Bayesian sparsity for statistical learning in high dimensions

Charles BOUVEYRON

Professor of Statistics
Chair Inria on "Data Science"

Laboratoire LJAD, UMR CNRS 7351
Equipe Epione, Inria Sophia-Antipolis
Université Côte d'Azur
charles.bouveyron@unice.fr

Joint work with P. Latouche & P.A. Mattei

Disclaimer

"Ce qui est simple est toujours faux.
Ce qui ne l'est pas est inutilisable."

Paul Valéry

Outline

Introduction

In statistical learning, the challenge nowadays is to learn from data which are:

- high-dimensional (p large),
- big or as stream (n large),
- evolutive (evolving phenomenon),
- heterogeneous (categorical, functional, networks, texts, ...)

In any case, the understanding of the results is essential:

- the practitioners are interested in visualizing their data,
- to have a selection of the relevant original variables for interpretation,
- and to have a probabilistic model supposed to have generated the data.

Introduction

Principal component analysis (PCA) is probably the most popular tool of statistical data analysis, with applications in a wide range range of fields:

- psychology: children test results (Hotelling, '33),
- finance: study of volatility dynamics (Egloff et al., '10),
- image processing: from eigenfaces (Turk and Pentland, '91) to deep learning (Chan et al., '15),
- mass spectrometry (Ostrowski et al., '04),
- genomics: DNA microarray data (Rignér, '08).

Introduction

Principal component analysis (PCA) is probably the most popular tool of statistical data analysis, with applications in a wide range range of fields:

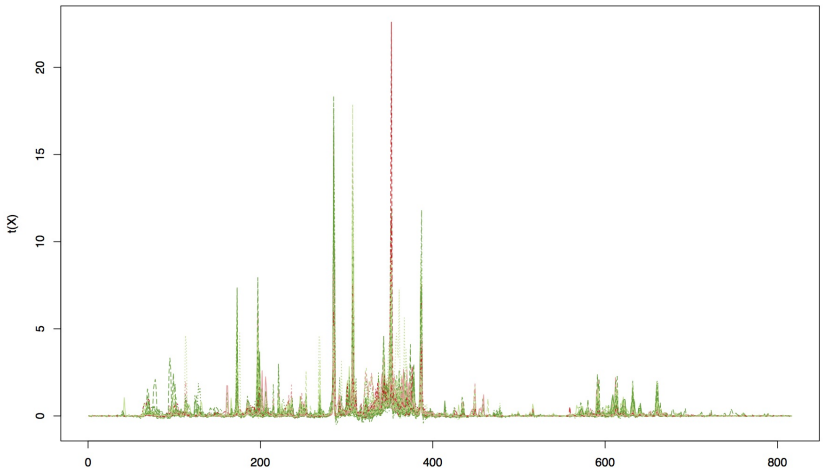
- psychology: children test results (Hotelling, '33),
- finance: study of volatility dynamics (Egloff et al., '10),
- image processing: from eigenfaces (Turk and Pentland, '91) to deep learning (Chan et al., '15),
- mass spectrometry (Ostrowski et al., '04),
- genomics: DNA microarray data (Rignér, '08).

Many modern applications fall into the "ultra-high dimension" case with much more variables than observations ($n << p$) !

A motivating example: NMR spectroscopy

Early prediction of Chronic Kidney Disease from Metabolomics:

- project with Renal Division of Hôpital Européen Georges Pompidou in Paris,
- urine samples from $n = 110$ patients measured with NMR spectroscopy,
- each spectrum is described by $p = 816$ variables.



The goal is to isolate some urinary metabolites (associated with variables) which are early-stage markers of the disease.

Outline

Principal component analysis

Let us consider a $n \times p$ data matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$ that one wants to project onto a "good" d-dimensional subspace.

Principal component analysis (PCA):

- the optimal choice is spanned by the top-d eigenvectors of $\mathbf{X}^T\mathbf{X}$,
- PCA can also be view as a factorization into a low-rank decomposition.

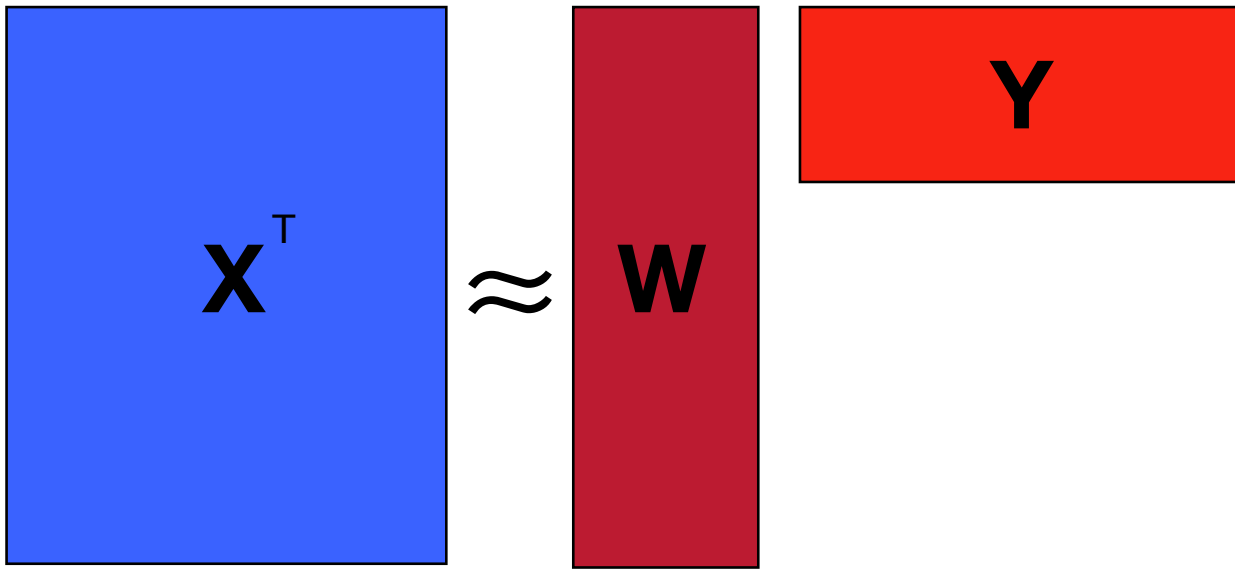


Figure: PCA viewed as a low-rank decomposition.

Sparse principal component analysis

However, regular PCA fails when p is large (Johnstone & Lu '09):

- sparse versions of PCA (SPCA, Zou et al., '06) have beed developed consequently,
- sparse PCA allows to regularize the problem but does not improve significantly the interpretation of the results.

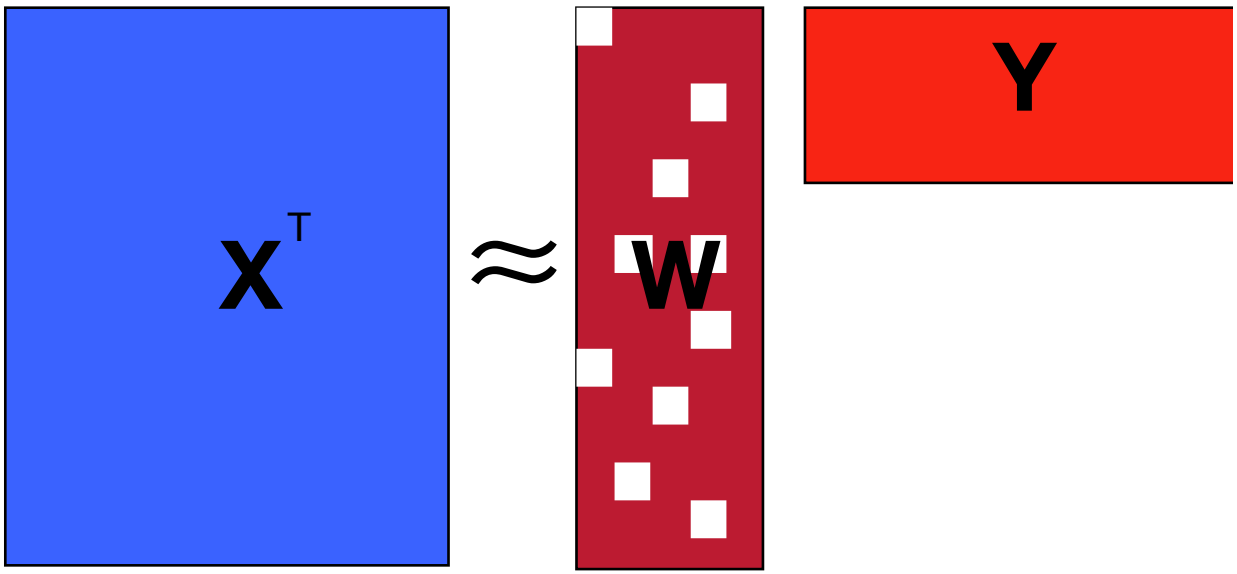


Figure: Sparse PCA viewed as a low-rank decomposition.

Globally sparse principal component analysis

Our objective is to truly perform unsupervised variable selection within PCA:

- the projection matrix W should be row-sparse, leading to the globally sparse PCA problem,
- this solution allows to identify the relevant original variables while reducing the dimensionality.

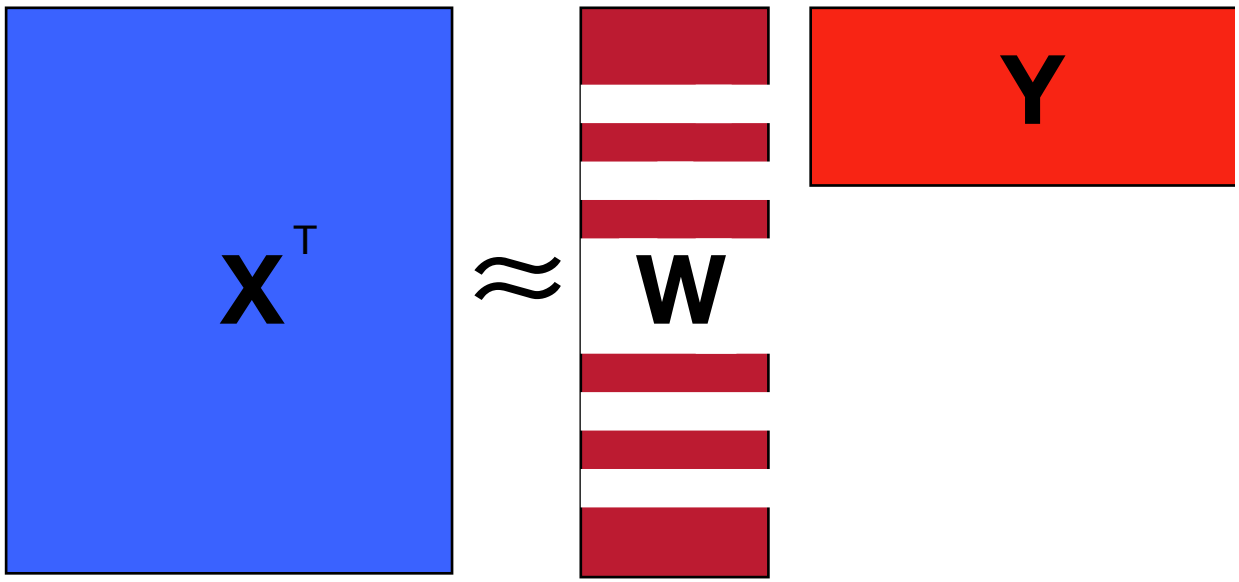


Figure: Globally sparse PCA viewed as a low-rank decomposition.

Outline

Outline

Probabilistic PCA

Let us consider probabilistic PCA (PPCA, Tipping & Bishop, '99) which assumes that each observation is generated by the following model:

$$\mathbf{x} = \mathbf{W}\mathbf{y} + \varepsilon \tag{1}$$

- where $\mathbf{y} \sim \mathcal{N}(0, \mathbf{I}_d)$ is a low-dimensional Gaussian latent vector,
- $\mathbf{W}$ is a $p \times d$ parameter matrix called the *loading matrix,*
- and $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$ is a Gaussian noise term.

Probabilistic PCA

Let us consider probabilistic PCA (PPCA, Tipping & Bishop, '99) which assumes that each observation is generated by the following model:

$$\mathbf{x} = \mathbf{W}\mathbf{y} + \varepsilon \tag{1}$$

- where $\mathbf{y} \sim \mathcal{N}(0, \mathbf{I}_d)$ is a low-dimensional Gaussian latent vector,
- $\mathbf{W}$ is a $p \times d$ parameter matrix called the *loading matrix,*
- and $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$ is a Gaussian noise term.

This model is equivalent to PCA in the following sense:

Theorem [Theobald '75, Tipping & Bishop '99]. If $\mathbf{A}$ is the $p \times d$ matrix of ordered principal eigenvectors of $\mathbf{X}^T\mathbf{X}$ and if $\mathbf{\Lambda}$ is the $d \times d$ diagonal matrix with corresponding eigenvalues, a maximum-likelihood estimator of $\mathbf{W}$ is

$$\mathbf{W}_{\text{ML}} = \mathbf{A}(\mathbf{\Lambda} - \sigma^2 \mathbf{I}_d)^{1/2}. \tag{2}$$

Bayesian variable selection for PPCA

We propose to handle variable selection within PPCA:

$$\mathbf{x} = \mathbf{VWy} + \varepsilon \tag{3}$$

- where $\mathbf{V} = \mathrm{diag}(\mathbf{v})$ such that the matrix $\mathbf{VW}$ is row-sparse, leading to global sparsity,
- the nonzero entries of the binary vector $\mathbf{v} \in \{0, 1\}^p$ correspond to relevant variables,
- and $q = ||\mathbf{v}||_0$ is the number of relevant variables.

Bayesian variable selection for PPCA

We propose to handle variable selection within PPCA:

$$\mathbf{x} = \mathbf{VWy} + \boldsymbol{\varepsilon} \tag{3}$$

- where $\mathbf{V} = \mathrm{diag}(\mathbf{v})$ such that the matrix $\mathbf{VW}$ is row-sparse, leading to global sparsity,
- the nonzero entries of the binary vector $\mathbf{v} \in \{0,1\}^p$ correspond to relevant variables,
- and $q = ||\mathbf{v}||_0$ is the number of relevant variables.

To perform Bayesian model selection:

- we impose Gaussian priors $w_{ij} \sim \mathcal{N}(0, \alpha^{-2})$ on the loadings,
- and chose the hyper-parameters that maximizes the marginal likelihood:

$$p(\mathbf{X}|\mathbf{v}, \alpha, \sigma) = \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{v}, \alpha, \sigma) = \prod_{i=1}^{n} \int_{\mathbb{R}^{p \times d}} p(\mathbf{x}_i|\mathbf{W}, \mathbf{v}, \alpha, \sigma) p(\mathbf{W}) d\mathbf{W}$$

Bayesian variable selection for PPCA

Classical Bayesian approximations are usually used:

- Laplace (Bishop '99, Minka '00),
- variational (Archambeau & Bach, '09).

Bayesian variable selection for PPCA

Classical Bayesian approximations are usually used:

- Laplace (Bishop '99, Minka '00),
- variational (Archambeau & Bach, '09).

In our case, we reach the following expression:

Theorem. The density of $\mathbf{x}$ is given by

$$p(\mathbf{x}|\mathbf{v},\alpha,\sigma) = e^{-\frac{||\mathbf{x}_{\bar{\mathbf{v}}}||_2^2}{2\sigma^2}} \sigma^{q-p}(2\pi)^{-p/2} \\ ||\mathbf{x}_{\mathbf{v}}||_2^{1-q/2} \int_0^\infty \frac{u^{q/2}e^{-\sigma^2 u^2}}{(1+(u/\alpha)^2)^{d/2}} J_{q/2-1}(u||\mathbf{x}_{\mathbf{v}}||_2)du \quad (4)$$

where J_ν is the 1st type Bessel function of order ν.

Bayesian variable selection for PPCA

Classical Bayesian approximations are usually used:

- Laplace (Bishop '99, Minka '00),
- variational (Archambeau & Bach, '09).

In our case, we reach the following expression:

Theorem. The density of $\mathbf{x}$ is given by

$$p(\mathbf{x}|\mathbf{v},\alpha,\sigma) = e^{-\frac{||\mathbf{x}_{\bar{\mathbf{v}}}||_2^2}{2\sigma^2}} \sigma^{q-p}(2\pi)^{-p/2} \\ ||\mathbf{x}_{\mathbf{v}}||_2^{1-q/2} \int_0^\infty \frac{u^{q/2} e^{-\sigma^2 u^2}}{(1+(u/\alpha)^2)^{d/2}} J_{q/2-1}(u||\mathbf{x}_{\mathbf{v}}||_2) du \quad (4)$$

where J_ν is the 1st type Bessel function of order ν.

Problem: the marginal likelihood $p(\mathbf{X}|\mathbf{v},\alpha,\sigma)$ is numerically intractable !

Outline

The noiseless PPCA model

PPCA allows to recover the principal components even in the limit noiseless setting $\sigma \to 0$! (Roweis '98)

In order to obtain a tractable likelihood, we therefore consider the following model (globally sparse PPCA):

$$\mathbf{x} = \mathbf{V}\mathbf{W}\mathbf{y} + \bar{\mathbf{V}}\boldsymbol{\varepsilon_1} + \mathbf{V}\boldsymbol{\varepsilon_2} \tag{5}$$

- $\boldsymbol{\varepsilon_1} \sim \mathcal{N}(0, \sigma_1^2 \mathbf{I}_p)$ is the noise of the inactive variables,
- $\boldsymbol{\varepsilon_2} \sim \mathcal{N}(0, \sigma_2^2 \mathbf{I}_p)$ is the noise of the active variables.

We want to investigate the noiseless case $\sigma_2 \to 0$.

The noiseless PPCA model

In the context of the globally sparse (noiseless) PPCA model, we demonstrate that:

Theorem. In the noiseless limit $\sigma_2 \to 0$, $\mathbf{x}$ converges in probability to a random variable $\tilde{\mathbf{x}}$ whose density is

$$p(\tilde{\mathbf{x}}|\mathbf{v}, \alpha, \sigma_1^2) = \mathcal{N}(\tilde{\mathbf{x}}_{\bar{\mathbf{V}}}|0, \sigma_1 \mathbf{I}_{p-q})\text{Bessel}(\tilde{\mathbf{x}}_{\mathbf{V}}|1/\alpha, (d-q)/2). \quad (6)$$

This theorem allows us to efficiently compute the noiseless marginal log-likelihood defined as

$$\mathcal{L}(\mathbf{X}, \mathbf{v}, \alpha, \sigma_1) = \sum_{i=1}^{n} \log \mathbb{P}(\tilde{\mathbf{x}} = \mathbf{x}_i|\mathbf{v}, \alpha, \sigma_1).$$

Hyperparameter optimization

It remains to propose estimates for hyper-parameters:

- For σ_1: we propose to simply use the ML estimator from the ideal non-noiseless PPCA model, which is the mean of the $p - d$ smallest eigenvalues of $\mathbf{X}^T\mathbf{X}$.
- For α: if $\mathbf{v}$ is known, the regularization parameter can be optimized efficiently using a gradient ascent approach (we proved that the objective function is univariate and concave !).

Hyperparameter optimization

It remains to propose estimates for hyper-parameters:

- For σ_1: we propose to simply use the ML estimator from the ideal non-noiseless PPCA model, which is the mean of the $p - d$ smallest eigenvalues of $\mathbf{X}^T\mathbf{X}$.
- For α: if $\mathbf{v}$ is known, the regularization parameter can be optimized efficiently using a gradient ascent approach (we proved that the objective function is univariate and concave !).

A last (big) issue:

- find the optimal model, we have to find the binary vector $\mathbf{v}$ which has the highest marginal likelihood,
- problem: there are 2^p possible models $\mathbf{v}$!

Our solution:

- relax the model and rank the candidate models,
- compute the marginal likelihood of a family of p nested models.

Outline

The relaxed gsPPCA model

We replace $\mathbf{v}$ by a continuous parameter $\mathbf{u} \in [0,1]^p$. Denoting $\mathbf{U} = \mathrm{diag}(\mathbf{u})$, the relaxed gsPPCA model becomes:

$$\mathbf{x} = \mathbf{UWy} + \boldsymbol{\varepsilon}. \tag{7}$$

The relaxed gsPPCA model

We replace $\mathbf{v}$ by a continuous parameter $\mathbf{u} \in [0,1]^p$. Denoting $\mathbf{U} = \mathrm{diag}(\mathbf{u})$, the relaxed gsPPCA model becomes:

$$\mathbf{x} = \mathbf{UWy} + \varepsilon. \tag{7}$$

We write the marginal log-likelihood as:

$$\log(p(X|\theta)) = \mathcal{L}(q(Z);\theta) + KL(q(Z)||p(Z|X,\theta)),$$

- where $\theta = (\mathbf{u}, \alpha, \sigma)$ and $Z = (Y, W)$ are the latent variables
- $\mathcal{L}(q(Z);\theta) = \int_Z q(Z) \log(p(X,Z|\theta)/q(Z))dZ$ is a lower bound,
- $KL(q(Z)||p(Z|X,\theta)) = -\sum_Z q(Z) \log(p(Z|X,\theta)/q(Z))$ is the KL divergence between $q(Z)$ and $p(Z|X,\theta)$.

The relaxed gsPPCA model

We replace $\mathbf{v}$ by a continuous parameter $\mathbf{u} \in [0,1]^p$. Denoting $\mathbf{U} = \mathrm{diag}(\mathbf{u})$, the relaxed gsPPCA model becomes:

$$\mathbf{x} = \mathbf{UWy} + \varepsilon. \tag{7}$$

We write the marginal log-likelihood as:

$$\log(p(X|\theta)) = \mathcal{L}(q(Z);\theta) + KL(q(Z)||p(Z|X,\theta)),$$

- where $\theta = (\mathbf{u}, \alpha, \sigma)$ and $Z = (Y, W)$ are the latent variables
- $\mathcal{L}(q(Z);\theta) = \int_Z q(Z) \log(p(X,Z|\theta)/q(Z))dZ$ is a lower bound,
- $KL(q(Z)||p(Z|X,\theta)) = -\sum_Z q(Z) \log(p(Z|X,\theta)/q(Z))$ is the KL divergence between $q(Z)$ and $p(Z|X,\theta)$.

The VEM algorithm:

- E step: $\mathcal{L}$ is maximized over q ($\log q_j^*(Z_j) = E_{i\neq j}[\log p(X,Z|\theta)] + c$),
- M step: $\mathcal{L}(q^*(Z), \theta^{old})$ is now maximized over θ

How to come back to the binary problem?

Once the VEM algorithm has converged, we still need to transform the continuous vector $\mathbf{u}$ into a binary one:

- a family of p nested models is built using the order of the coefficients of $\hat{\mathbf{u}}$ as a way of ranking the variables,
- the marginal likelihood of the non-relaxed model (computed using the formula of Theorem 3) is then maximized over α for this family of models,
- the model $\hat{\mathbf{v}}$ with the largest marginal likelihood is kept.

The gsPPCA algorithm

Algorithm 1: GSPPCA algorithm for unsupervised variable selection

Input: data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, dimension of the latent space $d \in \mathbb{N}^*$
Output: sparsity pattern $\mathbf{v} \in \{0,1\}^p$

// VEM algorithm to infer the path of models
Initialize $\mathbf{u}, \alpha, \sigma, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_n, \mathbf{m}_1, ..., \mathbf{m}_p, \mathbf{S}_1, ..., \mathbf{S}_p$ and $\boldsymbol{\Sigma}$;
repeat
  E-step from Proposition 5;
  M-step from equations (13),(14),(15);
until *convergence of the variational free energy*;

// Model selection using the exact marginal likelihood
Compute σ_1 ;
for $k = 1..p$ **do**
  Compute $\mathbf{v}^{(k)}$;
  Find $\alpha_k = \text{argmax}_{\alpha>0}\{\alpha \mapsto \mathcal{L}(\mathbf{X}, \mathbf{v}^{(k)}, \alpha, \sigma_1)\}$ using gradient ascent ;

$q = \text{argmax}_{1 \leq k \leq p} \mathcal{L}(\mathbf{X}, \mathbf{v}^{(k)}, \alpha_k, \sigma_1)$;
$\mathbf{v} = \mathbf{v}^{(q)}$;

Outline

An introductory example

As an educational example:

- we simulated a data set according to model (3),
- with $n = 50$, $p = 30$, $d = 5$ and $q = 10$,
- such that $\mathbf{v} = (\underbrace{1, ..., 1}_{p}, \underbrace{0, ..., 0}_{p-q})$.

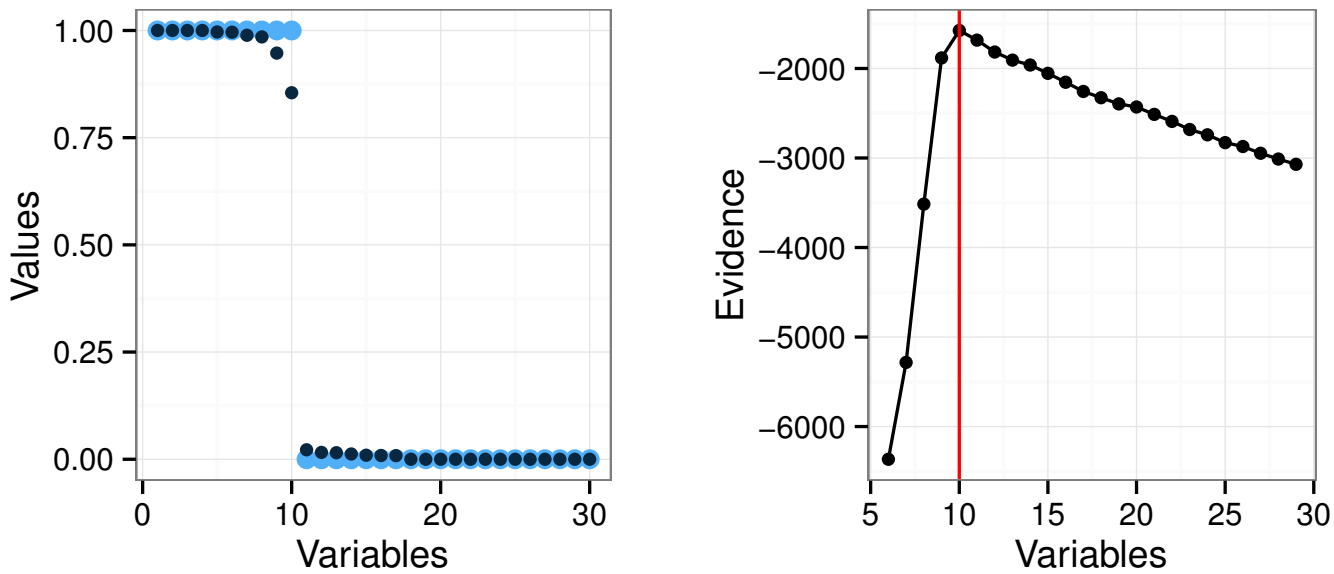


Figure: Variable selection with gsPPCA on the introductory example.

Benchmark: model selection

We here compare gsPPCA with reference methods:

- we simulated a series of data sets with a Toeplitz correlation structure and either a Gaussian or Laplace noise,
- with $p = 200$, $d = 10$, $q = 20$ and varying sample sizes $n = 40, ..., 200$,
- comparison with SPCA (Zou et al., '06) and SSPCA (Jenatton et al., '09).

Table 1: F-score×100 for the model selection experiment of subsection 3.3 with Gaussian noise

	$n = p/5$	$n = p/4$	$n = \lfloor p/3 \rfloor$	$n = p/2$	$n = p$
SPCA	20.7 ± 0.7	21.2 ± 0.7	21.5 ± 0.7	21.7 ± 0.5	25.2 ± 2.1
SSPCA	66.7 ± 21.4	71.5 ± 20	86.7 ± 14.2	95.6 ± 8.9	98.2 ± 7.2
GSPPCA	$\mathbf{86.8 \pm 7.06}$	$\mathbf{93.9 \pm 3.66}$	$\mathbf{97.2 \pm 2.55}$	$\mathbf{99.2 \pm 1.4}$	$\mathbf{1 \pm 0}$

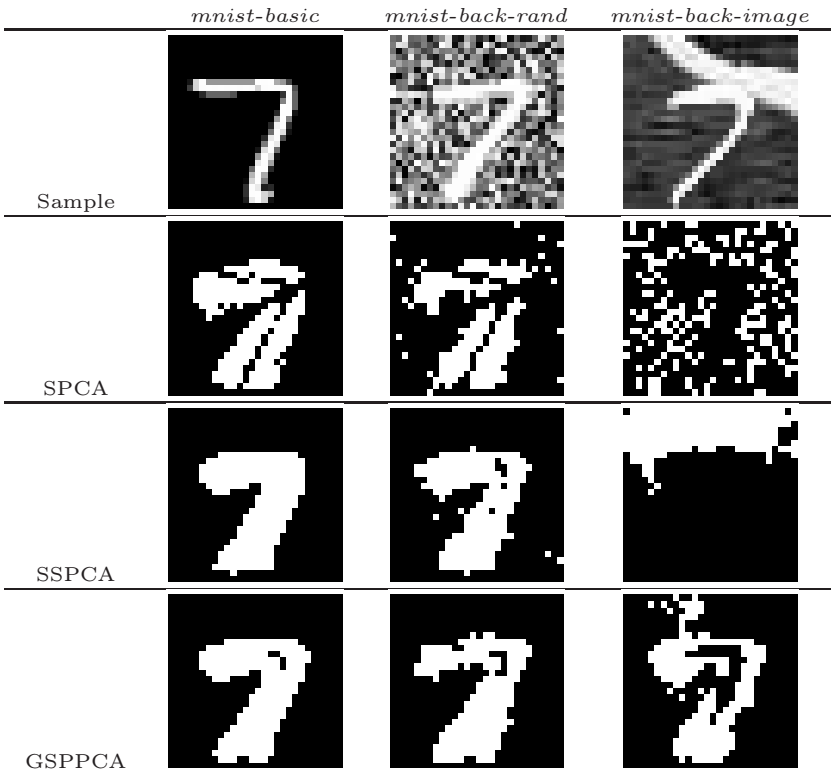
Table 2: F-score×100 for the model selection experiment of subsection 3.3 with Laplacian noise

	$n = p/5$	$n = p/4$	$n = \lfloor p/3 \rfloor$	$n = p/2$	$n = p$
SPCA	20.8 ± 0.6	21.3 ± 0.6	21.6 ± 0.8	21.8 ± 0.6	25.3 ± 1.7
SSPCA	60.6 ± 22.4	63.9 ± 25.2	82.7 ± 18.1	$\mathbf{94.2 \pm 10.2}$	97.4 ± 9.5
GSPPCA	$\mathbf{74.2 \pm 10}$	$\mathbf{77.6 \pm 9.09}$	$\mathbf{79.7 \pm 8.38}$	88 ± 5.95	$\mathbf{99.2 \pm 1.4}$

Benchmark: global vs. local sparsity

Here, we illustrate the difference between global and local sparsity:

- on a set of OCR images from Larochelle et al. (2007), which are variations around the famous MNIST data,
- we have 3 types of images:
  - handwritten 7s,
  - 7s with noisy background,
  - 7s with natural image background,
- in each case, $n = 500$ and $p = 758$.



Outline

Problem and data

We focus here on the diagnostic of chronic kidney disease (CKD):

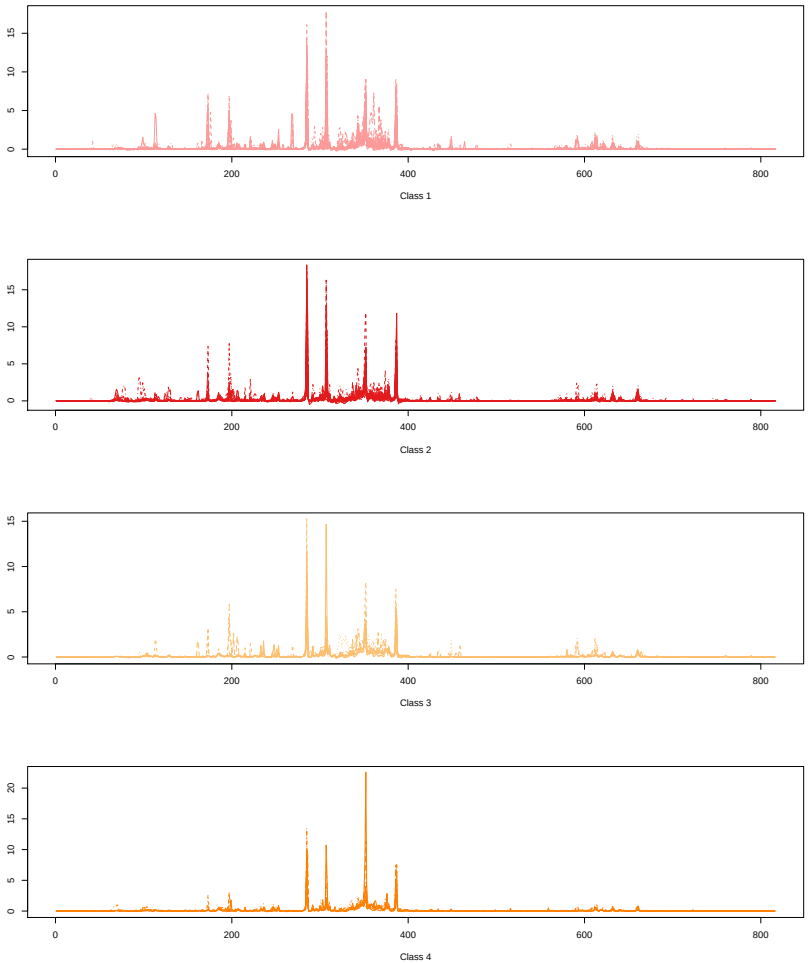
- data come from the Nephrology Department of the Georges Pompidou European Hospital (Paris, France),
- collaboration with Pr. Ph. Beaune, Dr. N. Pallet (Biochimie, HEGP) & Dr. G. Bertho (Plateforme RMN, Paris Descartes),
- the cohort is made of 110 CKD patients followed for renal biopsy between 2013 and 2014,
- for each patient, urine and serum samples were collected.

Beyond the diagnostic problem, the main goal of study is to isolate some urinary metabolites which are early-stage markers of the disease (class #3).

Problem and data

The CKD data set:

- 4 stages of CKD severity are defined according to creatinine rates,
- each urine sample was measured at 300K on a Bruker Avance II spectrometer,
- we end up with 110 spectra of 816 variables split into 4 classes.



Globally sparse HDDA

We introduce a new globally sparse discriminant analysis technique:

- HDDA (Bouveyron et al, '06) is a discriminant analysis method for high-dimensional data which assumes that the data of each class live in a specific low-dimensional subspace,
- gsHDDA combines the idea of gsPPCA and HDDA.

Globally sparse HDDA

We introduce a new globally sparse discriminant analysis technique:

- HDDA (Bouveyron et al, '06) is a discriminant analysis method for high-dimensional data which assumes that the data of each class live in a specific low-dimensional subspace,
- gsHDDA combines the idea of gsPPCA and HDDA.

The gsHDDA model assumes that:

$$Y_{|Z=k} = V_k W_k X + \mu_k + \epsilon,$$

- where $\|\mathrm{diag}(V_k)\|_0 = q_k$, $X_{|Z=k} \sim \mathcal{N}(0, \Delta_k)$ and $\epsilon_{|Z=k} \sim \mathcal{N}(0, \beta_k I_p)$,
- such that $Y_{|Z=k} \sim \mathcal{N}(\mu_k, \Sigma_k)$ where $\Sigma_k = W_k \Delta_k W_k^t + \beta_k I_p$ has a low-rank structure.

Classification results

We first evaluate the diagnostic ability of gsHDDA, in comparison with reference methods:

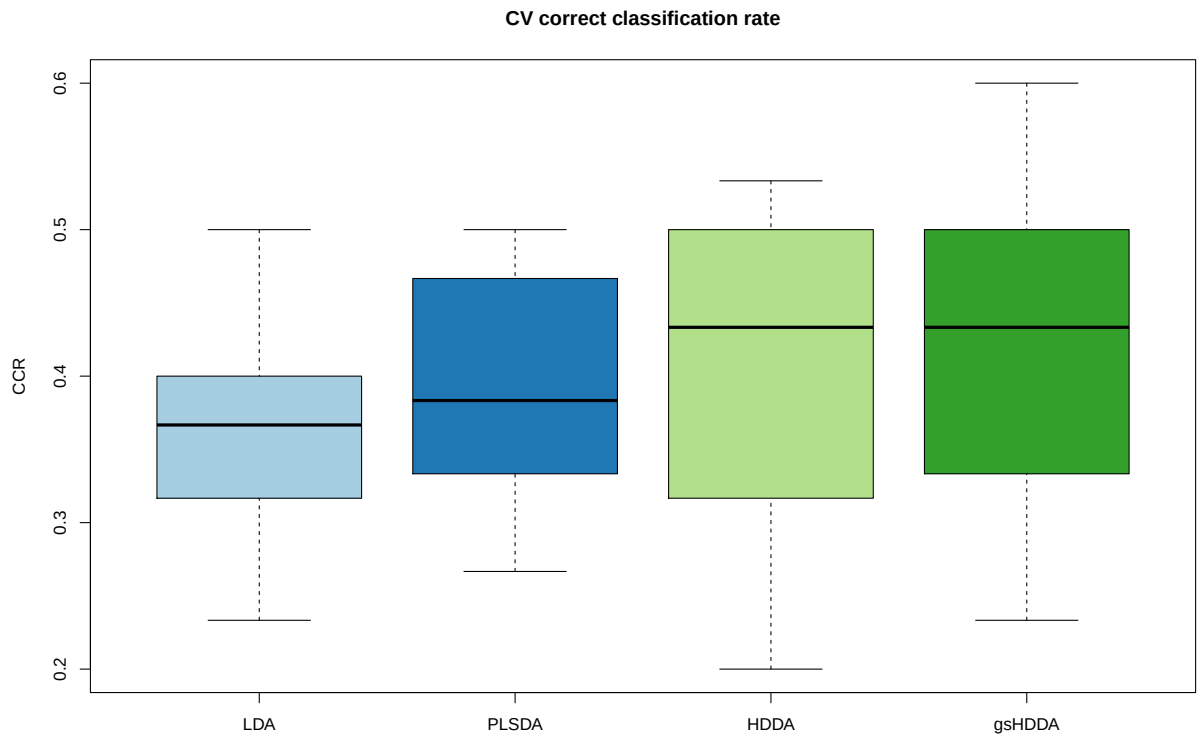


Figure: Correct classification rate for gsHDDA and competitors on CKD data.

Variable selection results

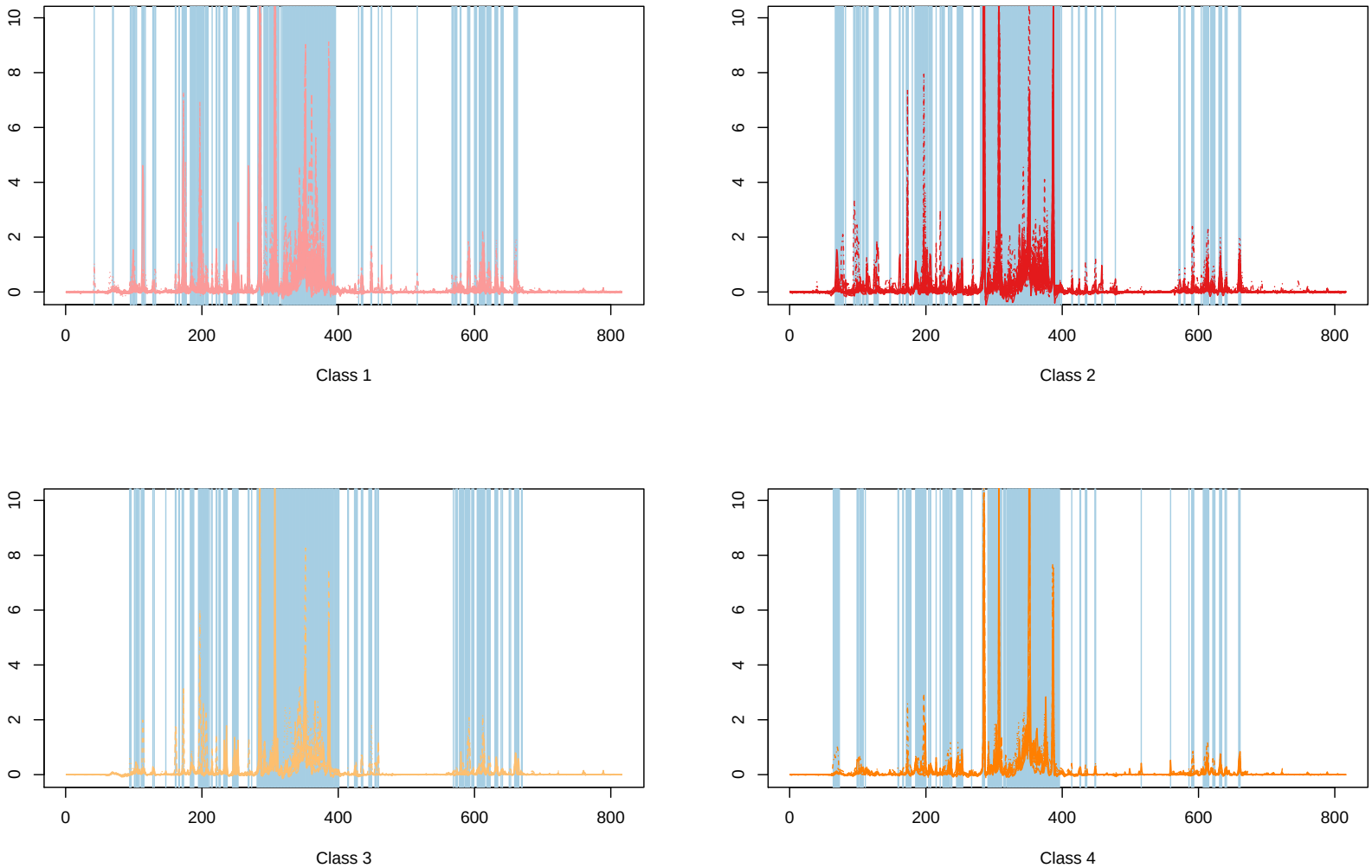


Figure: Variable selection for each class of the CKD data using gsHDDA.

Variable selection results

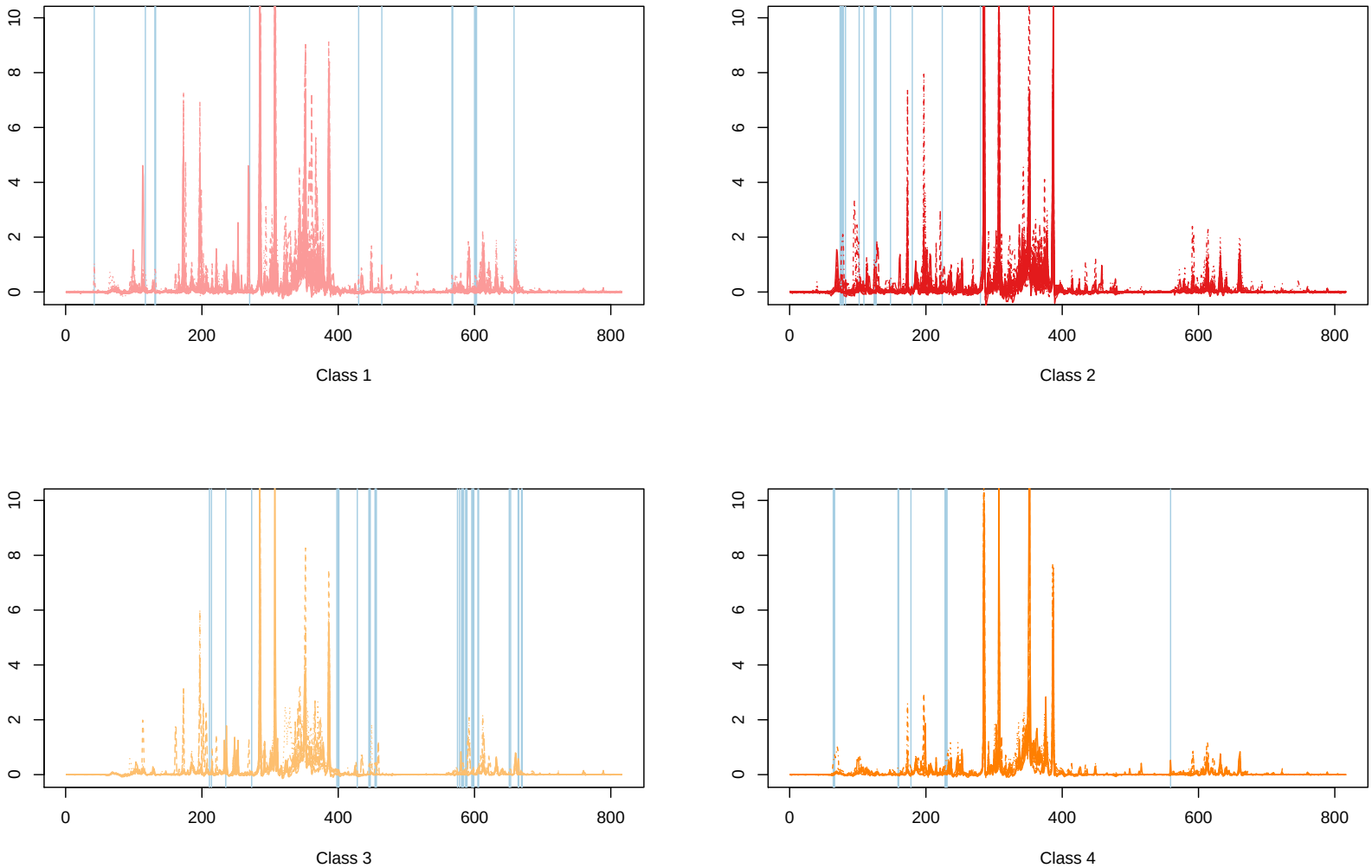


Figure: Specific variable selection for each class of the CKD data using gsHDDA.

Outline

Conclusion and further work

The proposed approach gsPPCA:

- we proposed a Bayesian procedure that allows to obtain several sparse components with the same sparsity pattern,
- this allows the practitioner to identify the original variables which are relevant to describe the data,
- we provided the first exact computation of the marginal likelihood of a Bayesian PCA model,
- a simple relaxation allows to find a path of models using a variational expectation-maximization algorithm.

Working in progress and further work:

- gsHDDA for class-specific variable selection and classification,
- selection of the intrinsic dimensionality d of the data (work in progress),
- mixture of gsPPCAs for clustering of high-dimensional data.

Preprint & R code: *up5.fr/GSPPCA*