Mixtures of Robust Probabilistic Principal Component Analyzers

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

1 - University College London, Dept. of Computer Science
Gower Street, London WC1E 6BT, United Kingdom
c.archambeau@cs.ucl.ac.uk

2 - Université catholique de Louvain, Machine Learning Group
3 Place du Levant, 1348 Louvain-la-Neuve, Belgium
nicolas.delannay@uclouvain.be, verleysen@dice.ucl.ac.be

Abstract. Discovering low-dimensional (nonlinear) manifolds is an important problem in Machine Learning. In many applications, the data are in a high dimensional space. This can be problematic in practice due to the curse of dimensionality. Fortunately, the core of the data lies often on one or several low-dimensional manifolds. A way to handle these is to pre-process the data by nonlinear data projection techniques. Another approach is to combine local linear models. In particular, mixtures of probabilistic principal component analyzers are very attractive as each component is specifically designed to extract the local principal orientations in the data. However, an important issue is the model sensitivity to data lying off the manifold, possibly leading to mismatches between successive local models. The mixtures of robust probabilistic principal component analyzers introduced in this paper heal this problem as each component is able to cope with atypical data while identifying the local principal directions. Moreover, the standard mixture of Gaussians is a particular instance of this more general model.

1 Introduction

Principal component analysis (PCA) is a well-known statistical technique for linear dimensionality reduction [1]. 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 distribution [2]. The Student-t distribution is a heavy tailed generalization of the Gaussian distribution. Similarly, the robust probabilistic reformulation of PCA generalizes standard probabilistic PCA [3, 4]. Increasing the robustness by replacing Gaussian distributions with Student-t distributions was already proposed in the context of finite mixture modeling [5].

*N.D. is research fellow of the F.N.R.S.

In contrast with previous robust approaches to PCA (see for example [6] and [7], 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 models instead of using nonlinear dimensionality reduction techniques [9].

In this paper mixtures of robust PPCAs are introduced. The method generalizes mixtures of standard PPCAs proposed in [10]. An additional 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 to be neglected to avoid numerical instabilities. In Section 2 and 3, we derive the model and then illustrate its use on the examples in Section 4.

2 Robust Probabilistic Principal Component Analysis

Principal component analysis (PCA) seeks a linear projection, which maps a set of observations $\{\mathbf{y}_n\}_{n=1}^N$ to a set of lower dimensional vectors $\{\mathbf{x}_n\}_{n=1}^N$ such that the variance in the projection space is maximized [1]. This can be formalized as a latent variable model with $\mathbf{y}_n = \mathbf{W}\mathbf{x}_n + \boldsymbol{\mu} + \epsilon_n$. Matrix $\mathbf{W} \in \mathbb{R}^{D\times d}$ is the (transposed) projection matrix. The data offset and the projection errors are respectively denoted by $\boldsymbol{\mu}$ and $\{\epsilon_n\}_{n=1}^N$. In probabilistic principal component analysis (PPCA) [3, 4], it is further assumed that the error terms, as well as the uncertainty on the latent vectors, are drawn from isotropic Gaussian distributions. As shown in [4], ML leads to a solution that is equivalent to PCA (up to a rotation). More specifically, the columns of the ML estimate of $\mathbf{W}$ span the same subspace as the d principal eigenvectors of the sample covariance matrix.

As its non-probabilistic counterpart, PPCA suffers from the following well-known problem: since the approach is based on Gaussian noise model, it is very sensitive to atypical observations such as outliers. Outliers occur quite often in practice. Compared to the Gaussian distribution, the Student-t distribution has an additional parameter, called the number of degrees of freedom ν. This parameter regulates the thickness of the distribution tails and therefore its robustness to atypical observations. As noted in [11], the Student-t distribution can be interpreted as the following latent variable model:

$$\mathcal{S}(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) = \int_0^{+\infty} \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, u\boldsymbol{\Lambda})\mathcal{G}(u|\tfrac{\nu}{2},\tfrac{\nu}{2})du, \tag{1}$$

where $u > 0$ is a latent scale variable, $\mathcal{N}(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Lambda}) \propto |\boldsymbol{\Lambda}|^{1/2} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{y}-\boldsymbol{\mu})}$ is the Gaussian distribution (with inverse covariance $\boldsymbol{\Lambda}$) and $\mathcal{G}(u|\alpha,\beta) \propto u^{\alpha-1}e^{-\beta u}$ is

the Gamma distribution. Hence, the Student-t distribution can be reformulated as an infinite mixture of Gaussian distributions with the same mean and where the prior on u is a Gamma distribution with parameters depending only on ν.

As shown in [2], PPCA can be made robust by using a Student-t model instead of a Gaussian one. This leads to the following robust reformulation:

$$p(\mathbf{x}_n) = \mathcal{S}(\mathbf{x}_n|\mathbf{0}, \mathbf{I}_d, \nu), \tag{2}$$

$$p(\mathbf{y}_n|\mathbf{x}_n) = \mathcal{S}(\mathbf{y}_n|\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu}, \tau\mathbf{I}_D, \nu), \tag{3}$$

where τ is the inverse noise variance. Integrating out the latent variable $\mathbf{x}_n$ leads to the marginal likelihood $p(\mathbf{y}_n) = \mathcal{S}(\mathbf{y}_n|\boldsymbol{\mu}, \mathbf{A}, \nu)$, with $\mathbf{A}^{-1} \equiv \mathbf{W}\mathbf{W}^{\mathrm{T}} + \tau^{-1}\mathbf{I}_D$. Next, we show how to combine multiple robust PPCAs to form a mixture.

3 Mixtures of Robust PPCAs

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

$$p(\mathbf{y}_n) = \textstyle\sum_m \pi_m p(\mathbf{y}_n|m). \tag{4}$$

where $\{\pi_m\}_{m=1}^M$ is the set of mixture proportions, with $\sum_m \pi_m = 1$ and $\forall m : \pi_m \geq 0$. The likelihood term $p(\mathbf{y}_n|m)$ is defined as a single robust PPCA, that is $p(\mathbf{y}_n|m) = \mathcal{S}(\mathbf{y}_n|\boldsymbol{\mu}_m, \mathbf{A}_m, \nu_m)$ with $\mathbf{A}_m^{-1} \equiv \mathbf{W}_m\mathbf{W}_m^{\mathrm{T}} + \tau_m^{-1}\mathbf{I}_D$.

In order to reconstruct the m^{th} robust PPCA model, we define a set of continuous lower dimensional latent variables $\{\mathbf{x}_{nm}\}_{n=1}^N$ and a set of latent scale variables $\{u_{nm}\}_{n=1}^N$. In addition, for each observation $\mathbf{y}_n$, there is a binary latent variable $\mathbf{z}_n$ indicating by which component $\mathbf{y}_n$ was generated. The resulting complete probabilistic model is defined as follows:

$$P(\mathbf{z}_n) = \textstyle\prod_m \pi_m^{z_{nm}}, \tag{5}$$

$$p(\mathbf{u}_n|\mathbf{z}_n) = \textstyle\prod_m \mathcal{G}(u_{nm}|\frac{\nu_m}{2}, \frac{\nu_m}{2})^{z_{nm}}, \tag{6}$$

$$p(\boldsymbol{\chi}_n|\mathbf{u}_n, \mathbf{z}_n) = \textstyle\prod_m \mathcal{N}(\mathbf{x}_{nm}|\mathbf{0}, u_{nm}\mathbf{I}_d)^{z_{nm}}, \tag{7}$$

$$p(\mathbf{y}_n|\boldsymbol{\chi}_n, \mathbf{u}_n, \mathbf{z}_n) = \textstyle\prod_m \mathcal{N}(\mathbf{y}_n|\mathbf{W}_m\mathbf{x}_{nm} + \boldsymbol{\mu}_m, u_{nm}\tau_m\mathbf{I}_D)^{z_{nm}}, \tag{8}$$

where $\mathbf{z}_n = (z_{n1}, \ldots, z_{nM})^{\mathrm{T}}$, $\mathbf{u}_n = (u_{n1}, \ldots, u_{nM})^{\mathrm{T}}$ and $\boldsymbol{\chi}_n = (\mathbf{x}_{n1}, \ldots, \mathbf{x}_{nM})^{\mathrm{T}}$. Integrating out all the latent variables leads to (4) as desired.

The (complete) log-likelihood is given by

$$\log \mathcal{L}(\{\pi_m\}, \{\boldsymbol{\mu}_m\}, \{\mathbf{W}_m\}, \{\tau_m\}, \{\nu_m\}) = \textstyle\sum_n \log p(\mathbf{y}_n, \boldsymbol{\chi}_n, \mathbf{u}_n, \mathbf{z}_n). \tag{9}$$

The EM algorithm [8] maximizes the expectation of this quantity, the expectation being taken with respect to the posterior distribution of the latent variables. Therefore, the E-step consists in estimating these posteriors, i.e. $P(z_{nm} = 1|\mathbf{y}_n)$, $p(u_{nm}|\mathbf{y}_n, z_{nm} = 1)$ and $p(\mathbf{x}_{nm}|\mathbf{y}_n, \mathbf{u}_n, z_{nm} = 1)$, which allow us to compute the expectations required in the M-step (details omitted):

$$\bar{\rho}_{nm} \equiv \mathrm{E}\{z_{nm}\} = \tfrac{\pi_m \mathcal{S}(\mathbf{y}_n|\boldsymbol{\mu}_m, \mathbf{A}_m, \nu_m)}{\sum_m \pi_m \mathcal{S}(\mathbf{y}_n|\boldsymbol{\mu}_m, \mathbf{A}_m, \nu_m)}, \tag{10}$$

$$\bar{u}_{nm} \equiv \mathrm{E}\{u_{nm}\} = \tfrac{D+\nu_m}{(\mathbf{y}_n - \boldsymbol{\mu}_n)^{\mathrm{T}}\mathbf{A}_m(\mathbf{y}_n - \boldsymbol{\mu}_m) + \nu_m}, \tag{11}$$

$$\log \tilde{u}_{nm} \equiv \mathrm{E}\{\log u_{nm}\} = \psi\left(\tfrac{D+\nu_m}{2}\right) - \log\left(\tfrac{(\mathbf{y}_n-\boldsymbol{\mu}_m)^{\mathrm{T}}\mathbf{A}_m(\mathbf{y}_n-\boldsymbol{\mu}_m)+\nu_m}{2}\right), \quad (12)$$

$$\bar{\mathbf{x}}_{nm} \equiv \mathrm{E}\{\mathbf{x}_{nm}\} = \tau_m \mathbf{B}_m^{-1}\mathbf{W}_m^{\mathrm{T}}(\mathbf{y}_n - \boldsymbol{\mu}_m), \quad (13)$$

$$\bar{\mathbf{S}}_{nm} \equiv \mathrm{E}\{u_{nm}\mathbf{x}_{nm}\mathbf{x}_{nm}^{\mathrm{T}}\} = \mathbf{B}_m^{-1} + \bar{u}_{nm}\bar{\mathbf{x}}_{nm}\bar{\mathbf{x}}_{nm}^{\mathrm{T}}, \quad (14)$$

where $\mathbf{B}_m \equiv \tau_m \mathbf{W}_m^{\mathrm{T}}\mathbf{W}_m + \mathbf{I}_d$ and $\psi(\cdot) \equiv \Gamma'(\cdot)/\Gamma(\cdot)$ is the digamma function.

The M-step maximizes the expected value of (9). This leads to the following update rules for the parameters:

$$\pi_m \leftarrow \tfrac{1}{N}\textstyle\sum_n \bar{\rho}_{nm}, \quad (15)$$

$$\boldsymbol{\mu}_m \leftarrow \tfrac{\sum_n \bar{\rho}_{nm}\bar{u}_{nm}(\mathbf{y}_n - \mathbf{W}_m\bar{\mathbf{x}}_{nm})}{\sum_n \bar{\rho}_{nm}\bar{u}_{nm}}, \quad (16)$$

$$\mathbf{W}_m \leftarrow \left(\textstyle\sum_n \bar{\rho}_{nm}\bar{u}_{nm}(\mathbf{y}_n - \boldsymbol{\mu})\bar{\mathbf{x}}_{nm}^{\mathrm{T}}\right)\left(\textstyle\sum_n \bar{\rho}_{nm}\bar{\mathbf{S}}_{nm}\right)^{-1}, \quad (17)$$

$$\tau_m^{-1} \leftarrow \tfrac{1}{DN\pi_m}\textstyle\sum_n \rho_{nm}\big(\bar{u}_{nm}\|\mathbf{y}_n - \boldsymbol{\mu}_m\|^2 - \mathrm{tr}\{\mathbf{W}_m\bar{\mathbf{S}}_{nm}\mathbf{W}_m^{\mathrm{T}}\}\big). \quad (18)$$

Similarly, the maximum likelihood estimate for each ν_m is found by solving the following expression by line search at each iteration:

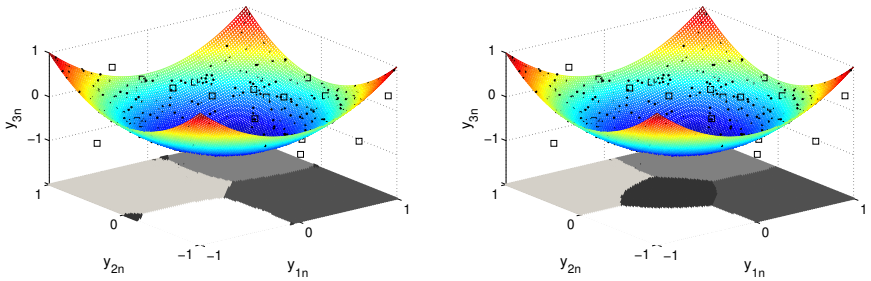
$$1 + \log\left(\tfrac{\nu_m}{2}\right) - \psi\left(\tfrac{\nu_m}{2}\right) + \tfrac{1}{N\pi_m}\textstyle\sum_n \bar{\rho}_{nm}\{\log\tilde{u}_{nm} - \bar{u}_{nm}\} = 0. \quad (19)$$

Using a low-dimensional representation in the latent space has a clear advantage over a standard mixture of Gaussians (or Student-*t*s). Indeed, the number of parameters to estimate each covariance matrix in the case of the latter is $D(D+1)/2$, while it is equal to $Dd+1-d(d-1)/2$ in the case of the mixture of (robust) PPCAs (taking the rotational invariance into account). The interesting feature of our approach is that the correlations between the principal directions are not neglected. By contrast, it is common practice to force the covariance matrices to be diagonal in order to avoid numerical instabilities.

4 Experiments

The two experiments considered in this section illustrate the use of the mixture of robust PPCAs. In the first one, it is shown how a low-dimensional nonlinear manifold spoiled by noisy data can still be estimated. In the second one, it allows finding and interpreting clusters of high-dimensional data.

Robust reconstruction of low-dimensional manifolds: The following 3-dimensional data set is used: $y_{3n} = y_{1n}^2 + y_{2n}^2 - 1 + \epsilon_n$. 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 $\{\epsilon_n\}_{n=1}^N$ are distributed according to $\mathcal{N}(\epsilon_n|0,\tau_\epsilon)$, with $\tau_\epsilon^{-1} = 0.01$. The data is located along a 2-dimensional paraboloid; 500 training data were genererated. The number of components was fixed to 5 and d was set to 2 (the true dimension of the manifold). Fig. 1 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



(a) Projection by mixt. PPCAs. (b) Projection by mixt. robust PPCAs.

Fig. 1: Noisy paraboloid data set. Each shaded region is associated with a different local linear model (or component). These regions can be projected on the 3-dimensonal paraboloid in order to identify which part is modeled by which linear model. The outliers are indicated by squares.

to each component (or local linear model) after projection. In other words, the data belonging to these regions are assigned to the component with highest responsibility. When the local models are nicely aligned with the manifold, the regions do not split. However, as shown in Fig. 1, only the mixture of robust PPCAs provides a satisfactory solution. Indeed, one of the components of the mixture of standard PPCAs is "lost"; it is not oriented along the manifold (and thus crosses the paraboloid) to account for the dispersion of outliers.

Analysis of high-dimensional data: In this experiment, the well-known USPS handwritten digit data[1] is considered. The data are black and white 16×16-pixels images of digits (0 to 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. We compared the mixture of PPCAs and the mixture of robust PPCAs in their ability to find the two main clusters assuming a 1-dimensional latent space. The result is shown in Fig. 2. Each row represents images generated along the principal directions. The mixture of robust RPPCAs 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

[1]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. 2: Mixture of 2 component PPCAs with 1-dimensional latent space to cluster USPS digit 2 and 3, and outliers digit 0. Right: robust, Left: standard.

entirely. This experiment shows that the mixture of robust PPCAs is able to model such noisy manifolds, which are common in practice.

5 Conclusion

Principal component analysis is an elementary and fundamental tool for exploratory data mining, data visualization and unsupervised learning. When tackling real-life problems, it is essential to take a robust approach. Here, the term "robust" is used to indicate that the performance of the algorithm is not spoiled by non-Gaussian noise (e.g., outliers). This property is obtained by exploiting the heavier distribution tails of the Student-t. In this paper, mixtures of robust probabilistic principal component analyzers were introduced. They provide a practical approach for discovering nonlinear relationships in the data by combining robust local linear models. More generally, they can also be used for robust mixture modeling, the multivariate Gaussian mixture model being a special case.

References

[1] Ian T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[2] C. Archambeau, N. Delannay, and M. Verleysen. Robust probabilistic projections. In W. W. Cohen and A. Moore, editors, *ICML 23*, pages 33–40. ACM, 2006.

[3] S. T. Roweis. EM algorithms for PCA and SPCA. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *NIPS 10*. The MIT Press, 1998.

[4] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society B*, 61:611–622, 1999.

[5] D. Peel and G. J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10:339–348, 2000.

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

[7] K. Huang, Y. Ma, and R. Vidal. Minimum effective dimension for mixtures of subspaces: A robust GPCA algorithm and its applications. In *CVPR 2004*, vol 2, pages 631–638.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–38, 1977.

[9] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[10] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.

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