Unsupervised Bayesian variable selection

Pierre Latouche
(Pierre-Alexandre Mattei)

Laboratoire MAP5, UMR CNRS 8145
Université Paris Descartes

Joint work with C. Bouveyron

O'Bayes, Warwick

Outline

On variable selection

Local and Global Sparsity for High-Dimensional PCA

Bayesian Variable Selection for PCA

- Global framework
- A closed-form likelihood for Roweis' noiseless PPCA model
- High-dimensional inference through a continuous relaxation

Applications

Conclusion

Variable selection

- consider standard statistical models
- look for prior distributions (+ choice of the parameters)
- $\rightarrow$ look for analytical expressions of the marginal likelihood
- propose algorithms for inference
- define a *path* of models
- optimize the marginal likelihood over the path

The linear model: a noisy linear system

A classical linear regression problem...

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\mathbf{Y} \in \mathbb{R}^n$ is a vector of n observed responses,
$\mathbf{X} \in \mathcal{M}_{n,p}$ is the design matrix with p input variables,
$\boldsymbol{\varepsilon}$ is a stochastic noise term of finite variance.
Goal : estimating $\boldsymbol{\beta} \in \mathbb{R}^p$.

...with a dimensionality issue...

n might be much smaller than p (maximum likelihood becomes is an ill-posed problem).

...and a sparsity assumption.

$\boldsymbol{\beta} \in \mathbb{R}^p$ is sparse (most of its coefficients are null).

Obtaining a sparse solution through penalization

Regularizing the maximum likelihood procedure

$$\hat{\beta}_{\text{penalized}} = \text{argmin}_{\beta \in \mathbb{R}^p} ||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda \text{pen}(\beta),$$

λ is a tuning parameter,
pen is an (often convex) function that penalizes larger models.

Examples

- $\text{pen}(\beta) = ||\beta||_0$ leads to NP-hard problems,
- $\text{pen}(\beta) = ||\beta||_1$ (lasso, Tibshirani, '96) is fast but not necessarily model-consistent,
- $\text{pen}(\beta) = \sum_{i=1}^{p} w_i|\beta|$ (adaptive lasso, Zou, '06) is asymptotically model-consistent,
- $\text{pen}(\beta) = \alpha||\beta||_2^2 + (1-\alpha)||\beta||_1$ (elastic net, Zou & Hastie, '06) can select more variables than the lasso,
- etc.

A model tailored for betting on sparsity

Plugging a sparsity pattern in the linear model

$$\begin{cases} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\beta} &= \mathbf{v} \odot \mathbf{w}, \end{cases}$$

- $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n/\gamma)$ is a Gaussian noise term
- $\mathbf{w} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p/\alpha)$ is a parameter vector with Gaussian prior

Consequence : Spike-and-Slab-like prior on $\boldsymbol{\beta}$

$$p(\boldsymbol{\beta}|\mathbf{v}, \alpha) = \prod_{j=1}^{p} p(\beta_j|v_j, \alpha) = \prod_{j=1}^{p} \delta_0(\beta_j)^{1-v_j} \mathcal{N}(\beta_j; 0, 1/\alpha)^{v_j}$$

(*related to Mitchell and Beauchamp, '88*)

An empirical Bayes framework...

$\mathbf{v}$, γ and α are estimated via maximum marginal likelihood (MML) :

$$(\hat{\mathbf{v}}, \hat{\gamma}, \hat{\alpha}) \in \text{argmax}_{\mathbf{v},\gamma,\alpha} \int_{\mathbb{R}^p} p(\mathbf{Y}|\mathbf{X}, \mathbf{w}, \mathbf{v}, \alpha, \gamma) p(\mathbf{w}|\alpha) d\mathbf{w}.$$

...leads to an automatic penalization of the likelihood

The MML approach implies an Occam factor which, by penalizing larger models, leads to an efficient model selection:

$$-\log p(\mathbf{Y}|\mathbf{v}, \alpha, \gamma) = \frac{\gamma}{2}||\mathbf{Y} - \mathbf{X}_\mathbf{v}\mathbf{m}_\mathbf{v}||_2^2 + \text{pen}(\mathbf{v}, \alpha, \gamma)$$

where

$$\text{pen}(\mathbf{v}, \alpha, \gamma) = \frac{\alpha}{2}\|\mathbf{m}\|_2^2 - \frac{\log \alpha}{2}\|\mathbf{m}\|_0 - \frac{1}{2}\log\det(\gamma \mathbf{X}_\mathbf{v}^T \mathbf{X}_\mathbf{v} + \alpha \mathbf{I}_q) \quad \text{a.s.}$$

On variable selection

Local and Global Sparsity for High-Dimensional PCA

Bayesian Variable Selection for PCA

- Global framework
- A closed-form likelihood for Roweis' noiseless PPCA model
- High-dimensional inference through a continuous relaxation

Applications

Conclusion

Principal component analysis

PCA aims at summarizing multivariate data.

Tons of applications over the last century...

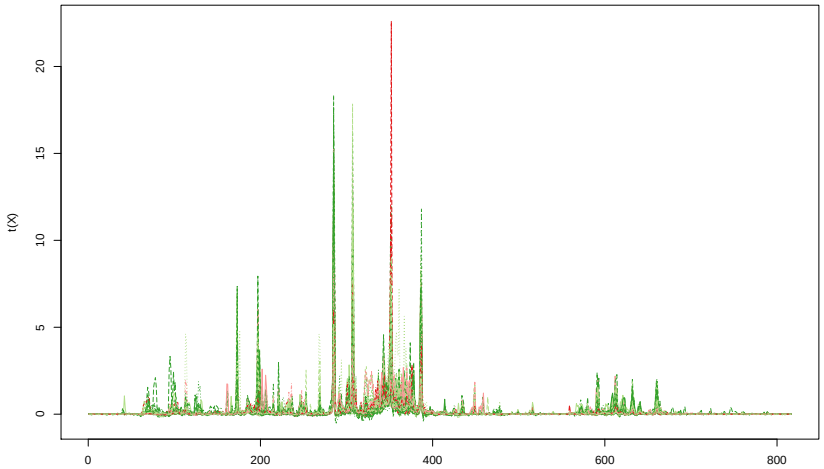
- children test results (Hotelling, '33)
- image processing, from eigenfaces (Turk and Pentland, '91) to deep learning (Chan et al., '15)
- mass spectrometry (Ostrowski et al., '04)
- DNA microarray data (Rignér, '08)

Many modern applications involve cases with much more variables than observations !

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.



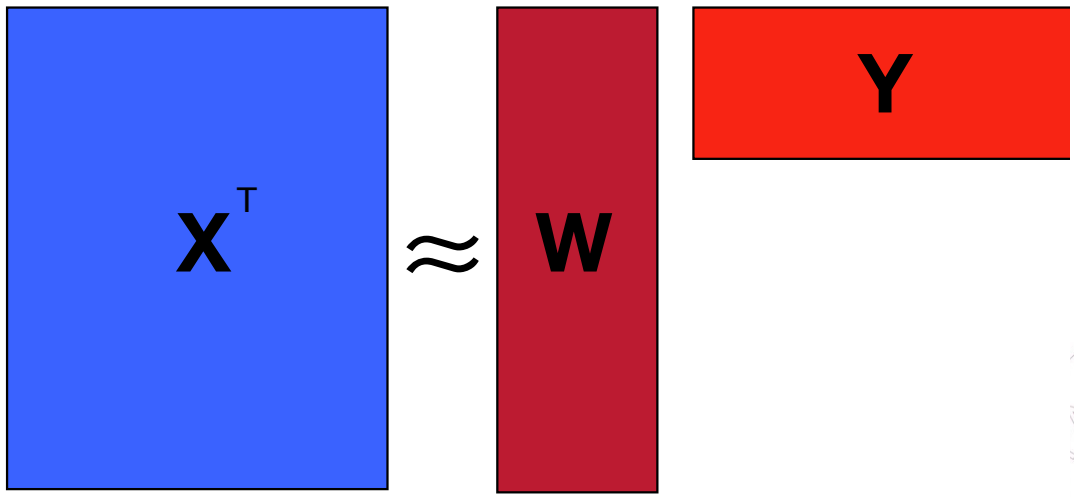
The goal is to isolate some important urinary metabolites (associated with variables) which are early-stage markers of the disease.

Principal component analysis

A $n \times p$ data matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$ is observed.

Goal: project it onto a "good" d-dimensional subspace.

The optimal choice is obtained by spanning the top-d eigenvectors of $\mathbf{X}^T\mathbf{X}$ or by factorizing into a low-rank decomposition

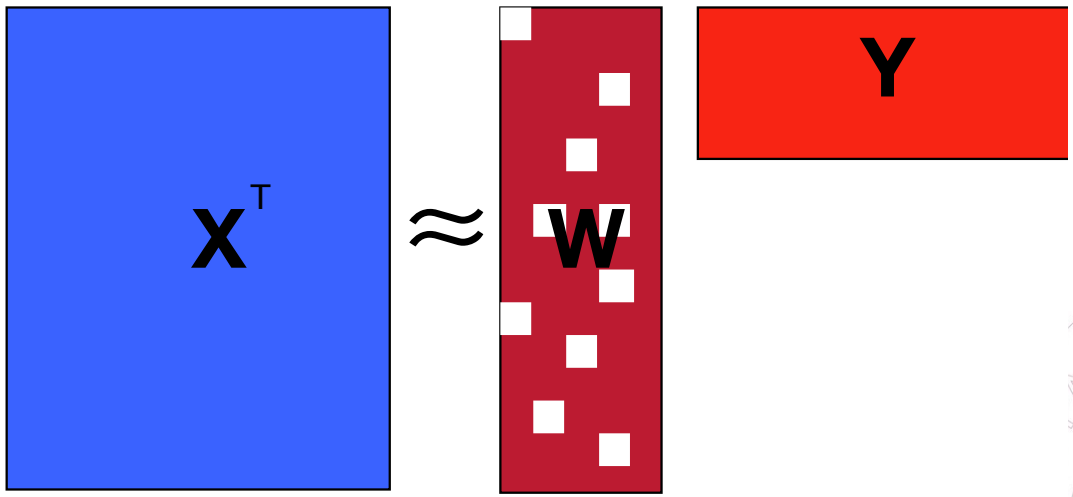


(Locally) Sparse Principal Component Analysis

A $n \times p$ data matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$ is observed.

Goal: project it onto a "good" d-dimensional subspace.

But regular PCA fails when p is large (Johnstone & Lu '09). Sparse versions of PCA have beed developed consequently.

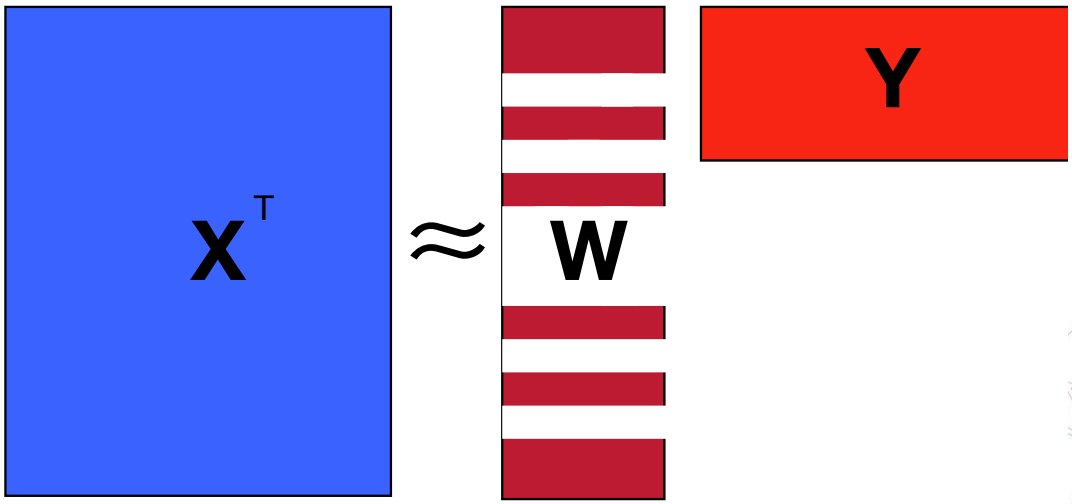


Globally Sparse Principal Component Analysis

A $n \times p$ data matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$ is observed.

Goal: project it onto a "good" d-dimensional subspace.

To truly perform unsupervised variable selection, the projection matrix $\mathbf{W}$ has to be row-sparse, leading to the globally sparse PCA problem.



Probabilistic PCA

PPCA assumes that each observation is driven by the following generative 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 PCA

We handle variable selection using a binary vector $\mathbf{v} \in \{0,1\}^p$ whose nonzero entries correspond to relevant variables. $q = ||\mathbf{v}||_0$ is the number of relevant variables. Consider the model

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

where $\mathbf{V} = \text{diag}(\mathbf{v})$, the matrix $\mathbf{VW}$ is row-sparse, leading to global sparsity.

To perform Bayesian model selection, we use Gaussian priors $w_{ij} \sim \mathcal{N}(0, 1/\alpha^2)$ and chose the hyperparameters that maximizes the marginal likelihood.

$$p(\mathbf{X}|\mathbf{v}, \alpha, \sigma)$$

Bayesian variable selection for PCA

Problem: The marginal likelihood $p(\mathbf{X}|\mathbf{v}, \alpha, \sigma)$ appears to be intractable !

Theorem (Bouveyron, Latouche & Mattei '16)

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

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

Classical Bayesian approximations are usually used: Laplace (Bishop '99, Minka '00), variational (Archambeau & Bach, '09)...

Is it possible to play with the PPCA model to obtain a tractable likelihood ?

Probabilistic PCA *à la* Roweis

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 consider the following model:

$$\mathbf{x} = \mathbf{V}\mathbf{W}\mathbf{y} + \cancel{\mathbf{V}}\boldsymbol{\varepsilon}_1 + \mathbf{V}\boldsymbol{\varepsilon}_2 \tag{5}$$

- $\boldsymbol{\varepsilon}_1 \sim \mathcal{N}(0, \sigma_1^2 \mathbf{I}_p)$ 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 $\sigma_2 \to 0$.

Theorem (Bouveyron, Latouche & Mattei '16)

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

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

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

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

Probabilistic PCA *à la* Roweis - hyperparameter optimization

For σ_1: what appears to work best is 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 gradient ascent. The objective function is univariate and concave !

Probabilistic PCA *à la* Roweis - model selection

To find the optimal model, we have to find the binary vector **v** which has the marginal likelihood.

Problem: there are 2^p possible models **v** !

Our solution is to only compute the marginal likelihood of a family of p nested models.

The relaxed model

We replace $\mathbf{v}$ by a continuous parameter $\mathbf{u} \in [0,1]^p$. Denoting $\mathbf{U} = \text{diag}(\mathbf{u})$, and $\boldsymbol{\theta} = (\mathbf{u}, \alpha, \sigma)$, this can be written

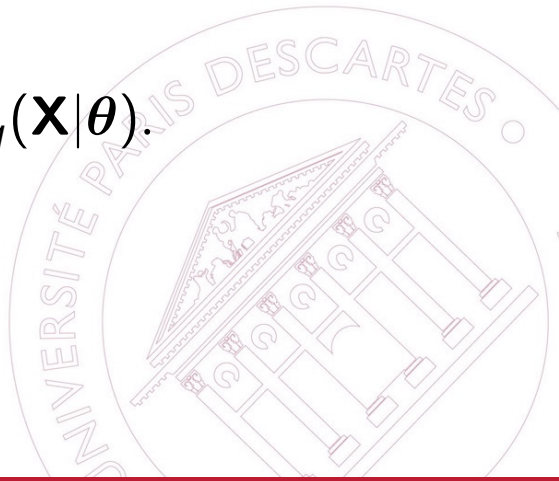
$$\mathbf{x} = \mathbf{UWy} + \varepsilon. \tag{7}$$

We follow a variational approach to minimize the free energy

$$\mathcal{F}_q(\mathbf{x_1}, ...\mathbf{x_n}|\boldsymbol{\theta}) = -\mathbb{E}_q[\ln p(\mathbf{X}, \mathbf{Y}, \mathbf{W}|\boldsymbol{\theta})] - H(q) \tag{8}$$

which is an upper bound to the negative log-evidence:

$$-\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta}) - \text{KL}(q||p(\cdot|\boldsymbol{\theta})) \leq \mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta}).$$



A VEM algorithm for the relaxed model - E step

If we make the mean-field approximation $q(\mathbf{Y}, \mathbf{W}) = q(\mathbf{Y})q(\mathbf{W})$, the variational posterior distribution of the latent variables which minimizes the free energy is given by

$$q^*(\mathbf{Y}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{y}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \;\; \textit{and} \;\; q^*(\mathbf{W}) = \prod_{k=1}^{p} \mathcal{N}(\mathbf{w}_k|\boldsymbol{m}_k, \mathbf{S}_k) \tag{9}$$

where, for all $i \in \{1, ..., n\}$ and $k \in \{1, ..., p\}$,

$$\boldsymbol{\mu}_i = \frac{1}{\sigma^2}\boldsymbol{\Sigma}\mathbf{M}^T\mathbf{U}\mathbf{x}_i, \; \mathbf{m}_k = \frac{u_k}{\sigma^2}\mathbf{S}_k \sum_{i=1}^{n} x_{i,k}\boldsymbol{\mu}_i,$$

$$\boldsymbol{\Sigma}^{-1} = \mathbf{I}_d + \frac{1}{\sigma^2}\mathbf{M}^T\mathbf{U}^2\mathbf{M} + \frac{1}{\sigma^2}\sum_{k=1}^{p} u_k^2\mathbf{S}_k, \; \mathbf{S}_k^{-1} = \alpha^2\mathbf{I}_d + \frac{nu_k^2}{\sigma^2}\boldsymbol{\Sigma} + \frac{u_k^2}{\sigma^2}\boldsymbol{\mathcal{M}}^T\boldsymbol{\mathcal{M}},$$

$$\mathbf{M} = (\mathbf{m}_1, ...\mathbf{m}_p)^T \;\; \textit{and} \;\; \boldsymbol{\mathcal{M}} = (\boldsymbol{\mu}_1, ...\boldsymbol{\mu}_n)^T.$$

A VEM algorithm for the relaxed model - M step

Minimizing the free energy leads to the following M-step updates:

$$\alpha^* = \left(\frac{1}{dp} \sum_{k=1}^{p} \mathrm{Tr}(\mathbf{S}_k + \mathbf{m}_k \mathbf{m}_k^T) \right)^{-1/2}, \tag{10}$$

$$\sigma^* = \sqrt{\frac{\mathrm{Tr}(\mathbf{X}\mathbf{X}^T + \mathbf{X}\mathbf{U}\mathbf{M}\mathcal{M})}{np} + \frac{1}{np} \sum_{i=1}^{n} \sum_{k=1}^{p} u_k^2 \mathrm{Tr}[(\mathbf{\Sigma} + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T)(\boldsymbol{S}_k + \boldsymbol{m}_i \boldsymbol{m}_i^T)]}, \tag{11}$$

and, for $k \in \{1, ..., p\}$,

$$u_k^* = \mathrm{argmin}_{u \in [0,1]} \frac{u^2}{2\sigma^2} \sum_{i=1}^{n} \mathrm{Tr}[(\mathbf{\Sigma} + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T)(\boldsymbol{S}_k + \boldsymbol{m}_i \boldsymbol{m}_i^T)] - u \sum_{i=1}^{n} x_{i,k} \mathbf{m}_k^T \boldsymbol{\mu}_i. \tag{12}$$

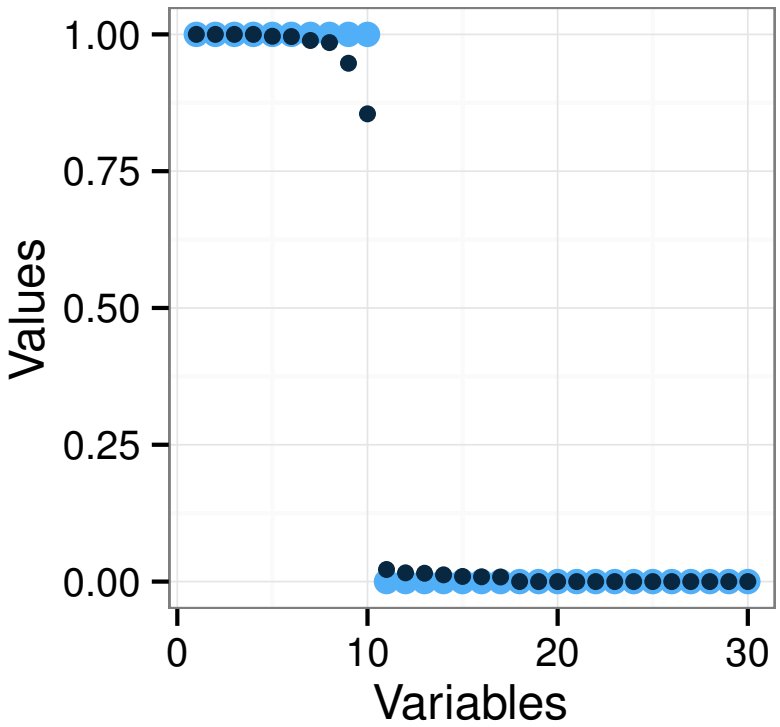
How to reverse the relaxation ?

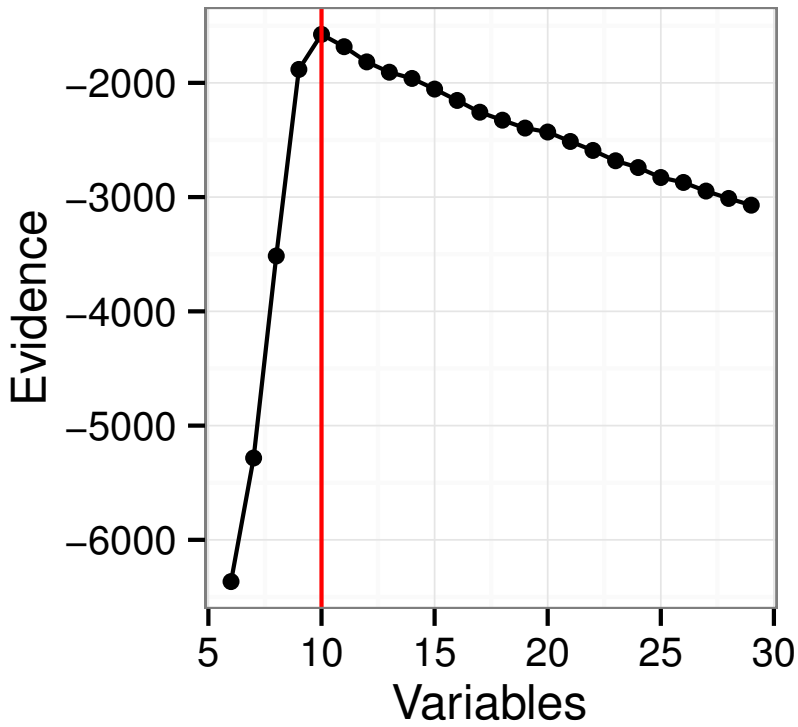
Once the VEM algorithm has converged, we still need to transform the continuous vector $\mathbf{u}$ into a binary one:

- 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 2) is then maximized over this family of models.
- the model $\hat{\mathbf{v}}$ with the largest marginal likelihood is kept.

Relaxation - binarization in practice

A simple example with $\mathbf{W}$ Gaussian, $p = 30$, $d = 5$, $q = 10$.





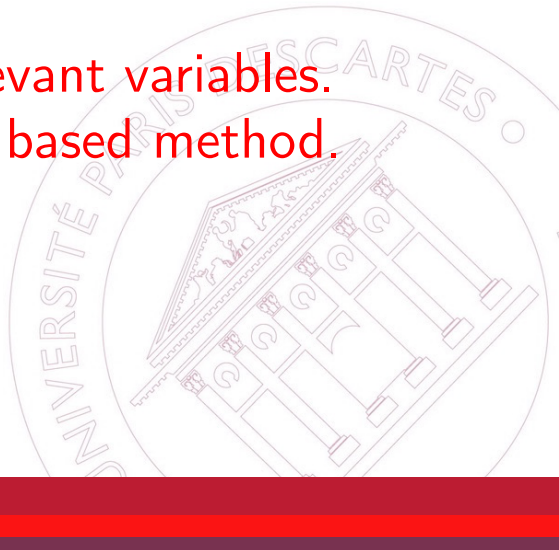
Comparisons with other methods

Simulation setup with $p = 200$, $q = 20$, $d = 10$, $\sigma = 1$.
SSPCA (Jenatton, Obozinski & Bach, '09) achives global sparsity with a $\ell_1 - \ell_2$ norm.

Table: F-score based on 50 runs

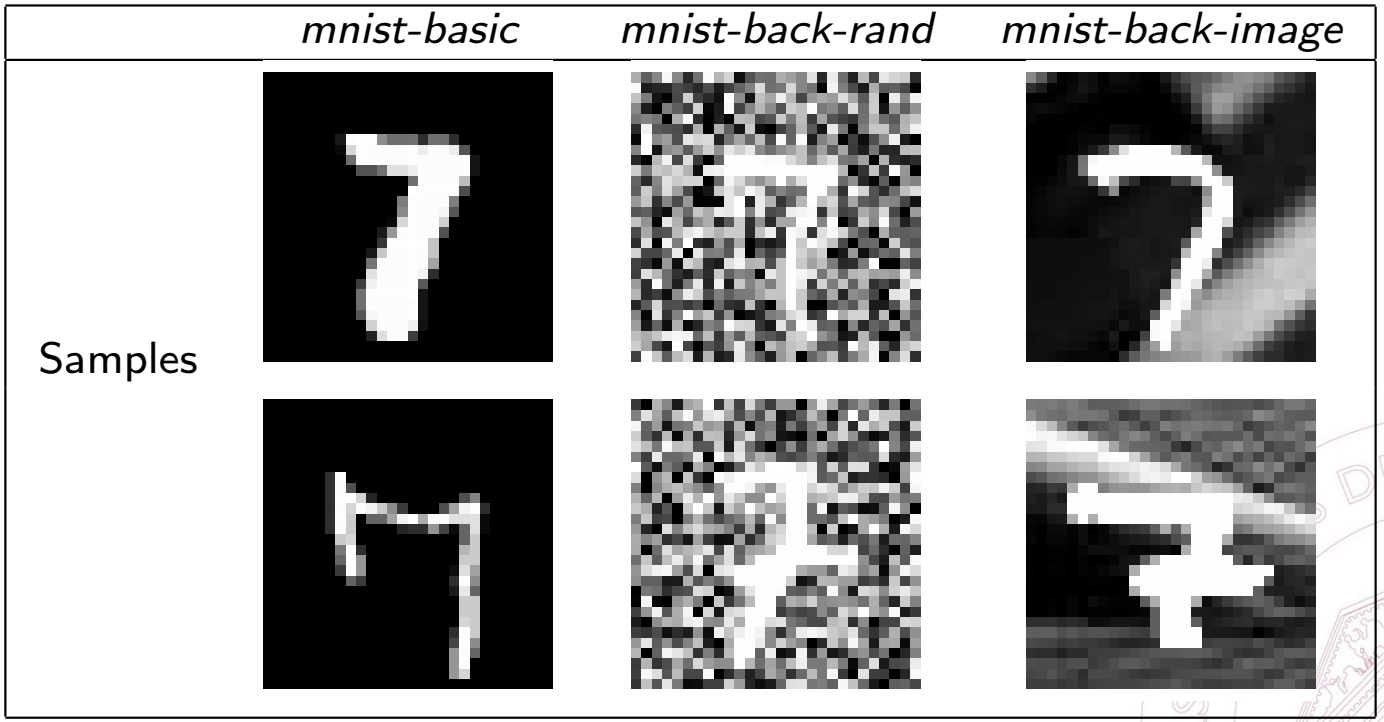
	$n = p/5$	$n = p/4$	$n = \lfloor p/3 \rfloor$	$n = p/2$	$n = p$
SPCA	20.7 ± 0.7	21.2 ± 0.7	21.5 ± 0.7	21.7 ± 0.5	25.2 ± 2.1
SSPCA	66.7 ± 21.4	71.5 ± 20	86.7 ± 14.2	95.6 ± 8.9	98.2 ± 7.2
GSPPCA	$\mathbf{86.8 \pm 7.06}$	$\mathbf{93.9 \pm 3.66}$	$\mathbf{97.2 \pm 2.55}$	$\mathbf{99.2 \pm 1.4}$	$\mathbf{1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

The local method (SPCA) is unable to select the relevant variables.
GSPPCA consistently outperforms the global $\ell_1 - \ell_2$ based method.

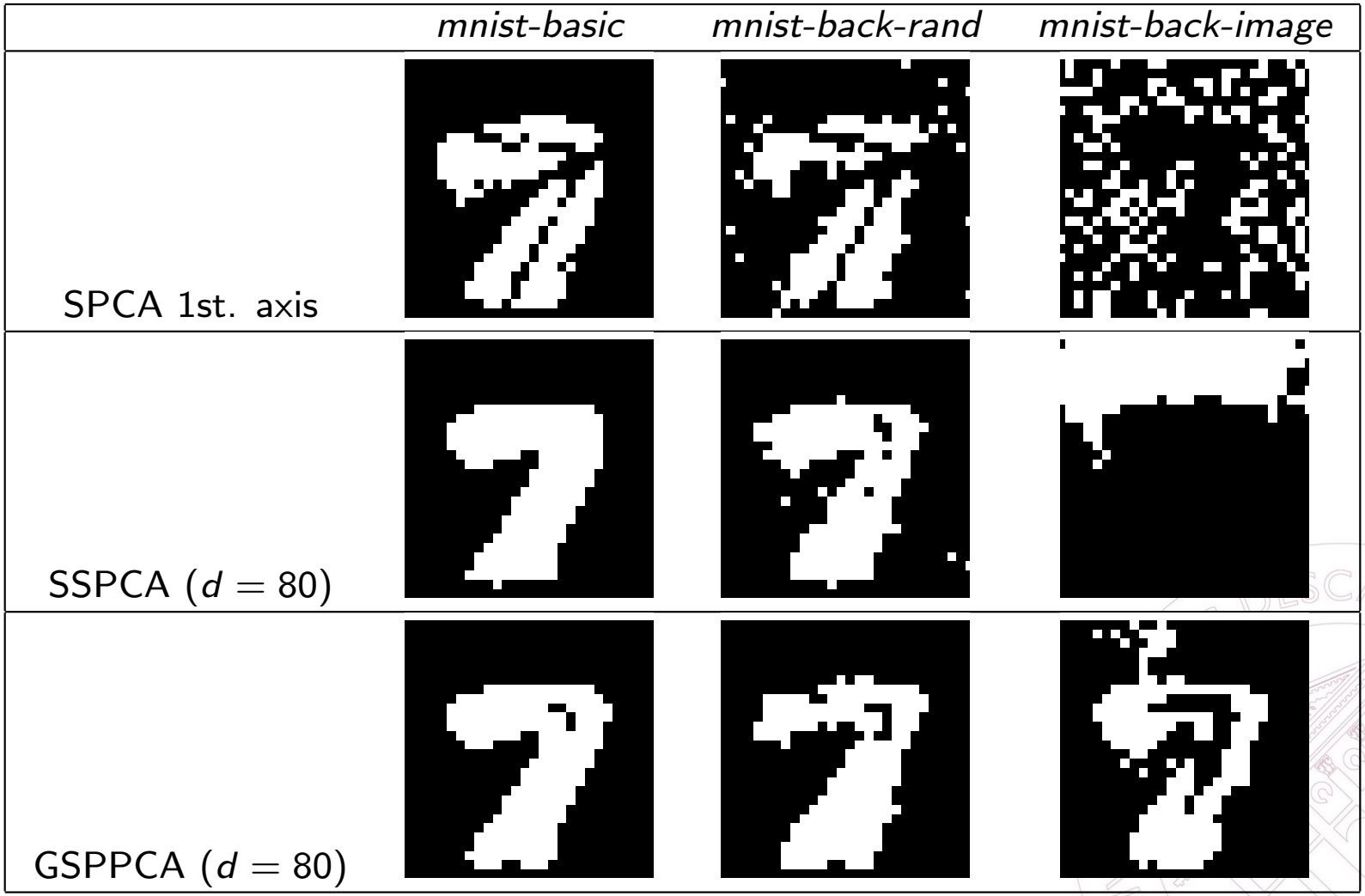


Global versus local - Variations on MNIST ($n = 500$, $p = 784$)

Goal: perform unsupervised variable selection for three datasets introduced by Larochelle, Erhan, Courville, Bergstra & Bengio ('07).



Global versus local - Variations on MNIST ($n = 500$, $p = 784$)

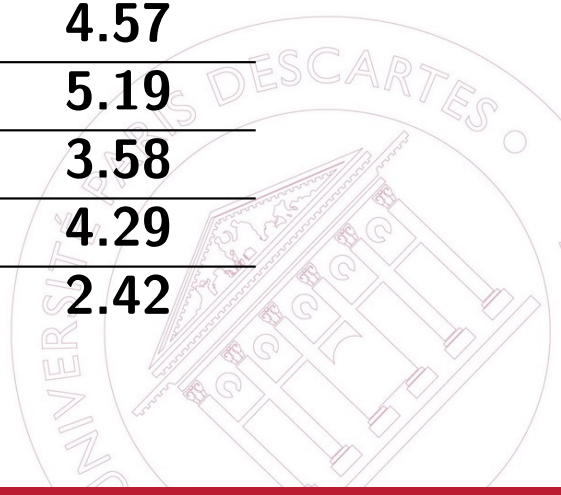


Global versus local - breast cancer data set ($n = 334$, $p = 5391$)

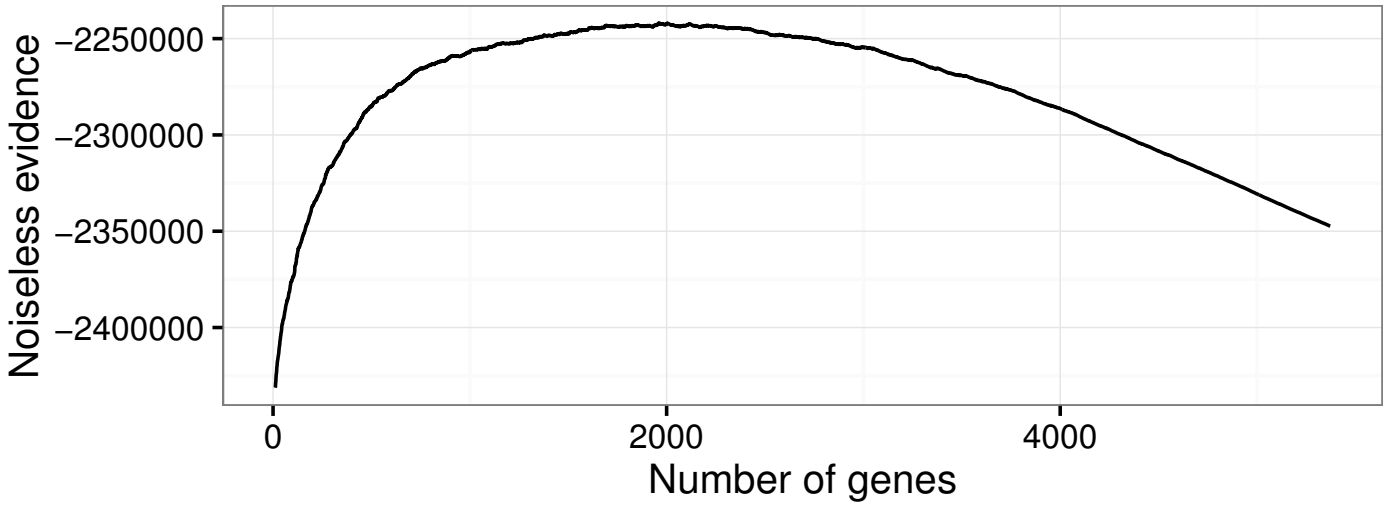
Microarray data from Wang et al. ('05) and Minn et al. ('07).
We can measure the biological significance using the pathway enrichment index (PEI) introduced by Teschendorff, Journée, Absil, Sepulchre & Caldas ('07).

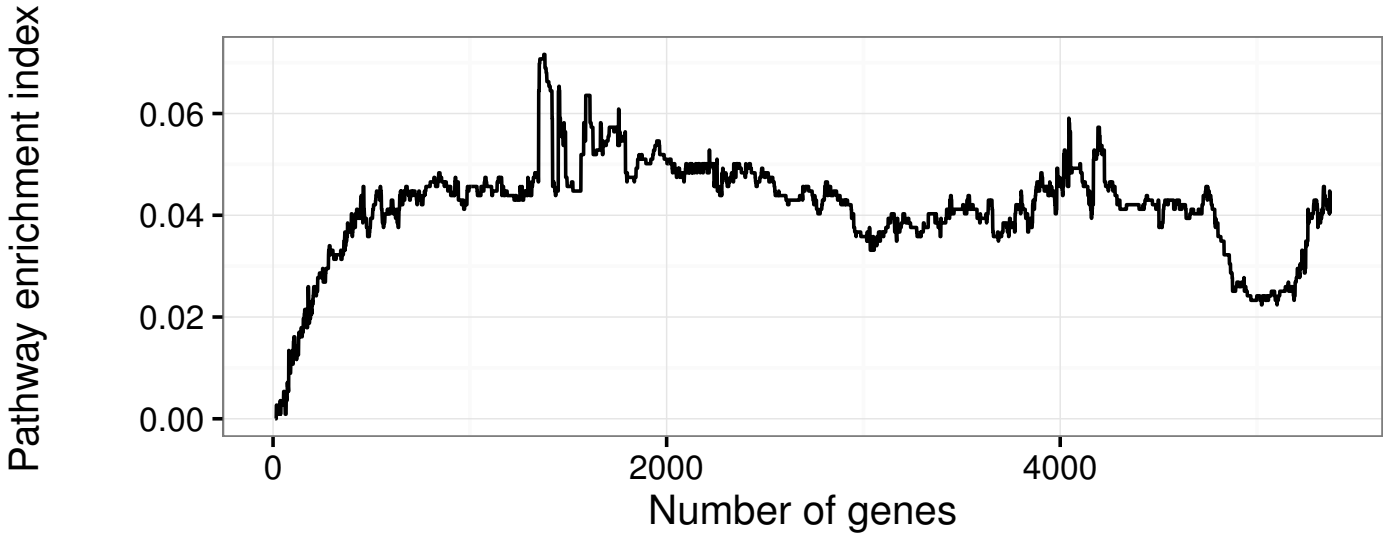
Table: PEI for several fixed cardinalities

Cardinality		tPCA	SPCA	GSPPCA
290	*selected by tPCA*	0.09	0.09	**3.22**
1000		1.88	1.88	**4.57**
1965	*selected by GSPPCA*	1.7	1.61	**5.19**
3000		1.16	1.43	**3.58**
4466	*selected by SPCA*	3.04	3.22	**4.29**
5000		1.79	1.88	**2.42**



Global versus local - breast cancer data set ($n = 334$, $p = 5391$)





Take-home message

For unsupervised feature selection, global sparsity works better than one-dimensional sparse approximations.

Bayesian Variable Selection for Globally Sparse Probabilistic PCA, Electronic Journal of Statistics, vol. 12 (2), pp. 3036-3070, 2018

More (including R code) on http://pamattei.github.io