PCA and Mixtures of PCA:
Improving the robustness to outliers

Joint work with
- Nicolas Delannay, UCL machine learning group
- Cédric Archambeau, University College London

26 October 2007

Overview

- Principal Component Analysis: a reminder
- Probabilistic PCA
- Robust probabilistic PCA
- Mixtures of (robust) probabilistic PCA
- Experiments

Overview

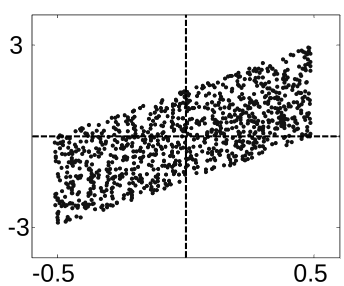
- Principal Component Analysis: a reminder
- Probabilistic PCA
- Robust probabilistic PCA
- Mixtures of (robust) probabilistic PCA
- Experiments

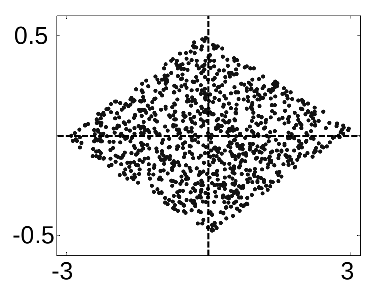
Maximum variance direction [1/2]

- We look for an orthogonal coordinate system such that the elements of **X** in the new system are uncorrelated

=

- We look for axes that maximize variance after projection





Maximum variance direction [2/2]

- We look for axes which maximise the variance after projection

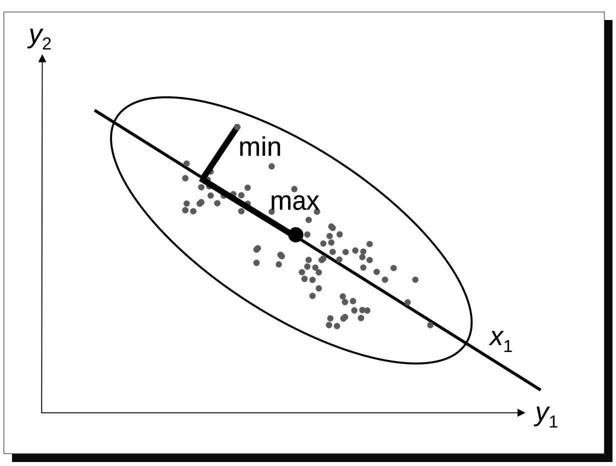
  i.e. along axis **u** with $||\boldsymbol{u}|| = \sqrt{\boldsymbol{u}\boldsymbol{u}^T} = 1$

- In the new coordinate system : $X_1 = \boldsymbol{u}\boldsymbol{Y}$
- Therefore :

$$Var(X_1) = \sigma^2_{X_1} = E[X_1 X_1] = \boldsymbol{u}^T E[\boldsymbol{Y}\boldsymbol{Y}^T]\boldsymbol{u}$$

- Maximum variance = min. distortion (LMS)

Maximum variance = minimum distortion



Choice of direction [1/3]

- Choice of **u** ?

$$\boldsymbol{e} = \underset{\boldsymbol{u}}{\arg\max}\, \boldsymbol{u}^T E[\boldsymbol{YY}^T]\boldsymbol{u}$$
$$= \underset{\boldsymbol{u}}{\arg\max}\, \boldsymbol{u}^T \boldsymbol{C}_{YY}\boldsymbol{u}$$
$$= \underset{\boldsymbol{u}}{\arg\max}\, \boldsymbol{u}^T \underbrace{\boldsymbol{\Theta\Lambda\Theta}^T}_{EVD(\boldsymbol{C_{YY}})} \boldsymbol{u}$$

$$\sigma^2_{X_1,\max} = \boldsymbol{e}^T \underbrace{\boldsymbol{\Theta\Lambda\Theta}^T}_{EVD(\boldsymbol{C_{YY}})} \boldsymbol{e}$$

Choice of direction [2/3]

- Choice of **u** ?

$$\boldsymbol{e} = \underset{\boldsymbol{u}}{\arg\max}\, \boldsymbol{u}^T E[\boldsymbol{YY}^T]\boldsymbol{u}$$

$$\mathbf{e} = \boldsymbol{\Theta}_1$$

where $\boldsymbol{\Theta}_1$ is the eigenvector corresponding to the largest eigenvalue of $\mathbf{C_{YY}}$

$$\sigma^2_{X_1,\max} = \boldsymbol{\Theta}_1{}^T \underbrace{\boldsymbol{\Theta\Lambda\Theta}^T}_{eig(\boldsymbol{C_Y})} \boldsymbol{\Theta}_1 = [1 \quad 0 \quad \dots \quad 0].\begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \lambda_n \end{bmatrix}.[1 \quad 0 \quad \dots \quad 0]^T$$

$$\sigma^2_{X_1,\max} = \lambda_1$$ where $\lambda_1 \ge \lambda_2 \ge \dots \ge \lambda_n$

Choice of direction [3/3]

- Classical result :

Best choice for $\mathbf{u}_1$ is the eigenvector Θ_1 associated to the largest eigenvalue λ_1 of matrix $\mathbf{C}_{\mathbf{YY}}$.

- In the space orthogonal to $\mathbf{u}_1$:

Best choice for $\mathbf{u}_2$ is the eigenvector Θ_2 associated to the largest eigenvalue λ_2 of matrix $\mathbf{C}_{\mathbf{YY}}$.

- And so on ...

Overview

- Principal Component Analysis: a reminder
- Probabilistic PCA
- Robust probabilistic PCA
- Mixtures of (robust) probabilistic PCA
- Experiments

Probabilistic PCA

- Idea #1: **generative probabilistic** model with latent variables ***x***
- Idea #2: **linear dependency** between y_n and x_n
- Idea #3: **Gaussian** distribution of **latent variables**
- Idea #4: **Gaussian** additive **noise**

$$\{y_n\}_{n=1}^{N} \subset \Re^{D}, \quad \{x_n\}_{n=1}^{N} \subset \Re^{J}, \quad J < N$$

$$\boxed{\begin{aligned} P(x) &= \mathrm{N}(x \mid O, I_J) \\ P(y \mid x) &= \mathrm{N}\left(y \mid Wx + \mu, \tau^{-1} I_D\right) \end{aligned}}$$

Probabilistic PCA

$$P(x) = \mathrm{N}(x \mid O, I_J)$$

$$P(y \mid x) = \mathrm{N}\left(y \mid Wx + \mu, \tau^{-1} I_D\right)$$

Isotropic: natural (what else?)

Arbitrary center and variance:
no problem because compensated
by μ and *W*

Gaussians: natural, and
mathematical convenience!

- Marginal distribution:

$$\boxed{P(y) = \int_{X} P(y \mid x) P(x)\, dx = \mathrm{N}(y \mid \mu, \Sigma)}$$

where $\Sigma = WW^T + \tau^{-1} I_D$

Probabilistic PCA training

- Finding parameters $\theta = \{W, \mu, \tau\}$ in

$$P(x) = \mathrm{N}(x|O, I_J)$$
$$P(y|x) = \mathrm{N}(y|Wx + \mu, \tau^{-1} I_D)$$

- How ? By finding the parameters that lead to maximum likelihood of the observations y_n

$$\theta_{\mathrm{ML}} = \arg\max_{\theta}\left(\prod_n P(y_n)\right) = \arg\max_{\theta}\left(\log\left(\prod_n P(y_n)\right)\right)$$
$$= \arg\min_{\theta}\left(-\sum_n \log(P(y_n))\right)$$
$$= \arg\min_{\theta}\left(\sum_n \frac{\tau}{2}(y_n - \mu)^T \Sigma^{-1}(y_n - \mu) + \frac{N}{2}\log 2\pi\tau^{-1}\right)$$

M. Verleysen
UCL
13

Probabilistic PCA training

- How to find θ_{ML} ?

$$\theta_{\mathrm{ML}} = \arg\min_{\theta}\left(\sum_n \frac{\tau}{2}(y_n - \mu)^T \Sigma^{-1}(y_n - \mu) + \frac{N}{2}\log 2\pi\tau^{-1}\right)$$

- $\equiv$ fitting a multivariate Gaussian distribution to the observations, with a constrained covariance matrix
$\Sigma = WW^T + \tau^{-1} I_D$

- How: EM algorithm

M. Verleysen
UCL
14

Probabilistic PCA

- Is it useful?

  (does the probabilistic formulation add something to the deterministic one?)

M. Verleysen
UCL
15

Probabilistic PCA

- Is it useful?

  (does the probabilistic formulation add something to the deterministic one?)

- Answer: no ☹

  Tipping and Bishop showed that the axes found by PPCA span the same subspace as those found by PCA

  (standard equivalence between ML + Gaussian noise hypothesis, and LMS criterion)

M. Verleysen
UCL
16

Probabilistic PCA

- Is it useful?

  (does the probabilistic formulation add something to the deterministic one?)

- Answer: yes ☺

  The probabilistic formulation makes it possible to introduce new, more realistic hypotheses

M. Verleysen
UCL
17

Probabilistic PCA

- Is it useful?

  (does the probabilistic formulation add something to the deterministic one?)

- Answer: yes ☺

  The probabilistic formulation makes it possible to introduce new, more realistic hypotheses

- Ex: LMS criterion (≡ Gaussian hyp.) does not correspond to natural goals!!!

M. Verleysen
UCL
18

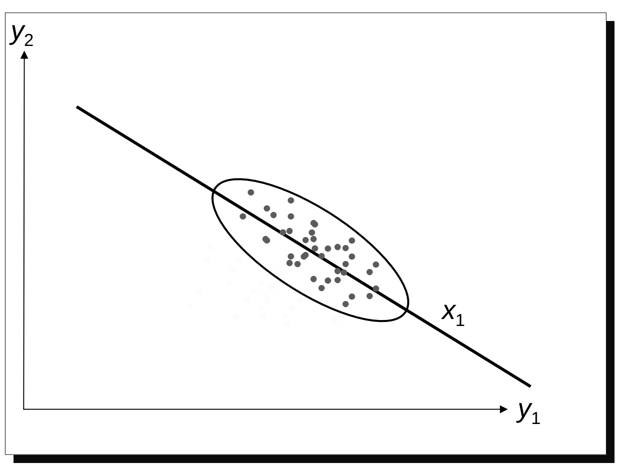
Overview

- Principal Component Analysis: a reminder
- Probabilistic PCA
- Robust probabilistic PCA
- Mixtures of (robust) probabilistic PCA
- Experiments

M. Verleysen
UCL
19

PCA and outliers

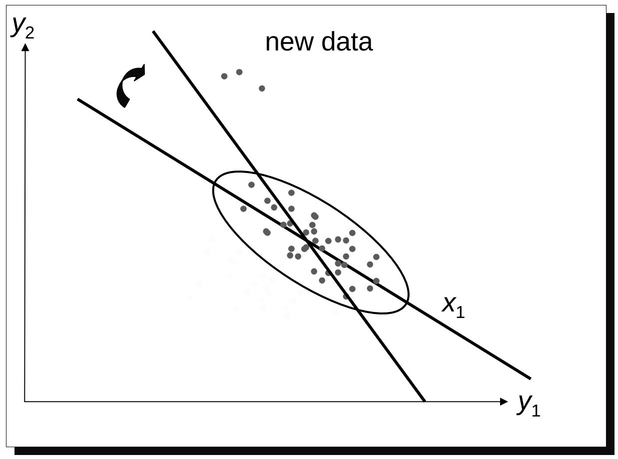
- With “easy” data



M. Verleysen
UCL
20

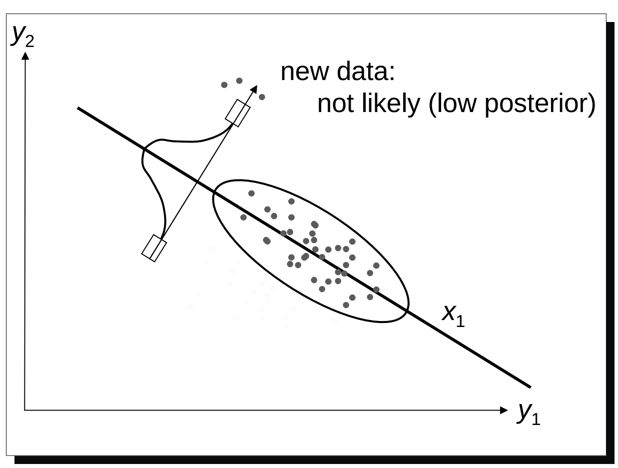
PCA and outliers

- With a few outliers



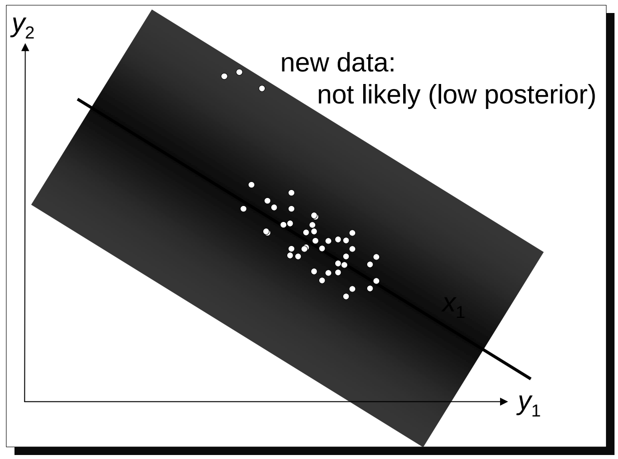
PCA and outliers

- Why is it the case?



PCA and outliers

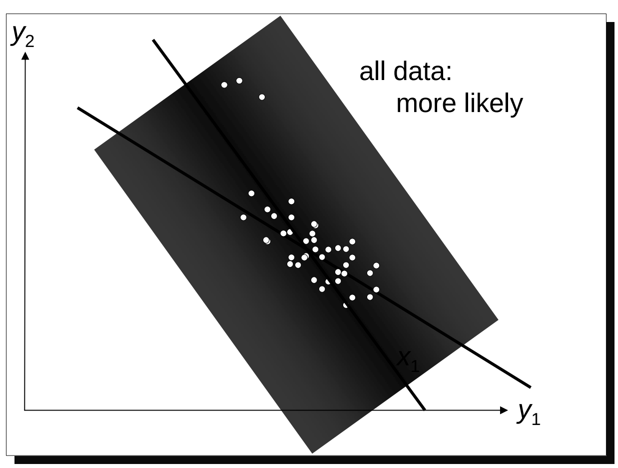
- Why is it the case?



M. Verleysen
UCL
23

PCA and outliers

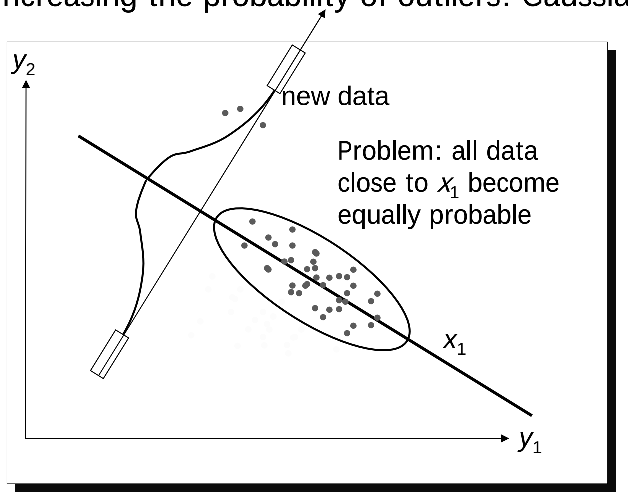
- Why is it the case?



M. Verleysen
UCL
24

PCA robust to outliers

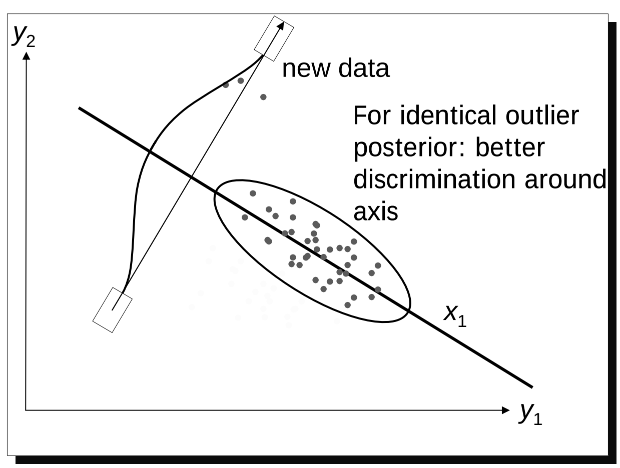
- Increasing the probability of outliers: Gaussian



M. Verleysen
UCL
25

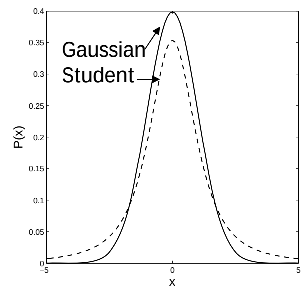
PCA robust to outliers

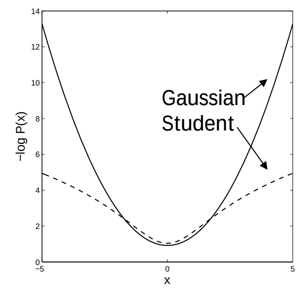
- Increasing the probability of outliers: Student



M. Verleysen
UCL
26

Student-*t* distribution





M. Verleysen
UCL
27

Robust PPCA

- Probabilistic PCA

$$P(x) = \mathrm{N}(x|O, I_J)$$
$$P(y|x) = \mathrm{N}(y|Wx + \mu, \tau^{-1} I_D)$$

- Robust Probabilistic PCA

$$P(x) = \mathrm{St}(x|O, I_J, \nu)$$
$$P(y|x) = \mathrm{St}(y|Wx + \mu, \tau^{-1} I_D, \nu)$$

Identical for simplicity !

M. Verleysen
UCL
28

Robust PPCA

- Model $\begin{cases} P(x) = \mathrm{St}(x | O, I_J, \nu) \\ P(y|x) = \mathrm{St}(y | Wx + \mu, \tau^{-1} I_D, \nu) \end{cases}$
- But Student-*t* = infinite mixture of Gaussians:

  $$\mathrm{St}(y|\mu,\Sigma,\nu) = \int_0^\infty \mathrm{N}\left(y \middle| \mu, \frac{1}{u}\Sigma\right) \mathrm{Ga}\left(u \middle| \frac{\nu}{2}, \frac{\nu}{2}\right) du, \ \nu > 0$$

  where Ga(u|.,.) is a Gamma distribution
- Therefore the generative model is

  $$\begin{aligned} P(u) &= \mathrm{Ga}\left(u \middle| \frac{\nu}{2}, \frac{\nu}{2}\right) \\ P(x|u) &= \mathrm{N}\left(x \middle| O, \frac{1}{u} I_J\right) \\ P(y|x,u) &= \mathrm{N}\left(y \middle| Wx + \mu, \frac{1}{u\tau} I_D\right) \end{aligned}$$

Robust PPCA

- Model

  $$\begin{aligned} P(u) &= \mathrm{Ga}\left(u \middle| \frac{\nu}{2}, \frac{\nu}{2}\right) \\ P(x|u) &= \mathrm{N}\left(x \middle| O, \frac{1}{u} I_J\right) \\ P(y|x,u) &= \mathrm{N}\left(y \middle| Wx + \mu, \frac{1}{u\tau} I_D\right) \end{aligned}$$

- Good news: the posterior is tractable

  $$P(y) = \int_0^\infty \int_X P(y|x,u) P(x|u) P(u) dx = \mathrm{St}(y|\mu,\Sigma,\nu)$$

  where again $\Sigma = WW^T + \tau^{-1} I_D$

Robust PPCA training

- Finding parameters $\theta = \{W, \mu, \tau, \nu\}$ in

$$P(u) = \mathrm{Ga}\left(u \middle| \frac{\nu}{2}, \frac{\nu}{2}\right)$$

$$P(x|u) = \mathrm{N}\left(x \middle| O, \frac{1}{u} I_J\right)$$

$$P(y|x,u) = \mathrm{N}\left(y \middle| Wx + \mu, \frac{1}{u\tau} I_D\right)$$

- How ? By finding the parameters that lead to maximum likelihood of the observations y_n

$$\theta_{\mathrm{ML}} = \arg\min_{\theta}\left(-\sum_n \log(P(y_n))\right)$$

M. Verleysen
UCL
31

Robust PPCA advantages

1. Only one parameter to fix in advance (the dimension of the latent –projection- space)
2. Natural framework for an extension to mixtures

M. Verleysen
UCL
32

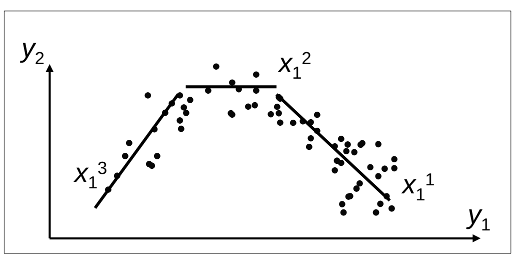
Overview

- Principal Component Analysis: a reminder
- Probabilistic PCA
- Robust probabilistic PCA
- Mixtures of (robust) probabilistic PCA
- Experiments

M. Verleysen
UCL
33

Mixtures of (robust) PPCA

- (P)PCA: linear dependencies in data only
- Mixtures of (P)PCA: nonlinear dependencies, through mixtures of linear manifolds

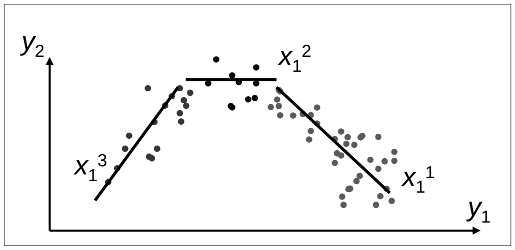


- Latent variable model, *no* common projection space

M. Verleysen
UCL
34

Mixtures of (robust) PPCA

- Concept of (hidden) cluster membership



$$P(y) = \sum_k \pi_k P_k(y)$$

Single PCA model

Mixture proportions $\pi_k > 0$

$$\sum_k \pi_k = 1$$

M. Verleysen
UCL
35

Mixtures of (robust) PPCA

- $z=[z_1, z_2,\ldots, z_K]$ with
  $z_k=1$ if y_n was generated by component k
  $z_k=0$ otherwise
- Latent variable model

$$P(z) = \prod_k \pi_k^{z_k}$$

$$P(u|z) = \prod_k \mathrm{Ga}\left(u_k \middle| \frac{\nu_k}{2}, \frac{\nu_k}{2} \right)^{z_k}$$

$$P(x|u,z) = \prod_k \mathrm{N}\left(x \middle| O, \frac{1}{u_k} I_J \right)^{z_k}$$

$$P(y|x,u,z) = \prod_k \mathrm{N}\left(y \middle| W_k x + \mu_k, \frac{1}{u_k \tau_k} I_D \right)^{z_k}$$

Possible different dimensionalities

M. Verleysen
UCL
36

Mixtures of (robust) PPCA training

- Finding parameters $\theta = \{(W_k, \mu_k, \tau_k, \nu_k, \pi_k)\}_{k=1\ldots K}$ in

$$P(z) = \prod_k \pi_k^{z_k}$$

$$P(u|z) = \prod_k \text{Ga}\left(u_k \middle| \frac{\nu_k}{2}, \frac{\nu_k}{2}\right)^{z_k}$$

$$P(x|u,z) = \prod_k \text{N}\left(x \middle| O, \frac{1}{u_k} I_J\right)^{z_k}$$

$$P(y|x,u,z) = \prod_k \text{N}\left(y \middle| W_k x + \mu_k, \frac{1}{u_k \tau_k} I_D\right)^{z_k}$$

- How ? By finding the parameters that lead to maximum likelihood of the observations y_n

$$\theta_{\text{ML}} = \arg\min_\theta \left(-\sum_n \log(P(y_n))\right)$$

M. Verleysen
UCL
37

Mixtures of (robust) PPCA training

- In practice: EM algorithm
- E-step: fix the model parameters, and compute *expectations* (over an approximate distribution) of latent variables
- M-step: update the model parameters to *maximize* the likelihood

- Two hyper-parameters:
  - the number of components
  - the dimensionality of the latent representations

M. Verleysen
UCL
38

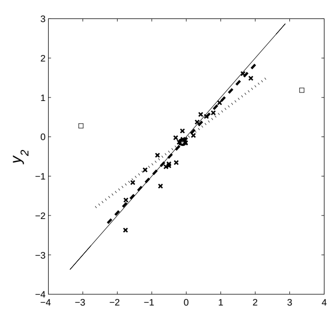
Overview

- Principal Component Analysis: a reminder
- Probabilistic PCA
- Robust probabilistic PCA
- Mixtures of (robust) probabilistic PCA
- Experiments

M. Verleysen
UCL
39

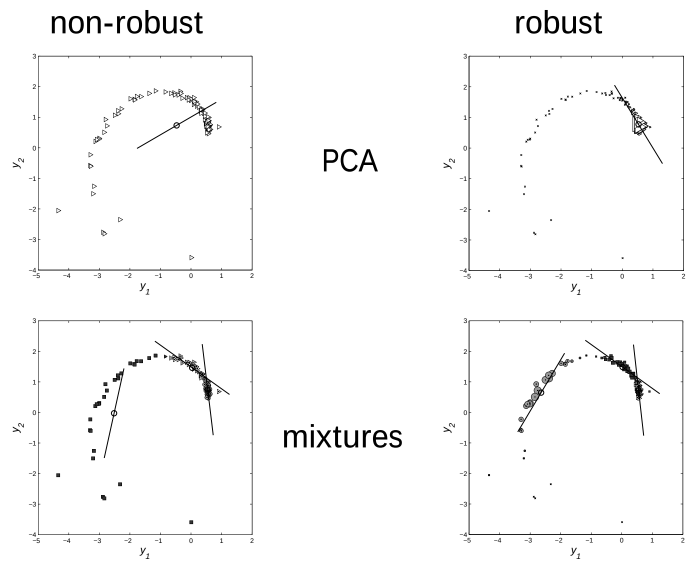
Experiments

- PCA and robust PCA



M. Verleysen
UCL
40

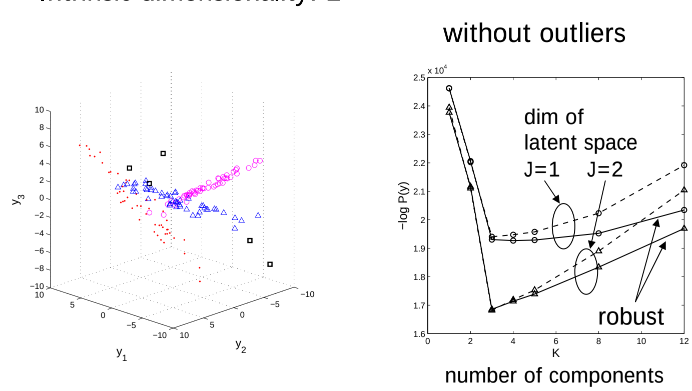
Experiments



M. Verleysen
UCL
41

Experiments

- Three 3-D Gaussian clusters
  - Diagonal covariance matrix: diag(5, 1, 0.2) before rotation
  - Intrinsic dimensionality: 2

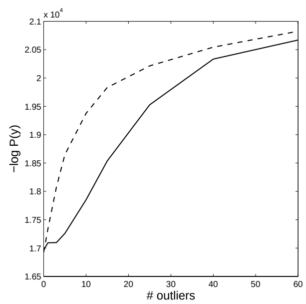


M. Verleysen
UCL
42

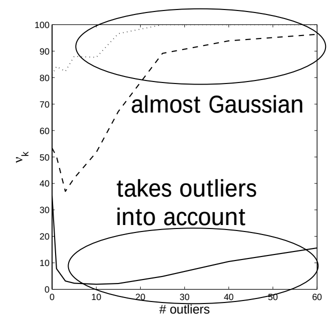
Experiments

- True model (K =3, J =2)
- Outliers

Performances on test set



Degree of freedom parameters



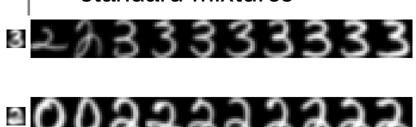
M. Verleysen
UCL
43

Experiments

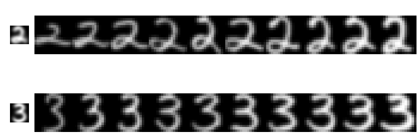
- USPS handwritten digit dataset
  - around 700 images of “2”
  - around 700 images of “3”
  - around 100 images of “0” (as outliers)

- Experiment: two 1-D components (two clusters)
- Illustrations: images close from the 1-D subspaces

standard mixtures



robust mixtures



M. Verleysen
UCL
44

Conclusions

- Probabilistic formulation of PCA:
  - identical to PCA
  - but possible to introduce other hypotheses
- Robust PCA:
  - replacing Gaussians by Student-t
  - makes outliers more likely → lower influence on the model
- Mixtures of PCA
  - latent variable model
  - better than mixtures of Gaussians through regularization (dimensionality of latent space)
- Mixtures of Robust PCA
  - replacing Gaussians by Student-t