\log P(\{\mathbf{y}_n\}) \geq \quad \mathbb{E}_Q\left\{\log P(\{\mathbf{y}_n\},\{\boldsymbol{\chi}_n\},\{\mathbf{u}_n\},\{\mathbf{z}_n\})\right\} \\ -\mathbb{E}_Q\left\{\log Q\left(\{\boldsymbol{\chi}_n\},\{\mathbf{u}_n\},\{\mathbf{z}_n\}\right)\right\} \ . \tag{8}$$

Equation (8) is valid for any distribution Q. The bound is tight when the distribution over the latent variable $Q(\{\boldsymbol{\chi}_n\},\{\mathbf{u}_n\},\{\mathbf{z}_n\})$ coincides with the posterior distribution. Fortunately, the posterior distribution of the mixture of robust PPCA model is still tractable. Indeed, applying the Bayes formula, one can show that the posterior is

$$P(\{\boldsymbol{\chi}_n\},\{\mathbf{u}_n\},\{\mathbf{z}_n\}|\{\mathbf{y}_n\}) = \prod_n\prod_k P(\mathbf{x}_{nk}|u_{nk}, z_{nk}=1,\mathbf{y}_n) \\ \cdot P(u_{nk}|z_{nk}=1,\mathbf{y}_n)P(z_{nk}=1|\mathbf{y}_n) \tag{9}$$

where the factor distributions are

$$P(\mathbf{x}_{nk}|u_{nk}, z_{nk}=1,\mathbf{y}_n) = \mathcal{N}(\mathbf{x}_{nk}|\tau_k\mathbf{C}_k\mathbf{W}_k^\top(\mathbf{y}_n-\boldsymbol{\mu}_k),\frac{1}{u_{nk}}\mathbf{C}_k) \tag{10}$$

$$P(u_{nk}|z_{nk}=1,\mathbf{y}_n) = \mathcal{G}a(u_{nk}|\alpha_k,\beta_{nk}) \tag{11}$$

$$P(z_{nk}=1|\mathbf{y}_n) = \frac{\pi_k\mathcal{S}t(\mathbf{y}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,\nu_k)}{\sum_k\pi_k\mathcal{S}t(\mathbf{y}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,\nu_k)} \tag{12}$$

and where we have defined $\mathbf{C}_k^{-1} = \tau_k\mathbf{W}_k^\top\mathbf{W}_k + \mathbf{I}_J$, $\alpha_k = (D+\nu_k)/2$, and $\beta_{nk} = ((\mathbf{y}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_n-\boldsymbol{\mu}_k)+\nu_k)/2$. Notice that there is only a single observation $\mathbf{y}_n$ appearing in each of these posterior factor distributions.

The EM algorithm then consists in two successive and repeated steps. The *E-step* consists in fixing Q to the distribution given by (9) and developing (8) accordingly. Note that only the first term of (8) (called the log-complete likelihood) has to be computed, as the second one does not depend on the values of the parameters. This leads to a somewhat complex expression, not detailed here for simplicity. Its evaluation necessitates to compute the following expectations:

$$\bar{\rho}_{nk} \equiv \mathbb{E}_Q\{z_{nk}\} = \tfrac{\pi_k\mathcal{S}t(\mathbf{y}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,\nu_k)}{\sum_k\pi_k\mathcal{S}t(\mathbf{y}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,\nu_k)}, \tag{13}$$

$$\bar{u}_{nk} \equiv \mathbb{E}_Q\{u_{nk}\} = \tfrac{\alpha_k}{\beta_{nk}}, \tag{14}$$

$$\log\tilde{u}_{nk} \equiv \mathbb{E}_Q\{\log u_{nk}\} = \psi\left(\alpha_k\right) - \log\left(\beta_{nk}\right), \tag{15}$$

$$\bar{\mathbf{x}}_{nk} \equiv \mathbb{E}_Q\{\mathbf{x}_{nk}\} = \tau_k\mathbf{C}_k\mathbf{W}_k^\top(\mathbf{y}_n-\boldsymbol{\mu}_k), \tag{16}$$

$$\bar{\mathbf{S}}_{nk} \equiv \mathbb{E}_Q\{z_{nk}u_{nk}\mathbf{x}_{nk}\mathbf{x}_{nk}^\top\} = \bar{\rho}_{nk}\mathbf{C}_k + \bar{\omega}_{nk}\bar{\mathbf{x}}_{nk}\bar{\mathbf{x}}_{nk}^\top, \tag{17}$$

where $\bar{\omega}_{nk} \equiv \bar{\rho}_{nk}\bar{u}_{nk}$ and $\psi(\cdot) \equiv \Gamma'(\cdot)/\Gamma(\cdot)$ is called the *digamma* function.

The log-complete likelihood of course depends on the model parameters; the *M-step* then consists in maximizing it with respect to the parameters, leading to a set of update rules for all k (tr$\{\cdot\}$ is the trace operator):

$$\pi_k \leftarrow \tfrac{1}{N} \textstyle\sum_n \bar{\rho}_{nk} \tag{18}$$

$$\boldsymbol{\mu}_k \leftarrow \frac{\sum_n \bar{\omega}_{nk}(\mathbf{y}_n - \mathbf{W}_k \bar{\mathbf{x}}_{nk})}{\sum_n \bar{\omega}_{nk}} \tag{19}$$

$$\mathbf{W}_k \leftarrow \left(\textstyle\sum_n \bar{\omega}_{nk}(\mathbf{y}_n - \boldsymbol{\mu}_k)\bar{\mathbf{x}}_{nk}^\top\right)\left(\textstyle\sum_n \bar{\mathbf{S}}_{nk}\right)^{-1} \tag{20}$$

$$\tau_k^{-1} \leftarrow \tfrac{1}{DN\pi_k} \textstyle\sum_n \left(\bar{\omega}_{nk}\|\mathbf{y}_n - \boldsymbol{\mu}_k\|^2 - \mathrm{tr}\{\mathbf{W}_k \bar{\mathbf{S}}_{nk} \mathbf{W}_k^\top\}\right) \quad . \tag{21}$$

In these updates rules, the contribution of each data point is weighted according to $\bar{\omega}_{nk}$, which accounts for both the effect of the responsibilities $\bar{\rho}_{nk}$ and the expected latent precision variables $\bar{u}_{nk}$. The latter ensures robustness as its value is small for $\mathbf{y}_n$ lying far from $\boldsymbol{\mu}_k$, such that the contribution in the M-step is small. For the non robust formulation ($\nu_k \to \infty$) we have $\bar{u}_{nk} = 1$ for all n and all k. Note also that these updates are coupled: one could cycle through these updates between each E-step until the M-step has converged.

There is no closed form update for $\{\nu_k\}_{k=1}^K$. Nevertheless, a solution can be computed by line search at each EM iteration [2]. Alternatively, a heuristic was proposed by Shoham [11] in the context of mixture modeling.

As the marginal likelihood of mixture models has local optima, it is recommended to repeat the optimization with different initializations. A good strategy to initialize the components is to set the centers $\boldsymbol{\mu}_k$ with a quantization algorithm and initialize the subspace orientation $\mathbf{W}_k$ from the first Principal directions in the Voronoi region of $\boldsymbol{\mu}_k$.

Two *hyper-parameters* still need to be set: the number of components and the dimensionalities of the latent representations. They can be set in a traditional way by cross-validation, or added in a Bayesian way to the probabilistic formulation; in the latter case however MCMC sampling techniques [9] or (mean field) variational approximation [14,4] must be used instead of the exact EM algorithm. Finally Automatic Relevance Determination was used in [5] to select the dimensionality of latent subspaces.

4 Experiments

In this section, the (robust) probabilistic models are applied first on two artificial examples, and then on the USPS high-dimensional real dataset, using the software available from http://www.ucl.ac.be/mlg/.

Figures 2(a)-(b) show an example where samples have been generated along a one-dimensional manifold, with higher density in the right end and higher noise at the other end. The PPCA estimates a global principal direction; the mean of the component lies in an empty region and is thus not representative of typical samples. On the other hand, the robust PPCA discards samples in order to concentrate on the higher density region of the manifold. Using three components in the model (Figures 2(c)-(d)), both the mixture of PPCA and robust PPCA estimate quite well the local principal directions. However one of the components of the mixture of PPCA (Figure 2(c)) tries to account for the noisy samples, forcing its mean to move away from the manifold.

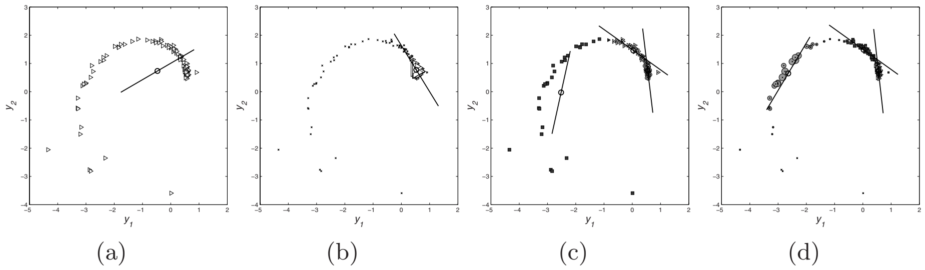


Fig. 2. Samples generated along a 1-dimensional manifold with additional atypical points. (a) Probabilistic PCA, (b) robust probabilistic PCA, (c) 3 components mixture of PPCA, (d) 3 components mixture of robust PPCA. The sizes of the markers represent their contribution to the estimation of the component parameters.

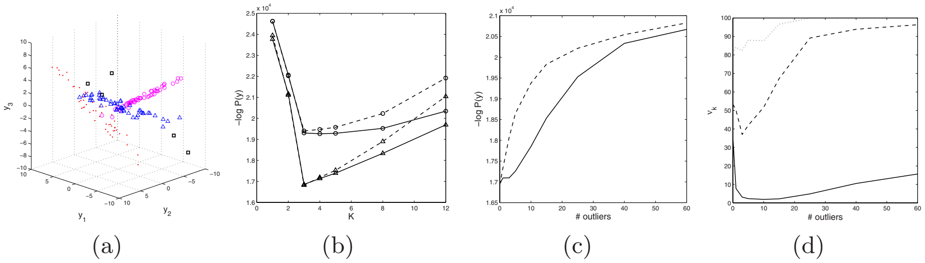
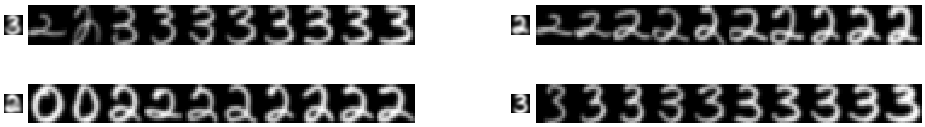


Fig. 3. (a) Synthetic example with 3 Gaussian clusters. The squares represent outliers. (b) Negative log likelihood of a validation set with respect to the number of components K and the dimensionality of the latent space (○: $J = 1$, △: $J = 2$). Dashed line: standard. Plain line: robust. (c) Negative log likelihood with respect to the number of outliers. (d) Degree of freedom parameters for the three components in the robust mixture model with respect to the number of outliers.

The next example consists in data arranged in three 3-dim. Gaussian clusters (see Figure 3(a)), with diagonal covariance matrices equal to diag$\{[5, 1, 0.2]\}$ before rotation around the second coordinate axis. Each component lies on an intrinsic two dimensional space as the variance in the third direction is significantly smaller. The two outer clusters make an angle of ±30 degrees with the middle one and are respectively shifted by ±5 units along the axis of rotation. For the first experiment, 30 data are generated for each cluster. The generalisation performances, measured as the log likelihood on a validation set averaged on 50 experiments, are plotted in Figure 3(b) for $K \in \{1, \ldots, 12\}$ components and $J \in \{1, 2\}$ latent space dimensions. As expected, the true model with $K = 3$ and $J = 2$ performs the best. Interestingly, we see that the standard and robust mixture models have comparable performances when the model underfits the data (i.e. $K < 2$) while the robust mixture has the edge when K increases. Overfitting is thus reduced with the robust formulation.



(a) Standard

(b) Robust

Fig. 4. Mixture of 2 component PPCAs with 1-dimensional latent space to cluster USPS digit 2 and 3, and outliers digit 0. (a) standard; (b) robust.

For the second experiment, K is set to 3 and J to 2 (their optimal values); we look at the sensitivity of the model to the number of outliers. The outliers are generated uniformly in the $[-10, 10]^3$ box. Again, 30 points are generated from each component; 1 to 60 outliers are added. The performances measured on a validation set without outliers, and averaged over 50 repetitions as above, are shown on Figure 3(c). Again, we see the increased robustness of the proposed model, in particular when there are few outliers. When the number of outliers increases to a significant proportion of the learning data the down-weighting of the outliers in the robust model is reduced, and the gap between the performances decreases. Figure 3(d) shows the average value of the degree of freedom parameters (ν_k for $k = 1 \ldots 3$). We note that the down-weighting of the outliers obtained with small value of ν_k, comes mainly from a single component.

The last example illustrates the robustness of the proposed method on high-dimensional data. The USPS handwritten digit dataset consists in 16×16 pixels images of digits (0 to 9). Only the (respectively 731 and 658) images of digits 2 and 3 are kept (they form the two dominant clusters), as well as 100 (randomly chosen) images of digit 0. We compare the mixtures of PPCAs and of robust PPCAs in their ability to find the two main clusters (thereby identifying the 0 as outliers) and to identify the main variability in these clusters with a one-dimensional latent space. Figure 4 shows sample images close to the one-dimensional subspace. The mixture of robust PPCAs completely ignores the smaller cluster of digits 0. On the other hand, the mixture of PPCAs cannot down-weight the contribution of the digits 0, influencing the two components.

5 Conclusion

This papers introduces the Mixtures of Robust Probabilistic PCA. The method is aimed to represent nonlinear manifolds and possibly identify clusters in data. All parameters of the method, with the exception of the number of clusters and the dimensionality of the latent space, are learned trough the use of a probabilistic latent formulation, and the optimization of the likelihood of the data. Compared to its non-robust parent, the method shows a strongly reduced sensitivity to outliers, even in high-dimensional spaces.

References

1. Archambeau, C.: Probabilistic Models in Noisy Environments - And their Application to a Visual Prosthesis for the Blind. PhD thesis, Université catholique de Louvain, Louvain-la-Neuve, Belgium (2005)
2. Archambeau, C., Delannay, N., Verleysen, M.: Robust probabilistic projections. In: Cohen, W.W., Moore, A. (eds.) 23rd International Conference on Machine Learning (ICML), pp. 33–40. ACM Press, New York (2006)
3. Archambeau, C., Delannay, N., Verleysen, M.: Mixtures of robust probabilistic principal component analysers. In: ESANN 2007, European Symposium on Artificial Neural Networks Advances in Computational Intelligence and Learning, Bruges, Belgium (2007)
4. Attias, H.: Inferring parameters and structure of latent variable models by variational bayes. In: Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-1999), pp. 21–30. Morgan Kaufmann, San Francisco (1999)
5. Bishop, C.M.: Bayesian pca. In: Proceedings of the 1998 conference on Advances in neural information processing systems II, pp. 382–388. MIT Press, Cambridge (1999)
6. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via EM algorithm. J. Royal Statistical Soc. B 39(1), 1–38 (1977)
7. Huang, K., Ma, Y., Vidal, R.: 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, pp. 631–638 (2004)
8. Peel, D., McLachlan, G.J.: Robust mixture modelling using the t distribution. Statistics and Computing 10, 339–348 (2000)
9. Richardson, S., Green, P.: On bayesian analysis of mixtures with an unknown number of components. J. Roy. Statist. Soc. 59, 731–792 (1996)
10. Roweis, S.T.: EM algorithms for PCA and SPCA. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (eds.) Advances in Neural Information Processing Systems 10 (NIPS). MIT Press, Cambridge (1998)
11. Shoham, S.: Robust clustering by deterministic agglomeration EM of mixtures of multivariate *t*-distributions. Pattern Recognition 35(5), 1127–1142 (2002)
12. Tipping, M.E., Bishop, C.M.: Mixtures of probabilistic principal component analyzers. Neural Computation 11(2), 443–482 (1999)
13. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society B 61, 611–622 (1999)
14. Waterhouse, S., MacKay, D., Robinson, T.: Bayesian methods for mixtures of experts. In: Touretzky, D.S., et al. (eds.) Advances in Neural Information Processing Systems, vol. 8, pp. 351–357. MIT Press, Cambridge (1996)
15. Xu, L., Yuille, A.L.: Robust principal component analysis by self-organizing rules based on statistical physics approach. IEEE Transactions on Neural Networks 6(1), 131–143 (1995)