# Probabilistic principal components and mixtures – how this works

CISIM 2015 Warsaw , 24– 27 September 2015

.

Anna M. Bartkowiak and Radoslaw Zimroz

University of Wroclaw PL and Technical Wroclaw University PL

# Intro

Keywords

- probabilistic principal components,
- latent (unobservable) variables
- multi-variate normal distribution,
- mixture models,
- un-mixing multivariate data,
- gearbox diagnostics,
- condition monitoring,
- healthy state,
- probabilities a posteriori

Real problem:

Building a model for recognizing healthy state of a gear box on the base of vibration data recorded for that gearbox.

# Mixtures

## The mixture model

We have N $d$-variate data vectors $\mathbf{x} = [x_1, x_2, ..., x_d]^T$ which are supposed co come from $M$ different groups of $d$-variate data.
The mixture model composed from these $M$ groups is defined as (see: Bishop, 1995; Nabney, 2002):

(1) $$p(\mathbf{x}) = \sum_{j=1}^{M} P(j) p(\mathbf{x} \mid j),$$

where:
$p(\mathbf{x} \mid j)$ denotes the probability (probability density function, p.d.f) that a given data vector **x** belongs to group $j,\ j = 1, \ldots, M$,

the $P(j)$ are called *mixing coefficients*; they have the properties:

$\sum_{j=1}^{M} P(j) = 1$, and $0 \leq P(j) \leq 1,\ \ j = 1, \ldots, M.$

## The mixture model – a Bayesian look

The p.d.f given in eq. (1) is called *the total probability density function* .
The mixing coefficients $P(j)$ are called *priors* or *probabilities* $\acute{a}$ *priori* .
The probabilities $p(\mathbf{x} \mid j)$ are called *group conditioned probabilities* .
Assuming normal (Gaussian) distribution, they may written as:

(2)
$$p(\mathbf{x} \mid j) = \frac{1}{(2\pi\sigma_j^2)^{d/2}}\, exp\{-\frac{||\mathbf{x} - \boldsymbol{\mu}_j||^2}{2\sigma_j^2}\}$$

To make inference about group membership of a given $\mathbf{x}$, one uses *posterior probabilities (posteriors)* defined as:

(3)
$$P(j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid j)P(j)}{p(\mathbf{x})}$$

# Modelling the mixture

## Probabilistic principal components

The observed data vector **x** is modelled as a linear combination of some latent variables defined in an (unobservable directly) latent variables space of dimension $q <= d$:

(4) $$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \mathbf{e}.$$

Meaning of symbols appearing in the assumed model:

**x** - the observed $d$-dimensional data vector,
**z** - $q$-dimensional latent factor variable, with **0** mean and unit isotropic variance; **z** is distributed as $N_q(\mathbf{0}, \mathbf{I})$,
**W** - so called projection matrix, of size $d \times q$, consists of constant real numbers playing the role of parameters in the assumed model;
$\boldsymbol{\mu}$ - some constants playing the role of shift parameters (means);
**e** - independent noise process distributed as $N_d(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I})$.

**The above dependency (eq. (4)) is set for each group of the mixture.**

# Modelling the mixture - cont.

## Probabilistic principal components - continued

Taking into account eq. (4), the probability density model for the **probabilistic principal component analysis (PPCA) model** reads:

(5)
$$p(\mathbf{x} \mid \mathbf{z}) = \frac{1}{(2\pi\sigma^2)^{d/2}} \, exp\{-\frac{||\mathbf{x} - \mathbf{W}\mathbf{z} - \boldsymbol{\mu}||^2}{2\sigma^2}\}$$

Tipping and Bishop [1999] have shown how to obtain estimates of the unknown parameters appearing in eq. (5). By integrating out from eq. (5) the latent variables **z**, they got that the distribution of the observed $d$-variate data vector **x** is distributed as

(6)
$$\mathbf{x} \sim N_d(\boldsymbol{\mu}, \mathbf{C}), \text{ where } \mathbf{C} = \mathbf{W}\mathbf{W}^{\mathbf{T}} + \sigma^{\mathbf{2}}\mathbf{I}.$$

This is a multi-variate Gaussian distribution with a specific covariance matrix **C**, which permits to estimate the elements of the matrix **W** pointing to **z**, the vector of the latent variables (see eq. (4)).

# Modelling the mixture - cont.

## Probabilistic principal components - cont.

The shift parameters $\boldsymbol{\mu}$ appearing in eq. (5) and (6) have to be estimated from observed data; the ML method yields here the group means as estimates.
Tipping and Bishop have shown also that the ML method yields the following estimates for the parameters **W** and $\sigma^2$ appearing in the probability model for PPCA shown in eq. (5):

(7) $$\mathbf{W}_{ML} = \mathbf{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R},$$

where $\sigma^2_{ML} = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j$, and $\boldsymbol{\Lambda}_q, \mathbf{U}_q$ contain, up to a rotation matrix **R**, the first $q$ largest eigenvalues and the connected with them eigenvectors of the covariance matrix **C**.
**Hence we are able to estimate all the parameters from the assumed PPCA model considering the observed data $\{\mathrm{x}_1, \ldots, \mathrm{x}_N\}$.**

**Let's go now to true data analysis**

# Modelling healthy gearbox data

**The analyzed data**

The analyzed data are described in Bartelmus and Zimroz [2009]. Observed vibration signal from a healthy gearbox was was cut into segments and processed by Fourier transform (PSD function in Matlab), yielding 15 power spectra variables. For a statistical data analysis, see: Bartkowiak and Zimroz [2011], Zimroz and Bartkowiak [2011, 2013]. A.o., big deviations from MVG model were stated. From the data gathered by Bartelmus and Zimroz we took for our analysis their set B, which we subdivided into two subsets::

**Training data set B500**, a matrix of size size $500 \times 15$; and

**Test data set Bres** , a matrix of size $951 \times 15$.

Each data vector has also an external variable ZWE (denoting load).

In the following we will assume, that the data may be modelled as a mixture of two Gaussians with embedded latent variables $\mathbf{z}$ of dimension q=2. *Each* component of the mixture will have *its own* MVG with probability density function shown in eqs. (5) and (6).

Matlab package **Netlab by Ian Nabney**[2002] will be used for further calculations.

# Computing the mixture model

Since we have a probabilistic model for the data, **we estimate the parameters** using the **maximum likelihood (ML) method**. The log-likelihood for the training data set B500 with rows composed from N=500 data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{500}$ is:

$$\mathcal{L} = \sum_{n=1}^{500} log \sum_{j=1}^{2} P(j) p(\mathbf{x}_n|j)$$

Our log-likelihood $\mathcal{L}$ includes 98 parameters for estimation from data:
$P(j=1), P(j=2)$ – two mixing coefficients (priors)
$\boldsymbol{\mu}_{j=1}, \boldsymbol{\mu}_{j=2}$ – centers, means of the mixture components, 30 params
$covar_{j=1}, covar_{j=2}$ – variances $\sigma_1^2$ and $\sigma_2^2$ appearing in $p(\mathbf{x}_n|j)$ - eq. (6)
$\mathbf{U}_{j=1}, \mathbf{U}_{j=2}$ – two pairs of eigenvectors, needed to estimate **W** in two groups (eq. (5) and (6)), together 15*2*2=60 params,
$\boldsymbol{\lambda}_{j=1}, \boldsymbol{\lambda}_{j=2}$ – eigenvalues needed for estimation of **W** in two groups (eq. (7)), together 4 params.

# Computing the mixture model

Netlab has packed the mixture model containing the Probabilistic PC's into a structure named **mixB500** with fields as shown below:

```
         type: 'gmm'
          nin: 15
     ncentres: 2
   covar_type: 'ppca'
     ppca_dim: 2
       priors: [0.1318 0.8682]
      centres: [2x15 double]
       covars: [8.6744e-004 0.0093]
            U: [15x2x2 double]
       lambda: [2x2 double]
         nwts: 98
```

It was necessary to estimate 98 parameters of the assumed model. This was done, applying the EM algorithm using the size $500 \times 15$ data matrix B500. Surprisingly, it has worked fast and gave reasonable results shown in the structure above.

# Results - Posteriors

As results we show the **posterior probabilities** evaluated from eq. (3). The formula reads:

$$P(j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid j)P(j)}{p(\mathbf{x})}$$

The posteriors $P(j \mid \mathbf{x})$ for the respective vectors **x** were evaluated using the parameters memorized in the structure mixB500 or mixBres (after training the structures with the EM algorithm). The posteriors were evaluated in three variants A, B, C, as specified in the headers of the three figures shown below. They were evaluated separately for groups labelled *j=1* and *j=2*.

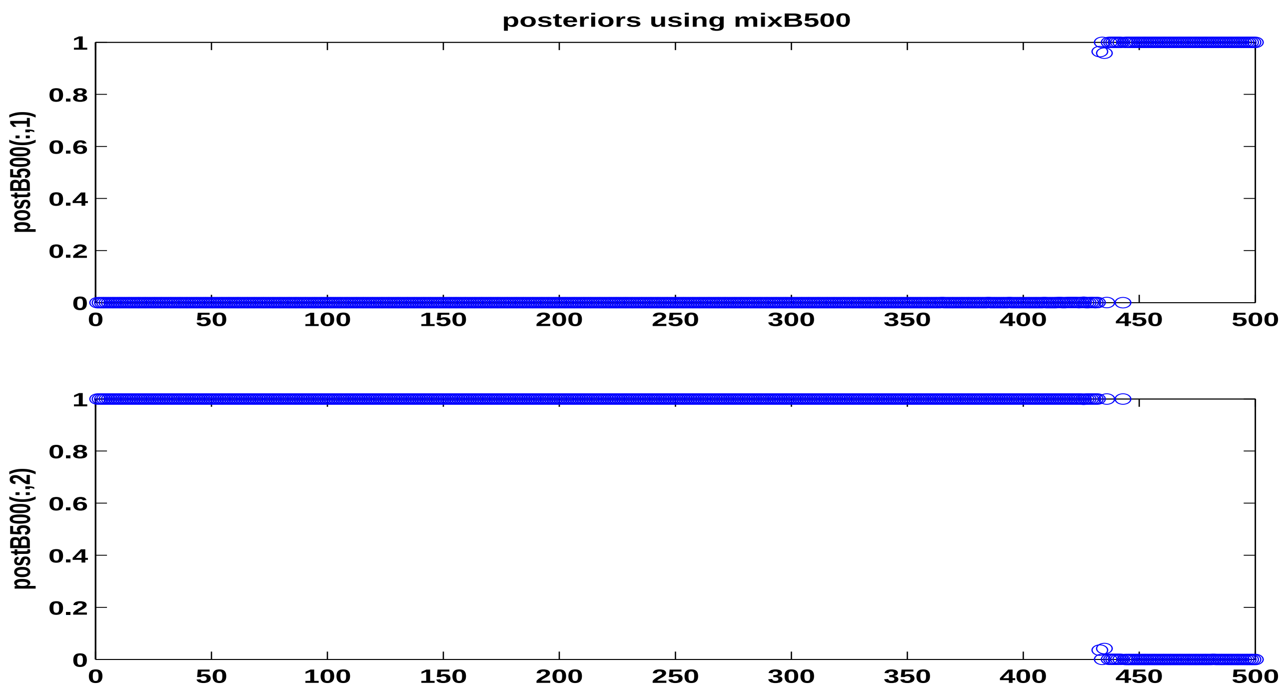
Each posterior may be sought as responsibility of the group j of producing the vector **x** appearing in formula $P(j \mid \mathbf{x})$ (Nabney 1999).

**The data vectors x**, for which the respective posteriors were evaluated, **are ordered according to** the decreasing values of their external variate **ZWE (momentary load of the machine)**.

Each figure is composed from 2 panels, corresponding to the posteriors obtained for group j=1 (top) and j=2 (bottom).

# Posteriors - variant A

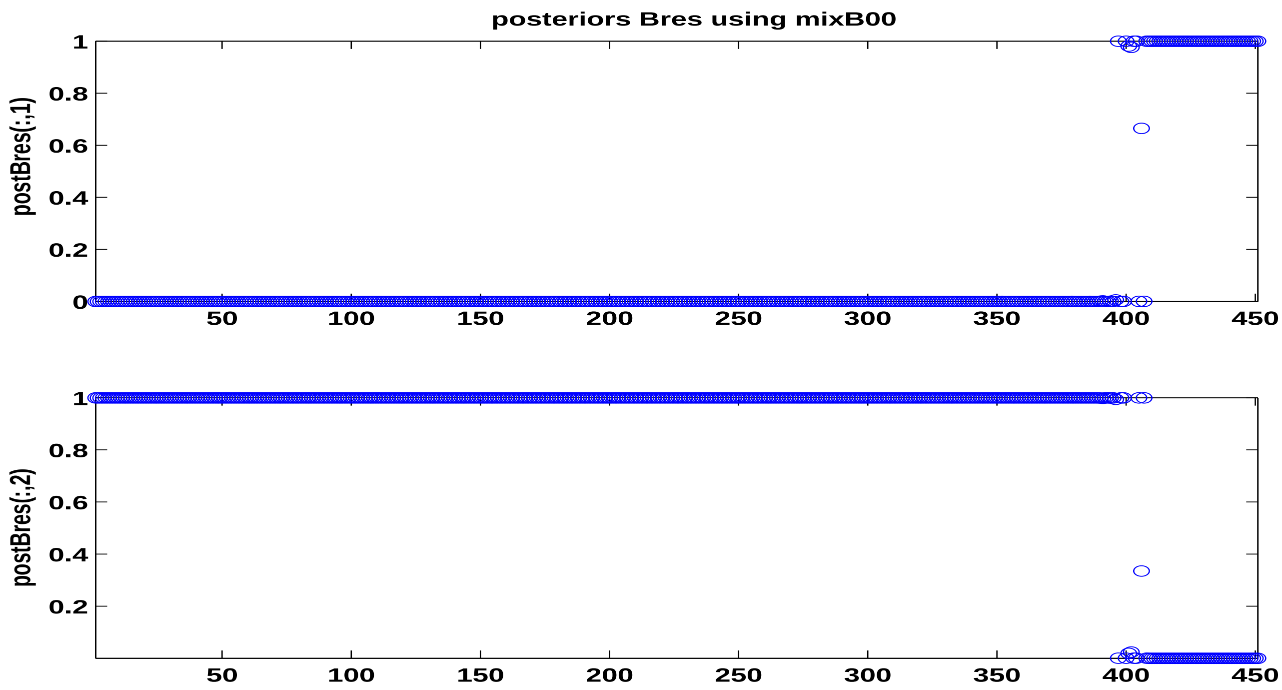
A. Posteriors for data vectors **x** $\in$ B500 training set
Posteriors evaluated from own structure mixB500.




*Top:* probabilities a posteriori $p(j|x)$ of belonging of the data vector **x** to sub-group numbered $j = 1$.
*Bottom:* probabilities a posteriori of belonging of the data vector **x** to sub-group numbered $j = 2$.

# Posteriors - variant B

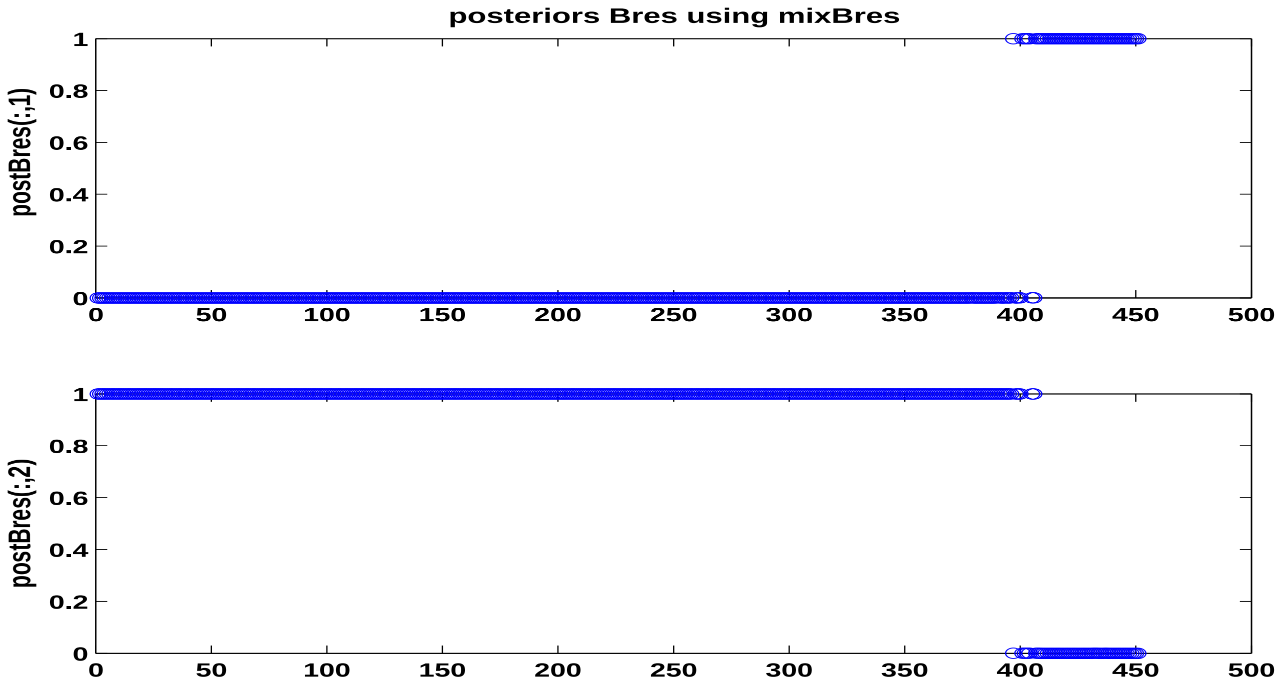
B. Posteriors for data vectors **x** $\in$ Bres test set
Posteriors evaluated from alien structure mixB500.




*Top:* probabilities a posteriori $p(j|x)$ of belonging of the data vector **x** to sub-group numbered $j = 1$.
*Bottom:* probabilities a posteriori of belonging of the data vector **x** to sub-group numbered $j = 2$.

# Posteriors - variant C

C. Posteriors for data vectors **x** $\in$ Bres test set
Posteriors evaluated from own structure mixBres.




*Top:* probabilities a posteriori $p(j|x)$ of belonging of the data vector **x** to sub-group numbered $j = 1$.
*Bottom:* probabilities a posteriori of belonging of the data vector **x** to sub-group numbered $j = 2$.

# Summary

We have considered only a very simple probabilistic principal component mixture model composed from Gaussian rank-2 sub-models with a spherical covariance matrix.

To our surprise, such a very simple model works amazingly well, both for the learning sample B500 and the test sample Bres of the healthy gearbox data. Indeed, we got an un-mixing of the entire data set B into two sub-groups, one of them corresponding to the heavy_load and the other to the light/none_load state of the instances belonging to set B.

Moreover, this was achieved using only sub-models of dimension $q = 2$ (the original data are 15-dimensional).

The computed *probabilities á posteriori* permit to establish a decision boundary delimiting two groups of data characterized by heavy and light/none load.

The *posteriors* assigning the given **x** to the true group:

A. $\mathbf{x} \in$ B500, using mixB500: All *posteriors* are equal to 1.

B. $\mathbf{x} \in$ Bres, using mixB500: All *posteriors* – except for one **x** – equal to 1.

C. $\mathbf{x} \in$ Bres, using mixBres: All posteriors are equal to 1.

Thus the statistical significance of the correct assignments is ultimate high.

# References


**Bartelmus, W., Zimroz, R.:** A new feature for monitoring the condition of gearboxes in non stationary operating systems, Mechanical Systems and Signal Processing 23 (5), 1528-1534 (2009).

**Bartkowiak, A., Zimroz, R.:** Outliers analysis and one class classification approach for planetary gearbox diagnosis. Journal of Physics: Conference Series 305 (1), art. no. 012031 (2011).

**Bishop, C.M.:** Neural Networks for Pattern Recognition.OUP (1995).

**Lei, Y., Lin, J., Zuo, M.J., He, Z.:** Condition monitoring and fault detection of planetary gearboxes: A review. Measurement 48, 292-306 (2014).

**Nabney, I.T.:** NETLAB, Algorithms for Pattern Recognition. Springer, London, Heidelberg (2002).

**Tipping, M.E., Bishop, C.M.:** Probabilistic principal component analysis, J. Roy. Statist. Soc. B 61, 611-622 (1999).

**Zimroz, R., Bartkowiak, A.:** Investigation on spectral structure of gearbox vibration signals by principal component analysis for condition monitoring purposes. Journal of Physics: Conference Series 305 (1), art. no. 012075 (2011).

**Zimroz, R., Bartkowiak, A.:** Two simple multivariate procedures for monitoring planetary gearboxes in non-stationary operating conditions. Mech. Syst. Signal Process. 38(1), 237-247 (2013).


♠ ♣ ♠ *Thanks for attention !* ♠ ♣ ♠