Robust Probabilistic Projections

Cédric Archambeau CEDRIC.ARCHAMBEAU@UCLOUVAIN.BE
Nicolas Delannay NICOLAS.DELANNAY@UCLOUVAIN.BE
Michel Verleysen VERLEYSEN@DICE.UCL.AC.BE
Université catholique de Louvain, Machine Learning Group, 3 Pl. du Levant, B-1348 Louvain-la-Neuve, Belgium.

Abstract

Principal components and canonical correlations are at the root of many exploratory data mining techniques and provide standard pre-processing tools in machine learning. Lately, probabilistic reformulations of these methods have been proposed (Roweis, 1998; Tipping & Bishop, 1999b; Bach & Jordan, 2005). They are based on a Gaussian density model and are therefore, like their non-probabilistic counterpart, very sensitive to atypical observations. In this paper, we introduce robust probabilistic principal component analysis and robust probabilistic canonical correlation analysis. Both are based on a Student-t density model. The resulting probabilistic reformulations are more suitable in practice as they handle outliers in a natural way. We compute maximum likelihood estimates of the parameters by means of the EM algorithm.

1. Introduction

Principal component analysis (PCA) is a standard statistical tool for dimensionality reduction (Hotelling, 1933; Jolliffe, 1986). It looks for a linear transformation which projects high-dimensional data into a low-dimensional subspace while preserving the data variance (i.e., it minimizes the mean squared reconstruction error). Therefore, PCA is used as a pre-processing step in many applications such as data compression, data visualization, image analysis, etc.

A related technique is canonical correlation analysis (CCA) (Hotelling, 1936). In general, CCA is used to analyze the (linear) relationship between two sets of variables. It looks for a *pair* of linear transformations which projects the two data sets into a common low-dimensional subspace while maximizing pairwise correlations.

Recently, PCA and CCA were reformulated as probabilistic latent variable models (Roweis, 1998; Tipping & Bishop, 1999b; Bach & Jordan, 2005). Defining a proper density model has a number of significant advantages. First, the associated likelihood measure allows us to compare probabilistic PCA (PPCA) and probabilistic CCA (PCCA) with other probabilistic techniques. Second, it is straightforward to extend PPCA and PCCA in order to handle missing data or to construct mixtures of PPCA (Tipping & Bishop, 1999a) or PCCA (Verbeek et al., 2004). This is important as it enables us to model non-linear relationships by aligning a collection of such local models. Another attractive feature is that they allow computing few principal/canonical axes in an efficient way (Roweis, 1998). Finally, although direct maximization of the data likelihood under these probabilistic models (or their extensions) is not always possible, local maxima of the likelihood function can in general be found by means of the EM algorithm (Dempster et al., 1977).

Nevertheless, (P)PCA and (P)CCA have severe limitations in practice. Both are based on a Gaussian density model. Therefore, atypical observations lead possibly to severe biases in the parameter estimates when using a maximum likelihood approach. In this paper, we propose to handle atypical observations in a principled and automatic way by using a Student-t density model. The Student-t distribution is a heavy tailed generalization of the Gaussian distribution. Replacing Gaussian distributions with Student-t distributions for increasing the robustness was already done by Peel and McLachlan (2000) and more recently by Archambeau (2005) in the context of finite mixture models.

Unlike previous robust approaches to linear projection, we use a probabilistic formalism, which has significant advantages: (i) the extension to the Bayesian frame-

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

work is possible (ii) constructing mixtures of robust PCA/CCA is straightforward and (iii) missing data can be dealt with quite easily. Besides, the proposed techniques are very practical for tackling real life problems, as they only require choosing the dimension of the projection space. In contrast, previous attempts need in general to optimise several additional parameters (e.g., Xu and Yuille (1995)'s robust PCA needs three). These are often difficult to adjust in practice.

This paper is organized as follows. In Section 2, we describe how the Student-t distribution can be interpreted as a latent variable model. Next, PPCA and PCCA are recalled. In Section 3, we introduce the robust probabilistic models and show how their parameters can be learnt by means of the EM algorithm. Finally, in Section 5 the models are validated experimentally.

2. Latent Variable View of the Student-t Distribution

Let $\mathbf{y}$ be a D-dimensional feature vector. The multivariate Student-t distribution is given by

$$\mathcal{S}(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) = \frac{\Gamma\left(\frac{D+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{\frac{D}{2}}}|\boldsymbol{\Lambda}|^{\frac{1}{2}} \times \left[1+\frac{1}{\nu}(\mathbf{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{y}-\boldsymbol{\mu})\right]^{-\frac{D+\nu}{2}}, \quad (1)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are respectively the mean and the precision (i.e., inverse covariance matrix), and $\Gamma(\cdot)$ denotes the gamma function. Parameter $\nu > 0$ is the degrees of freedom. It regulates the thickness of the distribution tails and therefore its robustness to atypical observations. When ν tends to infinity, the Gaussian distribution is recovered.

As noted by Liu and Rubin (1995), the Student-t distribution can be interpreted as the following latent variable model:

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

where $u > 0$ is a latent scale variable. The associated graphical model is shown in Figure 1. The Gaussian and the Gamma distribution are respectively given by

$$\mathcal{N}(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Lambda}) = (2\pi)^{-\frac{D}{2}}|\boldsymbol{\Lambda}|^{\frac{1}{2}} \times \exp\left\{-\tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\mathbf{y}-\boldsymbol{\mu})\right\}, \quad (3)$$

$$\mathcal{G}(u|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}u^{\alpha-1}\exp(-\beta u). \quad (4)$$

From (2) we see that the Student-t distribution can be viewed as an infinite mixture of Gaussian distributions with the same mean and where the prior distribution on u is a Gamma distribution with parameters depending only on ν.

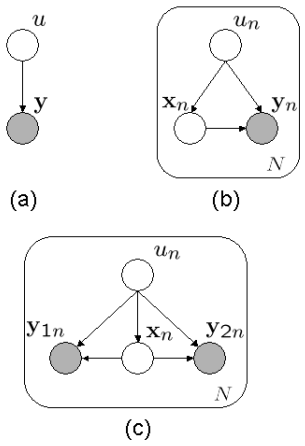


Figure 1. Graphical models of (a) the Student-t distribution, (b) robust probabilistic PCA and (c) robust probabilistic CCA. The shaded nodes are observed, arrows represent conditional dependencies between random variables and plates denote repetitions.

3. Linear Probabilistic Projections

Principal component analysis (PCA) and canonical correlation analysis (CCA) are both exploratory linear data projection techniques. PCA seeks for a linear projection $\mathbf{W} \in \mathbb{R}^{D\times d}$, which maps a set of observations $\{\mathbf{y}_n\}_{n=1}^N$ to a set of lower dimensional latent vectors $\{\mathbf{x}_n\}_{n=1}^N$ such that the variance in the projection space is maximized (Hotelling, 1933). By contrast, CCA investigates the relationship between *two* sets of variables (Hotelling, 1936) and determines how many dimensions are needed to account for that relationship. It seeks for a pair of linear projections $\mathbf{W}_1 \in \mathbb{R}^{D_1\times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{D_2\times d}$, which map two sets of observations $\{\mathbf{y}_{1n}\}_{n=1}^N$ and $\{\mathbf{y}_{2n}\}_{n=1}^N$ to a set of lower dimensional latent vectors $\{\mathbf{x}_n\}_{n=1}^N$ such that, in the projection space, one component within each set is maximally correlated with a single component of the other set.

3.1. Probabilistic PCA

In general, PCA assumes that each d-dimensional latent vector $\mathbf{x}_n$ is a linear projection of a D-dimensional feature vector $\mathbf{y}_n$, with $D \geq d$. The latent variable model is defined as follows:

$$\mathbf{y}_n = \mathbf{W}\mathbf{x}_n + \boldsymbol{\mu} + \epsilon_n, \quad (5)$$

where $\boldsymbol{\mu}$ is the data offset and $\{\epsilon_n\}_{n=1}^N$ are the projection errors. In PPCA, these error terms are assumed to be drawn from an isotropic Gaussian distribution

with inverse variance equal to τ and the uncertainty on the latent vectors is modeled by a unit isotropic Gaussian distribution. Tipping and Bishop (1999b) showed that maximizing the resulting incomplete data log-likelihood leads to PCA (up to a rotation). In other words, the columns of the maximum likelihood (ML) solution for the projection matrix $\mathbf{W}$ span the same subspace as the d principal eigenvectors of the sample covariance matrix $\bar{\mathbf{\Sigma}}$:

$$\widehat{\mathbf{W}}_{\mathrm{ML}} = \mathbf{U}_d(\mathbf{\Upsilon}_d - \tau^{-1}\mathbf{I}_d)^{\frac{1}{2}}\mathbf{R}, \tag{6}$$

where $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is the d-dimensional unit matrix, $\mathbf{U}_d \in \mathbb{R}^{D\times d}$ is the matrix of the d principal eigenvectors of the sample covariance matrix $\bar{\mathbf{\Sigma}} = \frac{1}{N}\sum_n(\mathbf{y}_n - \bar{\boldsymbol{\mu}})(\mathbf{y}_n - \bar{\boldsymbol{\mu}})^{\mathrm{T}}$, $\mathbf{\Upsilon} \in \mathbb{R}^{d\times d}$ is the diagonal matrix of the corresponding eigenvalues and $\mathbf{R} \in \mathbb{R}^{d\times d}$ is an arbitrary rotation matrix. Note that the true principal axes can be recovered by post-multiplying $\widehat{\mathbf{W}}_{\mathrm{ML}}$ by $\mathbf{R}^{\mathrm{T}}$, which is the matrix of the eigenvectors of $\widehat{\mathbf{W}}_{\mathrm{ML}}^{\mathrm{T}}\widehat{\mathbf{W}}_{\mathrm{ML}}$.

3.2. Probabilistic CCA

More recently, Bach and Jordan (2005) cast CCA as a probabilistic model in similar manner. The latent variable model for PCCA is defined as follows:

$$\begin{cases} \mathbf{y}_{1n} = \mathbf{W}_1\mathbf{x}_n + \boldsymbol{\mu}_1 + \boldsymbol{\epsilon}_{1n}, \\ \mathbf{y}_{2n} = \mathbf{W}_2\mathbf{x}_n + \boldsymbol{\mu}_2 + \boldsymbol{\epsilon}_{2n}. \end{cases} \tag{7}$$

Note that $\min\{D_1, D_2\} \geq d$. Again a unit isotropic Gaussian distribution is assumed for the d-dimensional latent vectors, but the error terms $\{\boldsymbol{\epsilon}_{1n}\}_{n=1}^N$ and $\{\boldsymbol{\epsilon}_{2n}\}_{n=1}^N$ are now considered to be drawn from two multivariate Gaussian distributions. Their inverse covariance matrices are respectively denoted by $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$. Maximizing the incomplete joint data log-likelihood leads then to CCA as the ML estimates of the projection matrices span the same subspace as the canonical directions. The ML estimates of the projection matrix parameters are given by

$$\widehat{\mathbf{W}}_1 = \bar{\mathbf{\Sigma}}_{11}\mathbf{U}_{1d}\mathbf{Q}_1, \tag{8}$$

$$\widehat{\mathbf{W}}_2 = \bar{\mathbf{\Sigma}}_{22}\mathbf{U}_{2d}\mathbf{Q}_2, \tag{9}$$

where $\bar{\mathbf{\Sigma}}_{11}$ and $\bar{\mathbf{\Sigma}}_{22}$ are the sample covariance matrices, and the columns of the associated matrices $\mathbf{U}_{1d} \in \mathbb{R}^{D_1\times d}$ and $\mathbf{U}_{2d} \in \mathbb{R}^{D_2\times d}$ are the d first canonical directions. The matrices $\mathbf{Q}_1 \in \mathbb{R}^{d\times d}$ and $\mathbf{Q}_2 \in \mathbb{R}^{d\times d}$ are arbitrary matrices such that $\mathbf{Q}_1\mathbf{Q}_2^{\mathrm{T}} = \mathbf{\Upsilon}_d$, where $\mathbf{\Upsilon}_d \in \mathbb{R}^{d\times d}$ is the diagonal matrix of the corresponding canonical correlations. As discussed in Appendix A, the true canonical directions (and correlations) can be recovered by a post-processing step, which removes the rotational ambiguity.

4. Robust Probabilistic Projections

Linear probabilistic projections, such as PPCA and PCCA (as well as their non-probabilistic counterpart) suffer from a common problem: since they use Gaussian density models they are very sensitive to atypical observations such as outliers. Outliers occur quite often in practice. For example in computer vision applications, they appear due to pixels that are corrupted by noise, occlusions or alignment errors (de la Torre & Black, 2001). Therefore, we propose to use Student-t density models instead of Gaussian ones. As discussed in Section 2, the robustness of the Student-t distribution can be tuned by means of its degrees of freedom.

4.1. Robust Probabilistic PCA

Instead of choosing a Gaussian noise model, we assume that the noise is drawn from an infinite mixture of Gaussian distributions, i.e., from a Student-t distribution. We also assume that outliers in the feature space will be outliers in the latent space. Therefore, we choose the prior distribution on the latent vectors to be a unit variance Student-t distribution. As a result, the effect of the outliers will be lowered in the feature as well as the latent space. This leads to the following probabilistic model:

$$p(\mathbf{x}_n) = \mathcal{S}(\mathbf{x}_n|\mathbf{0}, \mathbf{I}_d, \nu), \tag{10}$$

$$p(\mathbf{y}_n|\mathbf{x}_n) = \mathcal{S}(\mathbf{y}_n|\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu}, \tau\mathbf{I}_D, \nu), \tag{11}$$

where $\boldsymbol{\mu}$ is the data offset and $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is the d-dimensional unit matrix. Unfortunately, unlike in PPCA, direct maximization of the incomplete data log-likelihood $\sum_n \log p(\mathbf{y}_n)$ with respect to the parameters is intractable.

However, we can define an alternative robust probabilistic model for PPCA using (2):

$$p(u_n) = \mathcal{G}(u_n|\tfrac{\nu}{2}, \tfrac{\nu}{2}), \tag{12}$$

$$p(\mathbf{x}_n|u_n) = \mathcal{N}(\mathbf{x}_n|\mathbf{0}, u_n\mathbf{I}_d), \tag{13}$$

$$p(\mathbf{y}_n|\mathbf{x}_n, u_n) = \mathcal{N}(\mathbf{y}_n|\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu}, u_n\tau\mathbf{I}_D). \tag{14}$$

Its graphical representation is shown in Figure 1. Note that integrating out u_n from (13) and (14) leads respectively to (10) and (11). In order to find ML estimates of the parameters, we follow an EM approach by maximizing the expected complete data log-likelihood (Neal & Hinton, 1998).

The complete log-likelihood is given by

$$\log\mathcal{L}(\boldsymbol{\mu}, \mathbf{W}, \tau, \nu) = \textstyle\sum_n \log p(\mathbf{y}_n, \mathbf{x}_n, u_n). \tag{15}$$

The E-step consists then in updating the posterior distribution of the latent variables. First, note that the

Gamma distribution is conjugate to the exponential family. This leads to the following posterior distribution for the latent scale variables:

$$p(u_n|\mathbf{y}_n) \propto p(\mathbf{y}_n|u_n)p(u_n), \quad (16)$$
$$= \mathcal{G}(u_n|\tfrac{D+\nu}{2}, \tfrac{(\mathbf{y}_n-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{A}(\mathbf{y}_n-\boldsymbol{\mu})+\nu}{2}),$$

where $\mathbf{A}^{-1} \equiv \mathbf{W}\mathbf{W}^{\mathrm{T}} + \tau^{-1}\mathbf{I}_D$. Second, the posterior distribution of the latent vectors is given by

$$p(\mathbf{x}_n|\mathbf{y}_n, u_n) \propto p(\mathbf{y}_n|\mathbf{x}_n, u_n)p(\mathbf{x}_n|, u_n) \quad (17)$$
$$= \mathcal{N}(\mathbf{x}_n|\tau\mathbf{B}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{y}_n - \boldsymbol{\mu}), u_n\mathbf{B}),$$

where $\mathbf{B} \equiv \tau\mathbf{W}^{\mathrm{T}}\mathbf{W} + \mathbf{I}_d$. The expectations needed to update the parameters in the M-step are then given by

$$\bar{u}_n = \tfrac{D+\nu}{(\mathbf{y}_n-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{A}(\mathbf{y}_n-\boldsymbol{\mu})+\nu}, \quad (18)$$
$$\log\tilde{u}_n = \psi\left(\tfrac{D+\nu}{2}\right) - \log\left(\tfrac{(\mathbf{y}_n-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{A}(\mathbf{y}_n-\boldsymbol{\mu})+\nu}{2}\right), \quad (19)$$
$$\bar{\mathbf{x}}_n = \tau\mathbf{B}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{y}_n - \boldsymbol{\mu}), \quad (20)$$
$$\bar{\mathbf{S}}_n = \mathbf{B}^{-1} + \bar{u}_n\bar{\mathbf{x}}_n\bar{\mathbf{x}}_n^{\mathrm{T}}, \quad (21)$$

where $\bar{u}_n \equiv \mathrm{E}\{u_n\}$, $\log\tilde{u}_n \equiv \mathrm{E}\{\log u_n\}$, $\bar{\mathbf{x}}_n \equiv \mathrm{E}\{\mathbf{x}_n\}$ and $\bar{\mathbf{S}}_n \equiv \mathrm{E}\{u_n\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\}$. In (19), $\psi(\cdot)$ denotes the digamma function.

Next, the M-step is found by maximizing the expectation of (15) with respect to the latent variables, resulting in the following update rules for the parameters:

$$\boldsymbol{\mu} \leftarrow \tfrac{\sum_n \bar{u}_n(\mathbf{y}_n - \mathbf{W}\bar{\mathbf{x}}_n)}{\sum_n \bar{u}_n}, \quad (22)$$
$$\mathbf{W} \leftarrow \left(\textstyle\sum_n \bar{u}_n(\mathbf{y}_n - \boldsymbol{\mu})\bar{\mathbf{x}}_n^{\mathrm{T}}\right)\left(\textstyle\sum_n \bar{\mathbf{S}}_n\right)^{-1}, \quad (23)$$
$$\tau^{-1} \leftarrow \tfrac{1}{N}\textstyle\sum_n \big\{\bar{u}_n\|\mathbf{y}_n - \boldsymbol{\mu}\|^2 - 2\bar{u}_n(\mathbf{y}_n - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{W}\bar{\mathbf{x}}_n$$
$$+ \mathrm{tr}\{\bar{\mathbf{S}}_n\mathbf{W}^{\mathrm{T}}\mathbf{W}\}\big\}. \quad (24)$$

Observe how the contribution of each data point is weighted according to the associated latent scale variable. At each iteration, a maximum likelihood estimate of ν is found by solving the following expression by line search:

$$1 + \log\left(\tfrac{\nu}{2}\right) - \psi\left(\tfrac{\nu}{2}\right) + \tfrac{1}{N}\textstyle\sum_n\{\log\tilde{u}_n - \bar{u}_n\} = 0. \quad (25)$$

4.2. Robust Probabilistic CCA

A similar approach can be used to construct robust PCCA. Consider the following probabilistic model:

$$p(\mathbf{x}_n) = \mathcal{S}(\mathbf{x}_n|\mathbf{0}, \mathbf{I}_d, \nu), \quad (26)$$
$$p(\mathbf{y}_{1n}|\mathbf{x}_n) = \mathcal{S}(\mathbf{y}_{1n}|\mathbf{W}_1\mathbf{x}_n + \boldsymbol{\mu}_1, \boldsymbol{\Psi}_1, \nu), \quad (27)$$
$$p(\mathbf{y}_{2n}|\mathbf{x}_n) = \mathcal{S}(\mathbf{y}_{2n}|\mathbf{W}_2\mathbf{x}_n + \boldsymbol{\mu}_2, \boldsymbol{\Psi}_2, \nu). \quad (28)$$

Again, making the latent scale variable explicit leads to the following (compact) reformulation:

$$p(u_n) = \mathcal{G}(u_n|\tfrac{\nu}{2}, \tfrac{\nu}{2}), \quad (29)$$
$$p(\mathbf{x}_n|u_n) = \mathcal{N}(\mathbf{x}_n|\mathbf{0}, u_n\mathbf{I}_d), \quad (30)$$
$$p(\mathbf{y}_n|\mathbf{x}_n, u_n) = \mathcal{N}(\mathbf{y}_n|\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu}, u_n\boldsymbol{\Psi}), \quad (31)$$

where $\mathbf{y}_n \equiv (\mathbf{y}_{n1}, \mathbf{y}_{n2})^{\mathrm{T}}$, $\boldsymbol{\mu} \equiv (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)^{\mathrm{T}}$, $\mathbf{W} \equiv (\mathbf{W}_1^{\mathrm{T}}, \mathbf{W}_2^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\Psi} \equiv \mathrm{diag}\{\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2\}$. The corresponding graphical model is shown in Figure 1. Next, we use the EM algorithm to find ML estimates of the parameters.

The complete log-likelihood is given by

$$\log\mathcal{L}(\boldsymbol{\mu}, \mathbf{W}, \boldsymbol{\Psi}, \nu) = \textstyle\sum_n \log p(\mathbf{y}_n, \mathbf{x}_n, u_n). \quad (32)$$

First, we update the joint posterior distributions of the latent scale variables and the latent vectors. These are respectively given by

$$p(u_n|\mathbf{y}_n) = \mathcal{G}(u_n|\tfrac{D+\nu}{2}, \tfrac{(\mathbf{y}_n-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{A}(\mathbf{y}_n-\boldsymbol{\mu})+\nu}{2}), \quad (33)$$
$$p(\mathbf{x}_n|\mathbf{y}_n, u_n) = \mathcal{N}(\mathbf{x}_n|\mathbf{B}^{-1}\mathbf{W}^{\mathrm{T}}\boldsymbol{\Psi}(\mathbf{y}_n - \boldsymbol{\mu}), u_n\mathbf{B}), \quad (34)$$

where $D = D_1 + D_2$, $\mathbf{A}^{-1} \equiv \mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi}^{-1}$ and $\mathbf{B} \equiv \mathbf{W}^{\mathrm{T}}\boldsymbol{\Psi}\mathbf{W} + \mathbf{I}_d$. From these posterior distributions, we obtain the expectations $\bar{u}_n \equiv \mathrm{E}\{u_n\}$ and $\log\tilde{u}_n \equiv \mathrm{E}\{\log u_n\}$, which are respectively defined by (18) and (19), and $\bar{\mathbf{x}}_n \equiv \mathrm{E}\{\mathbf{x}_n\}$ and $\bar{\mathbf{S}}_n \equiv \mathrm{E}\{u_n\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\}$:

$$\bar{\mathbf{x}}_n = \mathbf{B}^{-1}\mathbf{W}^{\mathrm{T}}\boldsymbol{\Psi}(\mathbf{y}_n - \boldsymbol{\mu}), \quad (35)$$
$$\bar{\mathbf{S}}_n = \mathbf{B}^{-1} + \bar{u}_n\bar{\mathbf{x}}_n\bar{\mathbf{x}}_n^{\mathrm{T}}. \quad (36)$$

Maximizing the expected complete data log-likelihood leads then to the same M-step for $\boldsymbol{\mu}$, $\mathbf{W}$ and ν as for robust PPCA, while for the noise matrices we obtain

$${\boldsymbol{\Psi}_i}^{-1} \leftarrow \left(\tfrac{1}{N}\textstyle\sum_n \big\{\bar{u}_n(\mathbf{y}_n - \boldsymbol{\mu})(\mathbf{y}_n - \boldsymbol{\mu})^{\mathrm{T}}\right. \quad (37)$$
$$\left. -2\bar{u}_n(\mathbf{y}_n - \boldsymbol{\mu})(\mathbf{W}\bar{x}_n)^{\mathrm{T}} + \mathbf{W}\bar{\mathbf{S}}_n\mathbf{W}^{\mathrm{T}}\big\}\right)_{ii},$$

where $i \in \{1, 2\}$. The $(\cdot)_{11}$ and $(\cdot)_{22}$ notations denote respectively the matrix upper left block of size $D_1 \times D_1$ and the matrix lower right block of size $D_2 \times D_2$.

5. Experimental Results

In this section, we show that the robust probabilistic projection models find the same principal or canonical directions in absence of outliers as the standard (probabilistic) models. In addition, they are also able to lower the effect of atypical observations whenever they occur in the data sets, and are thus able to recover the true principal or canonical directions in this case as well. Note that in all the experiments, the degrees of freedom ν regulates the robustness of the model and is optimized automatically.

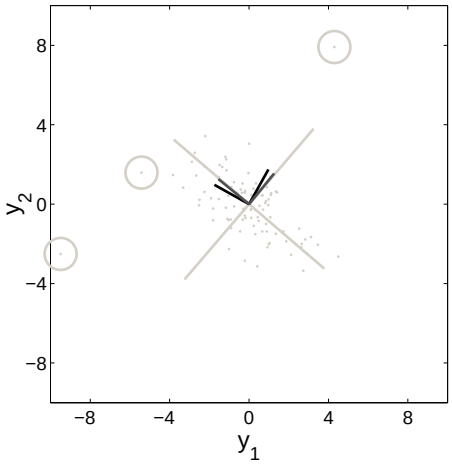


Figure 2. Principal axes found by PPCA (black) or standard PCA, and the ones found by robust PPCA (dark grey) in presence of 3 outliers (uniform random noise in the interval [-10,10] in each direction; indicated by circles). The data is drawn from a multivariate 2-dimensional Gaussian distribution. The light grey lines indicate the principal axes found by PPCA and robust PPCA in absence of outliers.

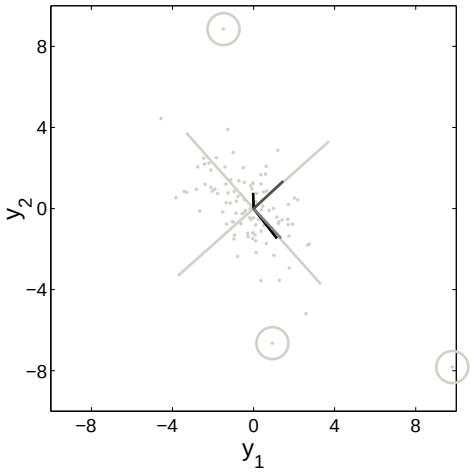


Figure 3. First two principal axes found by standard PCA and PPCA (black) and the ones found by robust PPCA (dark grey) in presence of 3 outliers (uniform random noise in the interval [-10,10] in each direction; indicated by circles). The data is drawn from a multivariate 4-dimensional Gaussian distribution (third and fourth dimension correspond to noise). The light grey lines indicate the principal axes found PPCA and robust PPCA in absence of outliers.

5.1. Experiments with Robust PPCA

Let us first consider a simple 2-dimensional data set of size 100 and which is drawn from a multivariate Gaussian distribution. The data is shown in Figure 2. Only 3 outliers are added. In presence of few outliers, PPCA is not able to recover the principal directions of the data. By contrast, robust PPCA is able to recover the same principal directions as PCA in absence of outliers.

The second data set that we consider is drawn from a 4-dimensional Gaussian distribution. The third and fourth dimensions have little thickness (small fraction of the standard deviation in the other directions) and can thus be viewed as noisy dimensions. Again 3 outliers are added. Figure 3 shows the principal directions in the $\{y_1, y_2\}$-subspace. Observe that here, instead of finding a rotated version of the principal directions, PCA and PPCA do not find the same subspace as in the absence of outliers. By contrast, robust PPCA does.

The third data set is the Old Faithful geyser data (see Figure 4 and 5), which is 2-dimensional. The data is normalized and then 20 outliers are added. In Figure 4, the outliers are drawn from a uniform distribution in the range $[-5, 5]$ in each direction. Again, there is an advantage to use robust PPCA. At first sight, one might think that the outliers have only a limited impact as their number is rather large. However, this is due to the fact that they are distributed in a balanced way around the data. By contrast, Figure 5 shows the situation where this not the case anymore. The outliers are now only located in the right portion of the input space. Clearly, the robust approach outperforms the standard ones.

5.2. Experiments with Robust PCCA

The first example we consider for CCA is such that

$$y_{1n,1} + y_{1n,2} = y_{2n,1} + y_{2n,2}, \tag{38}$$

where $\mathbf{y}_{1n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ and $\mathbf{y}_{2n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$. This data set has a single canonical correlation which is equal to 1 and the canonical directions are both $(1, 1)^{\mathrm{T}}$. Note that we add some Gaussian noise to the data. In order to asses the quality of CCA, we investigate how well the single cannonical correlation is identified. In Figure 6, it can be observed that in absence of outliers standard and robust PCCA find the same canonical direction, which is nicely aligned with the true one. However, when few outliers are added PCCA is not able to recover the true canonical direction, resulting in a projection mismatch.

The second example to illustrate robust probabilistic

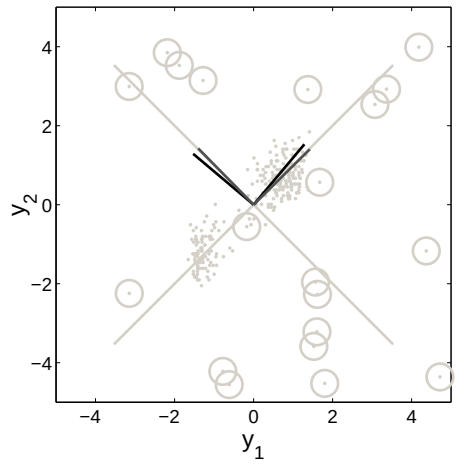


Figure 4. Principal axes found by standard PCA and PPCA (black) and the ones found by robust PPCA (dark grey) in presence of 20 outliers (uniform random noise in the interval [-5,5] in each direction; indicated by circles). The data set is the the normalized 2-dimensional Old Faithful geyser data. The light grey lines indicate the principal axes found by PPCA and robust PPCA in absence of outliers.

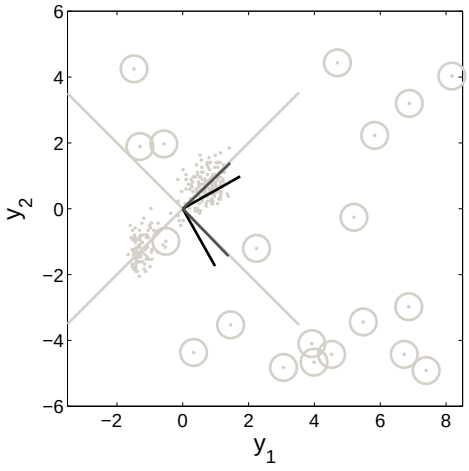


Figure 5. Principal axes found by standard PCA and PPCA (black) and the ones found by robust PPCA (dark grey) in presence of 20 outliers (uniform random noise in the interval [-3.5,8.5] in horizontal direction and [-5,5] in vertical direction; indicated by circles). The data set is the the normalized 2-dimensional Old Faithful geyser data. The light grey lines indicate the principal axes found by standard and PPCA and robust PPCA in absence of outliers.

CCA is the following:

$$\mathbf{x}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2),\ \mathbf{x}_2 = \mathbf{x}_1 + (0, \epsilon_x)^{\mathrm{T}}, \tag{39}$$

$$\mathbf{y}_1 \sim \mathcal{N}(\mathbf{W}_1\mathbf{x}_1, \mathbf{\Psi}_1),\ \mathbf{y}_2 \sim \mathcal{N}(\mathbf{W}_2\mathbf{x}_2, \mathbf{\Psi}_2), \tag{40}$$

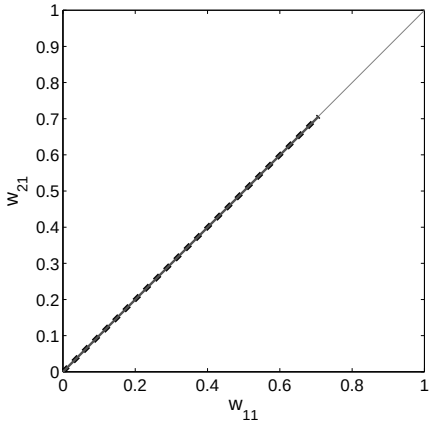
where $\epsilon_x \sim \mathcal{N}(0, \tau_x)$. The projections matrices $\mathbf{W}_1 \in \mathbb{R}^{2\times 2}$, $\mathbf{W}_2 \in \mathbb{R}^{5\times 2}$, as well as the noise matrices $\mathbf{\Psi}_1 \in \mathbb{R}^{2\times 2}$ and $\mathbf{\Psi}_2 \in \mathbb{R}^{5\times 5}$, are arbitrary symmetric positive definite with eigenvalues greater than a specific bound τ_y. Additional atypical points are generated from the independent latent variables $\mathbf{x}'_1, \mathbf{x}'_2 \sim \mathcal{U}(-2, 2)$. We investigate how well the projections from each data space into the low-dimensional subspace match. Figure 7 shows the 2-dimensional projections. Without outliers the models perform similarly. Unlike robust PCCA, standard PCCA is not able to recover the true canonical directions when there are outliers. This results in a projection mismatch (cf. the projections are not nicely aligned with the noiseless projections). Nevertheless, note that part of the correlation is also lost by robust PCCA in the second canonical direction (lower panel).

6. Conclusion

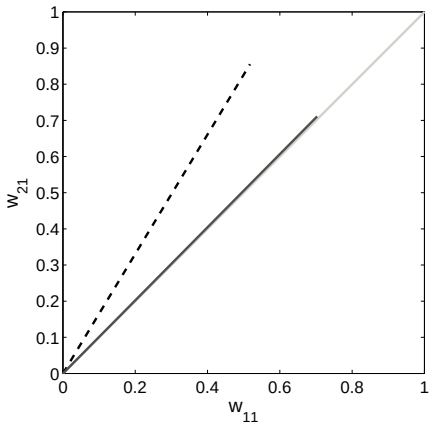
Many probabilistic models rely on a Gaussian assumption. In practice, however, this crude assumption may seem unrealistic as the resulting models are very sensitive to non-Gaussian noise processes. A possible approach is to embed the Gaussian distribution in a wider family of elliptical symmetric distributions, the Student-t distribution.

In this paper, we have shown that robust versions of probabilistic PCA and probabilistic CCA can be constructed based on a Student-t noise distribution. Recently, an increasing number of works have used a similar approach in other contexts (Peel & McLachlan, 2000; Archambeau, 2005). The Student-t distribution enables us to lower the effect of outliers. As a result, the low-dimensional latent subspace is recovered with a higher confidence.

In order to find tractable solutions for the parameters, we view the Student-t distribution as an additional latent variable model and find a local maximum of the likelihood by an EM scheme. The approach works well on several illustrative examples. Of course, the models could be further tested on other real world data sets. Future work includes the extension to mixtures of robust PPCA and PCCA, as well as looking for nonlinear relationships in the data sets with the kernel extension to PCA and CCA.

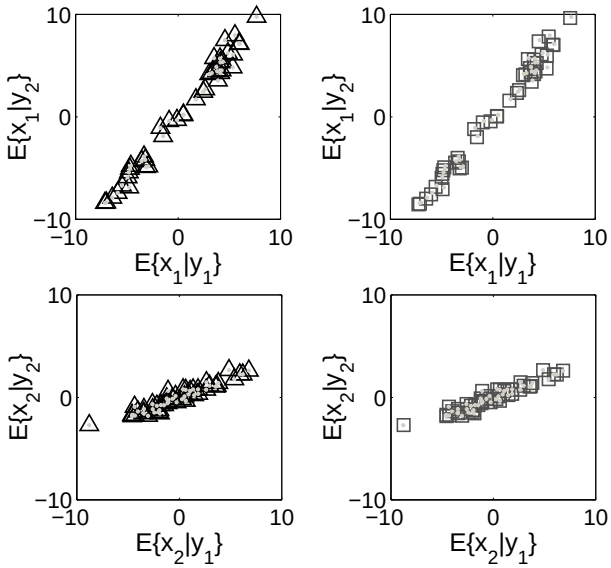


(a) No outliers.

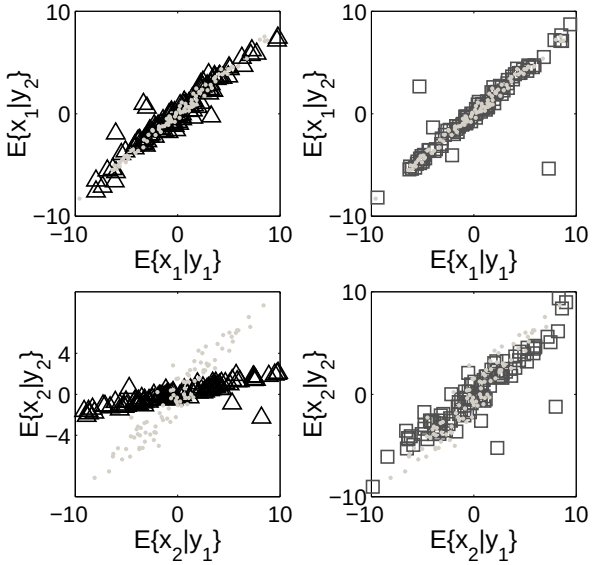


(b) With outliers.

Figure 6. The first canonical correlation obtained for (P)CCA (broken black line) and robust PCCA (dark grey solid line). The light grey line indicates the true cannonical correlation. It can be observed that in presence of outliers PCCA is not able to identify the correct cannonical correlation.



(a) No outliers.



(b) With outliers.

Figure 7. The left panels are the projection obtained for (P)CCA and the right panels are the ones obtained for robust PCCA. The black dots are the projections obtained by standard CCA without outliers (noiseless projections).

Acknowledgements

Nicolas Delannay and Michel Verleysen are respectively Research Fellow and Research Director of the Belgian National Fund for Scientific Research (FNRS).

A. Appendix

CCA is generally resolved by a generalized eigenvalue problem. Computing the reduced singular value decomposition

$$\bar{\mathbf{\Sigma}}_{11}^{-\frac{1}{2}}\bar{\mathbf{\Sigma}}_{12}\bar{\mathbf{\Sigma}}_{22}^{-\frac{1}{2}} = \mathbf{V}_1\mathbf{\Upsilon}\mathbf{V}_2^{\mathrm{T}}, \tag{41}$$

one can show that the canonical direction are given by

$$\begin{cases} \mathbf{U}_{1d} = \bar{\mathbf{\Sigma}}_{11}^{-\frac{1}{2}}\mathbf{V}_1, \\ \mathbf{U}_{2d} = \bar{\mathbf{\Sigma}}_{22}^{-\frac{1}{2}}\mathbf{V}_2. \end{cases} \tag{42}$$

Matrix $\mathbf{\Upsilon}$ is the diagonal matrix of the corresponding correlation coefficients. Working with the probabilistic CCA model, the ML covariance estimates are

$$\bar{\mathbf{\Sigma}}_{11} = \widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_1^{\mathrm{T}} + \mathbf{\Psi}_1^{-1}, \tag{43}$$

$$\bar{\mathbf{\Sigma}}_{12} = \widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_2^{\mathrm{T}}, \tag{44}$$

$$\bar{\mathbf{\Sigma}}_{21} = \widehat{\mathbf{W}}_2\widehat{\mathbf{W}}_1^{\mathrm{T}}, \tag{45}$$

$$\bar{\mathbf{\Sigma}}_{22} = \widehat{\mathbf{W}}_2\widehat{\mathbf{W}}_2^{\mathrm{T}} + \mathbf{\Psi}_2^{-1}. \tag{46}$$

It is possible to identify the canonical direction from these estimates, noting that

$$\begin{aligned} \mathbf{V}_1\mathbf{\Upsilon}^2\mathbf{V}_1^{\mathrm{T}} &= \bar{\mathbf{\Sigma}}_{11}^{-\frac{1}{2}}\widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_2^{\mathrm{T}}(\widehat{\mathbf{W}}_2\widehat{\mathbf{W}}_2^{\mathrm{T}} + \Psi_2)^{-1}\widehat{\mathbf{W}}_2\widehat{\mathbf{W}}_1^{\mathrm{T}}\bar{\mathbf{\Sigma}}_{11}^{-\frac{1}{2}} \\ &= \bar{\mathbf{\Sigma}}_{11}^{-\frac{1}{2}}\widehat{\mathbf{W}}_1(\mathbf{I}_d - \mathbf{B}_2^{-1})\widehat{\mathbf{W}}_1^{\mathrm{T}}\bar{\mathbf{\Sigma}}_{11}^{-\frac{1}{2}} \\ &= \widetilde{\mathbf{V}}_1(\mathbf{I}_d - \mathbf{B}_1^{-1})(\mathbf{I}_d - \mathbf{B}_2^{-1})\widetilde{\mathbf{V}}^{\mathrm{T}} \\ &= \widetilde{\mathbf{V}}_1\mathbf{R}\widetilde{\mathbf{\Upsilon}}^2\mathbf{R}^{\mathrm{T}}\widetilde{\mathbf{V}}_1^{\mathrm{T}} \end{aligned} \tag{47}$$

where we made use of the Woodburry inversion formula and defined $\mathbf{B}_i \equiv \widehat{\mathbf{W}}_i\mathbf{\Psi}_i\widehat{\mathbf{W}}_i^{\mathrm{T}} + \mathbf{I}_d$ and $\widetilde{\mathbf{V}}_1 \equiv \bar{\mathbf{\Sigma}}_{11}^{-\frac{1}{2}}\widehat{\mathbf{W}}_1(\mathbf{I}_d - \mathbf{B}_1^{-1})^{-\frac{1}{2}}$, which is an orthogonal matrix. The matrix $\mathbf{R}$ contains the eigenvectors of $(\mathbf{I}_d - \mathbf{B}_1^{-1})(\mathbf{I}_d - \mathbf{B}_2^{-1})$ and $\widetilde{\mathbf{\Upsilon}}^2$ the corresponding eigenvalues. Identifying the first and the last equalities, we find $\mathbf{V}_1 = \widetilde{\mathbf{V}}_1\mathbf{R}$ and $\mathbf{\Upsilon} = \widetilde{\mathbf{\Upsilon}}$. Finally, doing the same development for $\mathbf{V}_2\mathbf{\Upsilon}^2\mathbf{V}_2^{\mathrm{T}}$, and using formula (42) we find

$$\begin{cases} \mathbf{U}_{1d} = \bar{\mathbf{\Sigma}}_{11}^{-1}\widehat{\mathbf{W}}_1(\mathbf{I}_d - \mathbf{B}_1^{-1})^{-\frac{1}{2}}\mathbf{R}, \\ \mathbf{U}_{2d} = \bar{\mathbf{\Sigma}}_{22}^{-1}\widehat{\mathbf{W}}_2(\mathbf{I}_d - \mathbf{B}_2^{-1})^{-\frac{1}{2}}\mathbf{R}. \end{cases} \tag{48}$$

References

Archambeau, C. (2005). *Probabilistic models in noisy environments and their application to a visual prosthesis for the blind*. Doctoral dissertation, Université catholique de Louvain, Belgium.

Bach, F. R., & Jordan, M. I. (2005). *A probabilistic interpretation of canoncial correlation analysis* (Technical Report 688). Department of Statistics, University of California, Berkeley.

de la Torre, F., & Black, M. J. (2001). Robust principal component analysis for computer vision. *Int. Conf. on Computer Vision* (pp. 362–369).

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

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417–441.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*, 321–377.

Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.

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

Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.), *Learning in graphical models*, 355–368. Kluwer.

Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, *10*, 339–348.

Roweis, S. T. (1998). EM algorithms for PCA and SPCA. *Advances in Neural Information Processining Systems 10*.

Tipping, M. E., & Bishop, C. M. (1999a). Mixtures of probabilistic principal component analyzers. *Neural Computation*, *11*, 443–482.

Tipping, M. E., & Bishop, C. M. (1999b). Probabilistic principal component analysis. *Journal of the Royal Statistical Society B*, *61*, 611–622.

Verbeek, J., Roweis, S., & Vlassis, N. (2004). Nonlinear CCA and PCA by alignment of local models. *Advances in Neural Information Processing Systems 16*.

Xu, L., & Yuille, A. L. (1995). Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Transactions on Neural Networks*, *6*, 131–143.