Sélection de modèles parcimonieux pour l'apprentissage statistique en grande dimension

Pierre-Alexandre Mattei

Laboratoire MAP5, UMR CNRS 8145
Université Paris Descartes

http://pamattei.github.io – @pamattei
pierre-alexandre.mattei@parisdescartes.fr

Thèse de doctorat dirigée par
Charles Bouveyron et Pierre Latouche

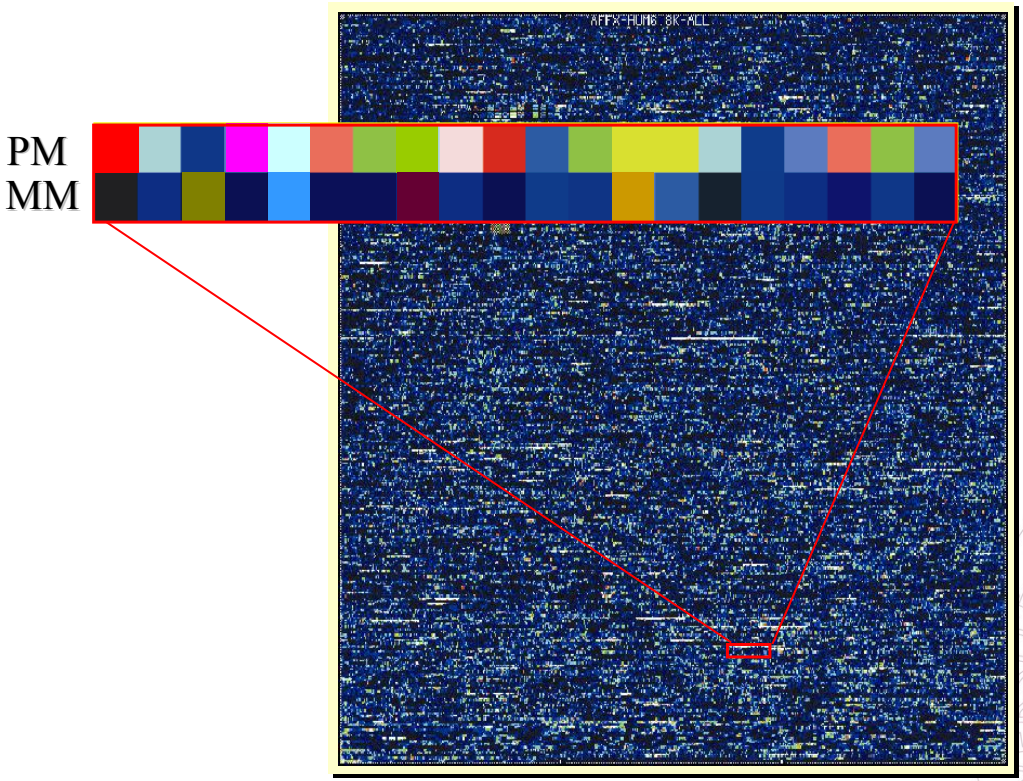
26/10/2017

High-dimensional data are ubiquitous

Traditional statistical paradigm:

- large number n of observations (patients, voters...),
- small number p of variables (medical measurements, answers in a survey...).

Modern (big) data: DNA microarray $n \approx 100, p \approx 10\,000$.

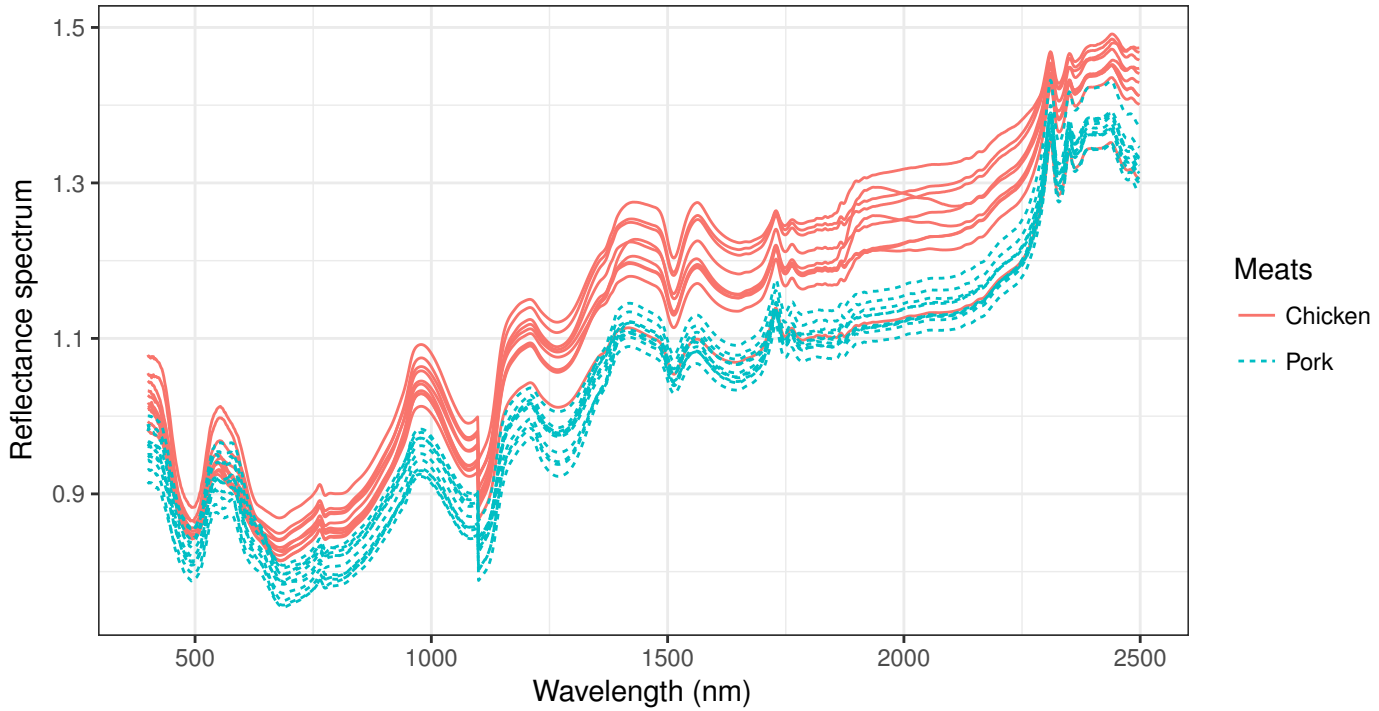


High-dimensional data are ubiquitous

Traditional statistical paradigm:

- large number n of observations (patients, voters...),
- small number p of variables (medical measurements, answers in a survey...).

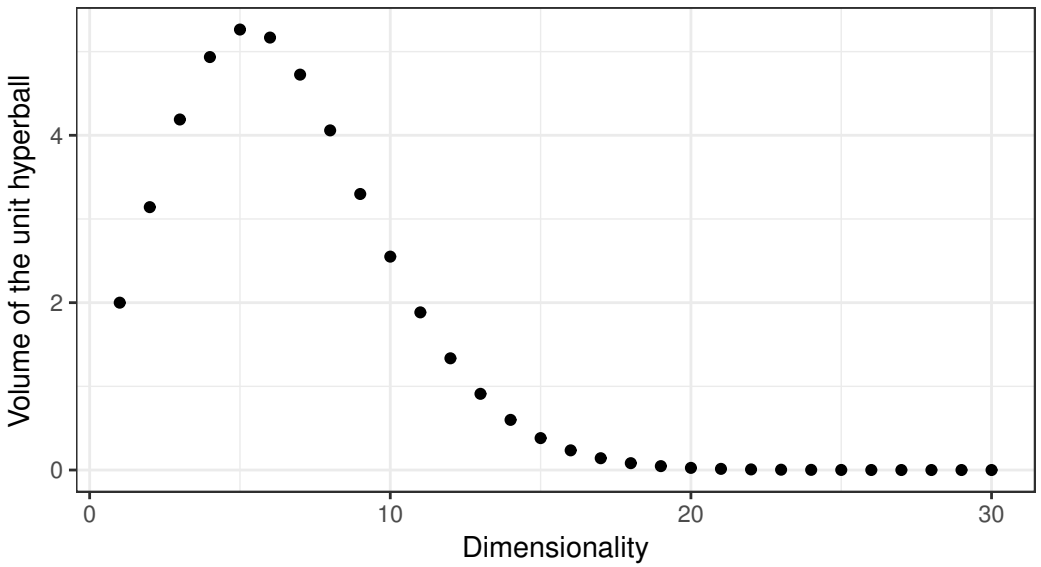
Modern (big) data: NMR spectra $n \approx 100, p \approx 1\,000$.



High-dimensional data are cursed

High-dimensional datasets are collections of points in high-dimensional spaces...
...and the geometry of high-dimensional spaces is rather peculiar.

High-dimensional Euclidean (hyper)balls are essentially empty!

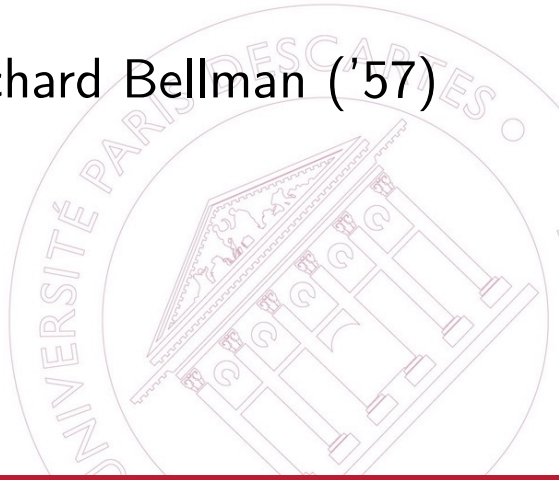


High-dimensional data are cursed

Since Gauss's and Legendre's least squares ($\approx$1810), most of classical statistics rely on Euclidean distances, which do not behave nicely in high-dimensions.

"All this [the problems related to high-dimensional geometry] *may be subsumed under the heading "the curse of dimensionality". Since this is a curse (...) there is no need to feel discouraged about the possibility of obtaining significant results despite it."*

Richard Bellman ('57)



Betting on sparsity

Parametric statistical models assume that the observed data $\mathbf{X} \in \mathbb{R}^{n \times p}$ comes from a density in a parametrized family $(p(\cdot|\boldsymbol{\theta}))_{\boldsymbol{\theta} \in \Theta}$.

The dimension of Θ usually grows with the dimensionality p of the data, which is another challenge of high-dimensional inference!

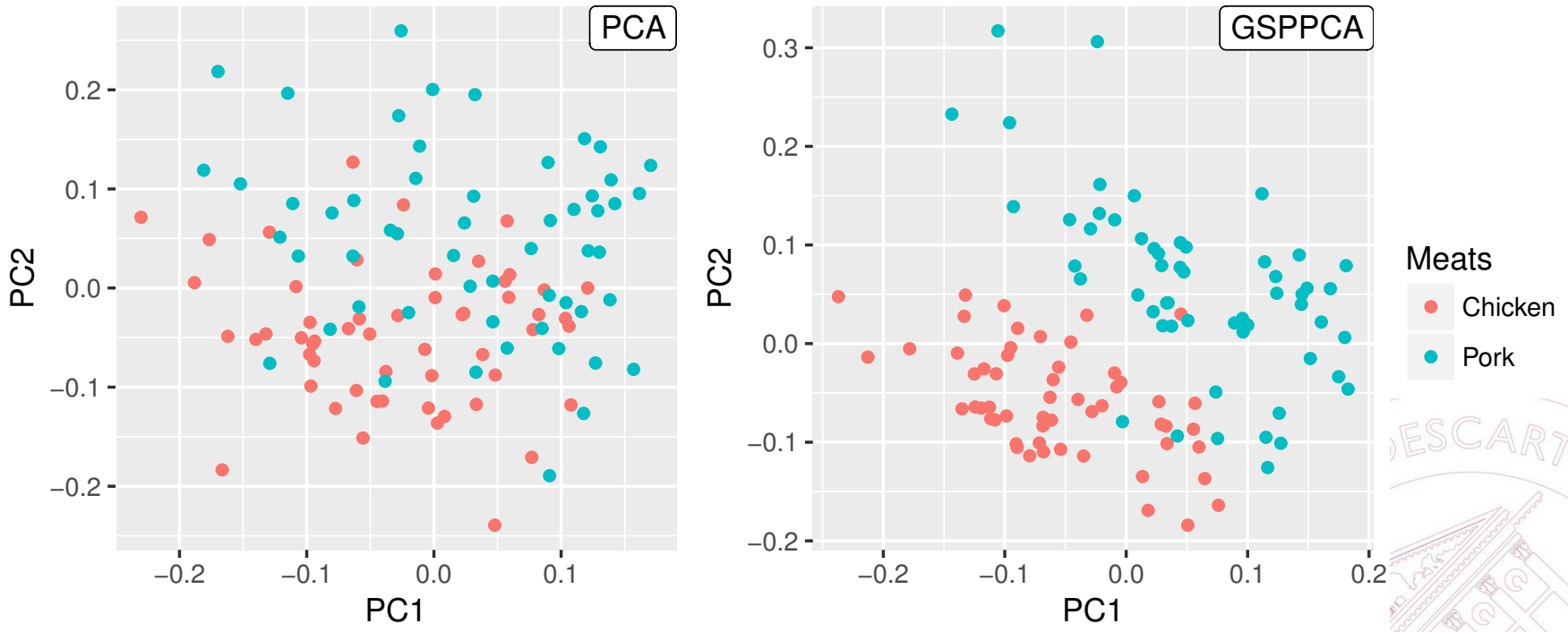
But most statistical/geometrical problems tend to disappear if we assume that $\boldsymbol{\theta}$ has few nonzero coefficients. We say that $\boldsymbol{\theta}$ is q-sparse, with $q \ll p$.

"This has been termed the "Bet on sparsity" principle: Use a procedure that does well in sparse problems, since no procedure does well in dense problems."

Hastie, Tibshirani & Wainwright ('15)

Visualizing data via PCA and the bet on sparsity

PCA aims at summarizing high-dimensional data using only two transformed variables. Without betting on sparsity, the results are much less interpretable.



Betting on sparsity via likelihood penalization

A natural way to find a sparse parameter is to maximize a penalized version of the likelihood

$$\hat{\boldsymbol{\theta}} \in \text{argmax}_{\boldsymbol{\theta} \in \Theta} \log p(\mathbf{X}|\boldsymbol{\theta}) - \lambda ||\boldsymbol{\theta}||_0.$$

This combinatorial problem lacks scalability, and is often replaced by

$$\hat{\boldsymbol{\theta}} \in \text{argmax}_{\boldsymbol{\theta} \in \Theta} \log p(\mathbf{X}|\boldsymbol{\theta}) - \lambda ||\boldsymbol{\theta}||_1.$$

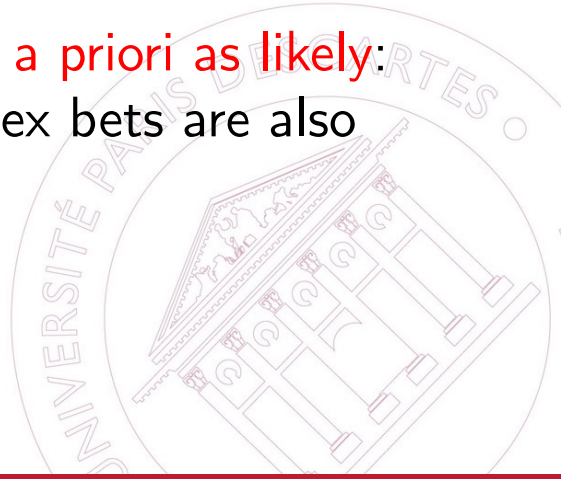
Such optimization problems – often called lasso problems, following Tibshirani ('96) – are highly scalable but hard to calibrate.

Since we should bet on sparsity, why not do it probabilistically ?

The Bayesian framework allows to express prior beliefs about $\boldsymbol{\theta}$ by treating it as a random variable. Here, our prior belief is that $\boldsymbol{\theta}$ might be sparse.

To translate that belief, we will use the sparsity pattern $\mathbf{v} \in \{0,1\}^p$ of $\boldsymbol{\theta}$, which is the binary vector that indicates which coefficients are nonzero. The number of nonzero coefficients is denoted by q.

A simple Bayesian bet on sparsity: all sparsity patters are a priori as likely: $p(\mathbf{v})$ is the uniform distribution over $\{0,1\}^p$. More complex bets are also possible.



Bayes's theorem updates the bet on sparsity

Using Bayes's theorem allows us to find which sparsity patters are more likely, by computing the posterior probabilities of such patterns:

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

where

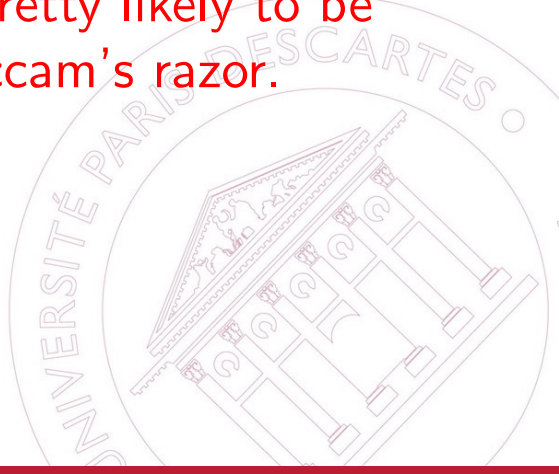
$$p(\mathbf{X}|\mathbf{v}) = \int_{\Theta} p(\mathbf{X}|\mathbf{v})p(\boldsymbol{\theta}|\mathbf{v})d\boldsymbol{\theta},$$

is the marginal likelihood or evidence of the data for the sparsity pattern $\mathbf{v}$. This sparsity pattern can be estimated via maximum a posteriori

$$\hat{\mathbf{v}} \in \text{argmax}_{\mathbf{v}\in\{0,1\}^p} p(\mathbf{X}|\mathbf{v}).$$

Betting on sparsity as Bayesian model uncertainty

- The Bayesian bet on sparsity is a particular instance of Bayesian model uncertainty. All possible sparsity patterns can be viewed as competing statistical models, over which we spread prior beliefs.
- The idea of spreading prior belief between models and computing consequently their posterior probabilities was independently developed by Harold Jeffreys & Dorothy Wrinch ($\approx$ 1920), and by Jack Good & Alan Turing ($\approx$ 1942).
- It embodies the fact that simpler models are a priori pretty likely to be useful, a philosophical principle often referred to as Occam's razor.



Two main challenges tackled in this thesis

The Bayesian framework provides a coherent way of expressing a bet on sparsity, and will be our main tool to study high-dimensional data. Two main issues lie however on top of these foundations:

- computing posterior probabilities of sparsity patterns imply computing challenging high-dimensional integrals,
- there are 2^p sparsity patterns that should be tested, which appears infeasible!

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

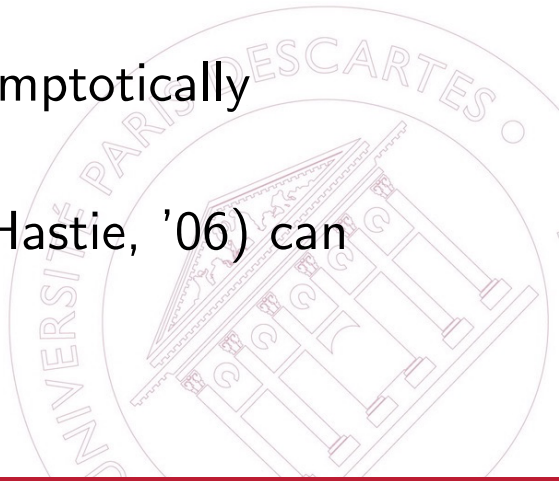
Regularizing the maximum likelihood procedure

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

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 $\boldsymbol{\beta}$

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

(*à la* 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

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

Towards scalable model selection: from discrete to continuous...

To tackle this combinatorial nature of the optimization problem, we replace $\mathbf{v}$ by a continuous parameter $\mathbf{u} \in [0,1]^p$ and use an EM algorithm (Dempster, Laird & Rubin, '77) to maximize the marginal likelihood of this relaxed model.

We end up with a continuous estimate $\hat{\mathbf{u}} \in [0,1]^p$.

...and to discrete again.

How to reverse the relaxation ?

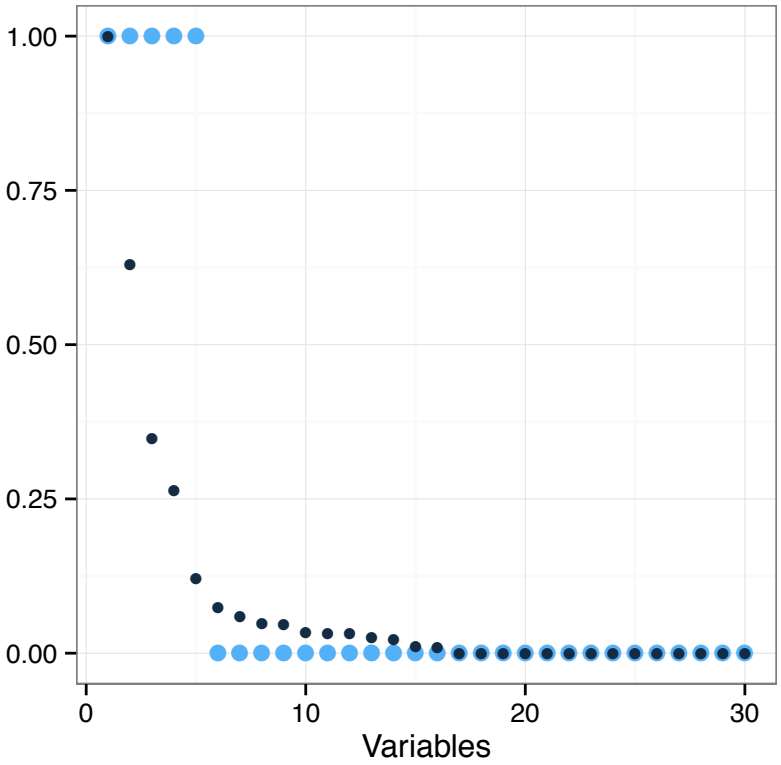
To lead to the right model, $\hat{\mathbf{u}}$ has to be binarized.

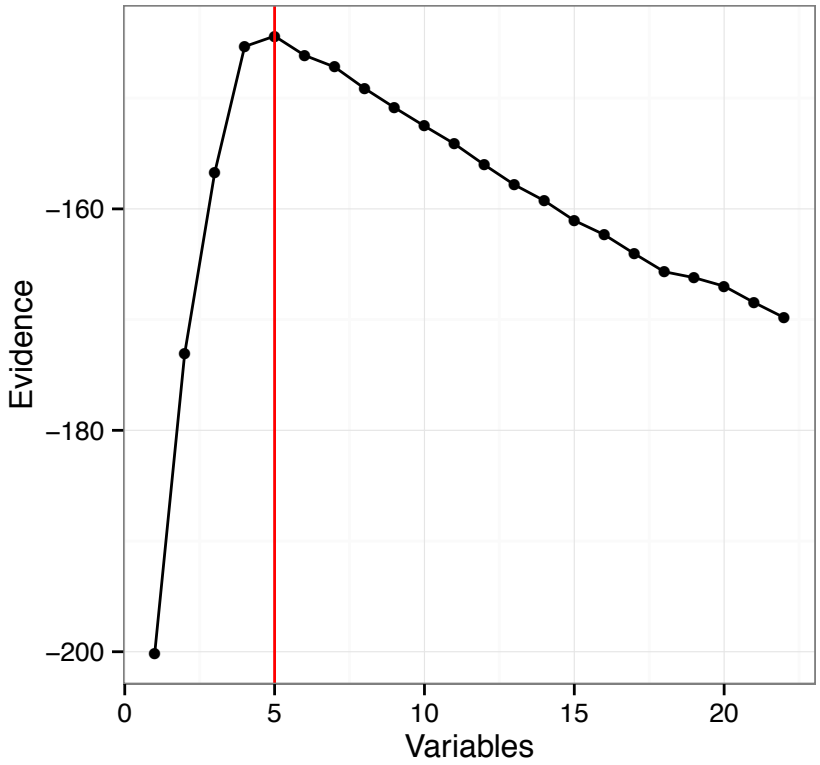
$\hat{\mathbf{u}}$ leads to a path of p models

We find $\hat{\mathbf{v}}$ by maximizing the marginal likelihood of the non-relaxed model over the path of p nested models implied by the ordering of the coefficients of $\hat{\mathbf{u}}$.

Relaxation-binarization in action (toy model)

Values of $\hat{\mathbf{u}}$ and actual binary values of $\mathbf{v}$ (left) and evidence computed over the path of models (right).





A (short) benchmark study ($p = 100$, $q = 40$)

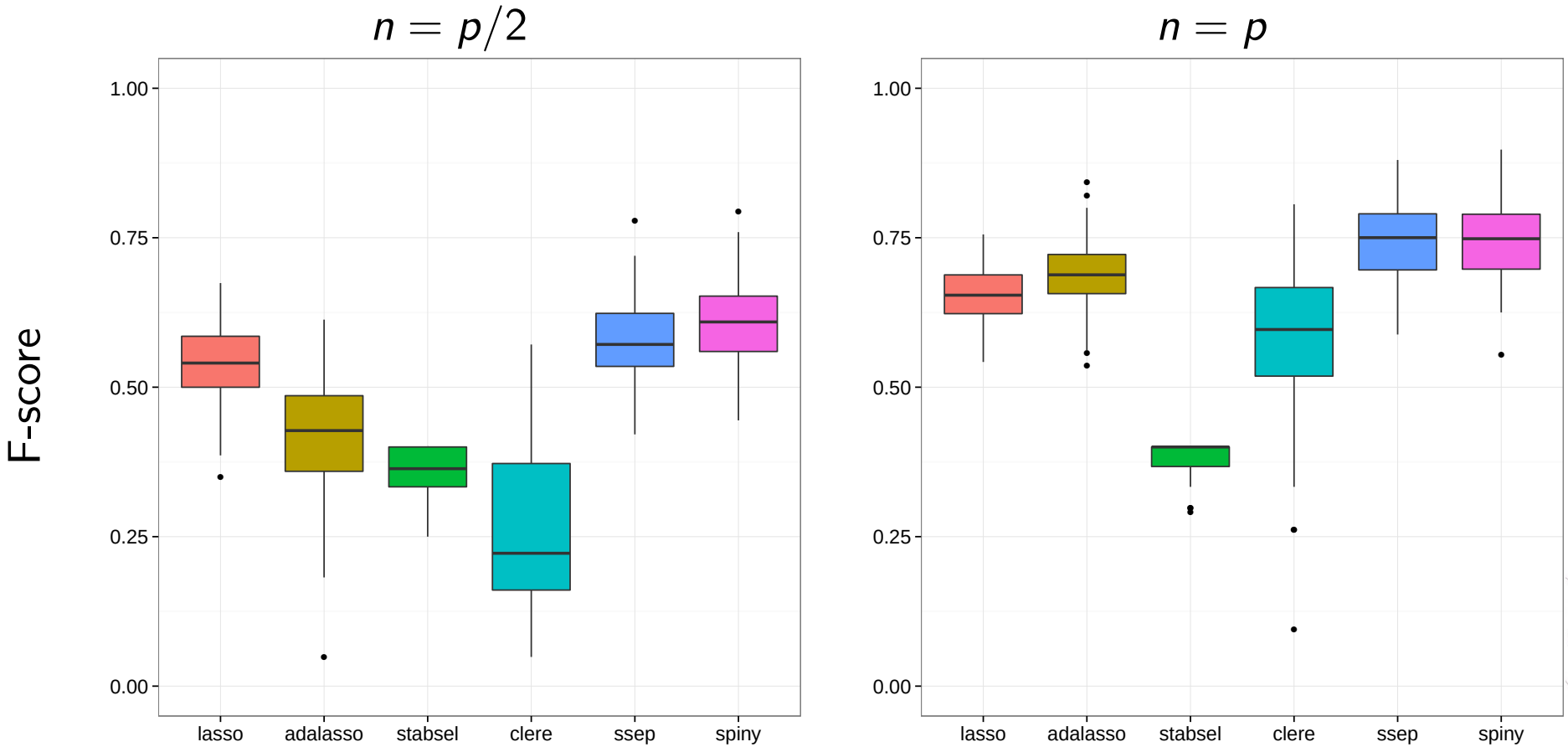
Highly correlated predictors with a Toeplitz covariance matrix **R** defined by

- $r_{ii} = 1$ for all $\in \{1, ..., p\}$,
- $r_{ij} = 0.75^{|i-j|}$ for $i, j \in \{1, ...p\}$ and $i \neq j$.

Evaluation metric: the F-score allows to measure the quality of a variable selection procedure by giving a score between 0 and 1.

18 different simulations schemes were done.

A (short) benchmark study ($p = 100$, $q = 40$)

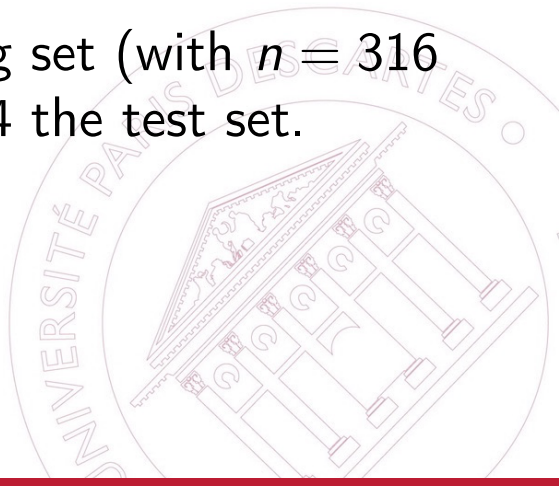


Predicting the number of visitors in the Orsay museum with bike-sharing data

The "OrsayVelib" database: a high-dimensional problem

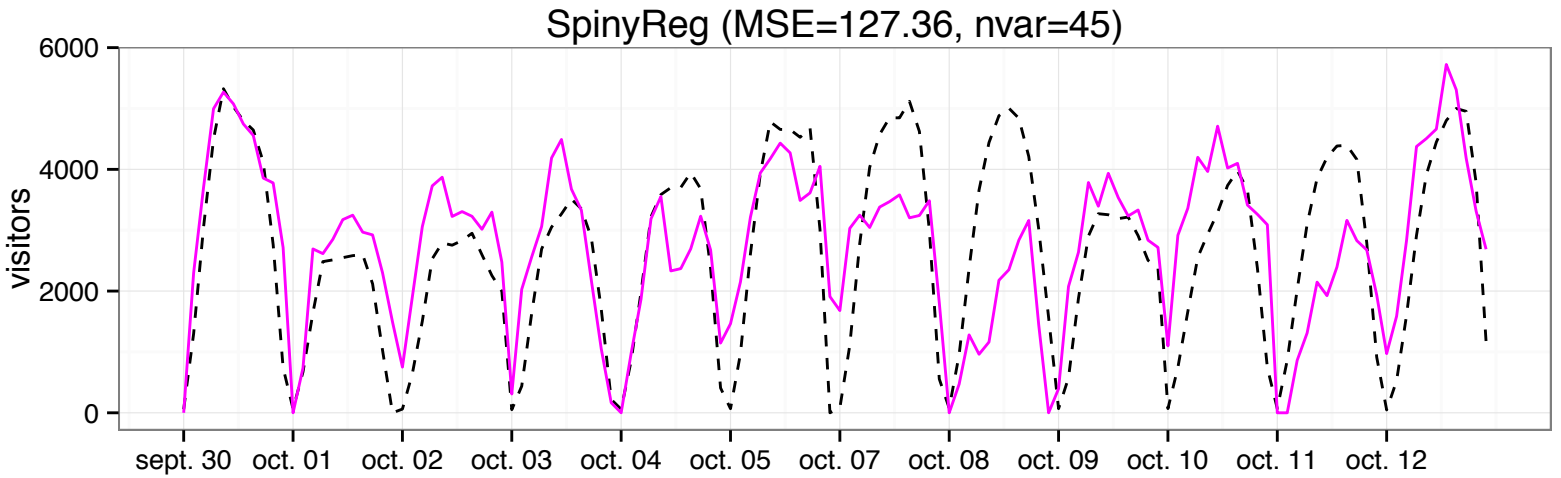
We wish to predict the number of visitors of the Orsay museum using the activity of the Paris bike-sharing system (*Vélib'*).

- At each hour, the number of visitors in the museum constitutes the response variable,
- The predictors are the loadings of the $\boldsymbol{p = 1158}$ *Vélib'* stations in Paris,
- The month of September 2014 constitutes the learning set (with $n = 316$ observations), and the first two weeks of October 2014 the test set.

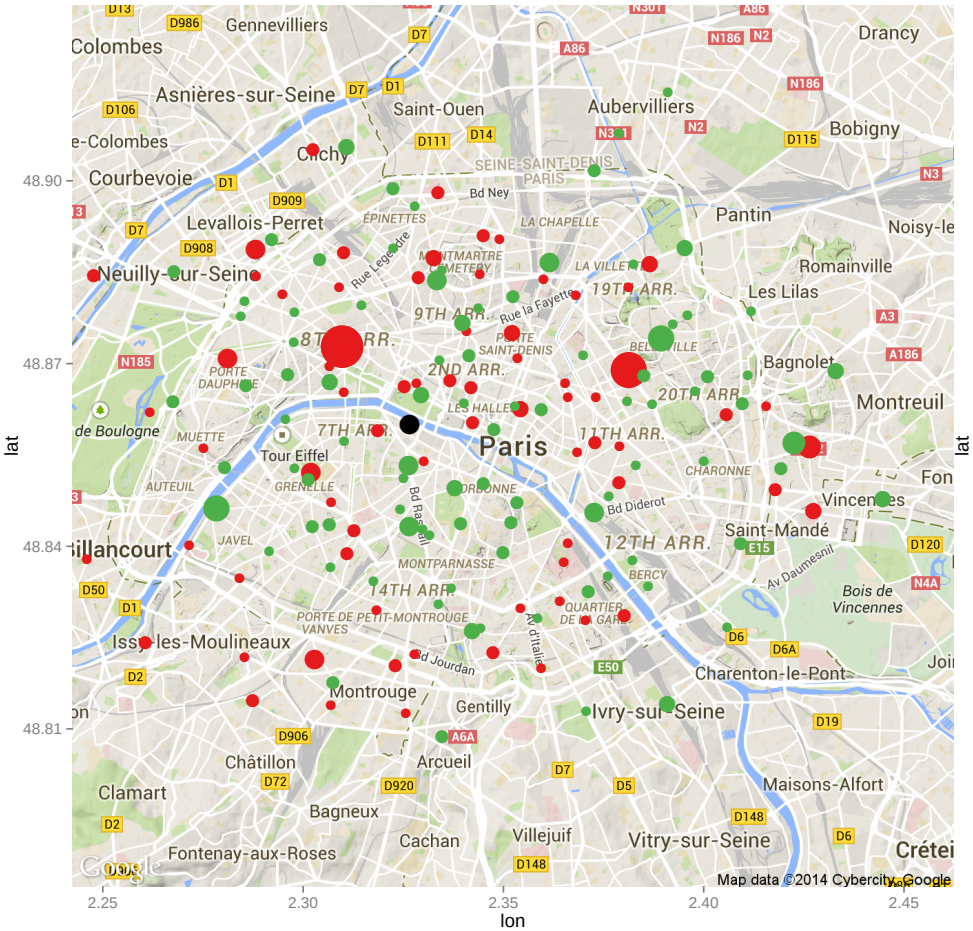


SpinyReg generalizes better and is sparser

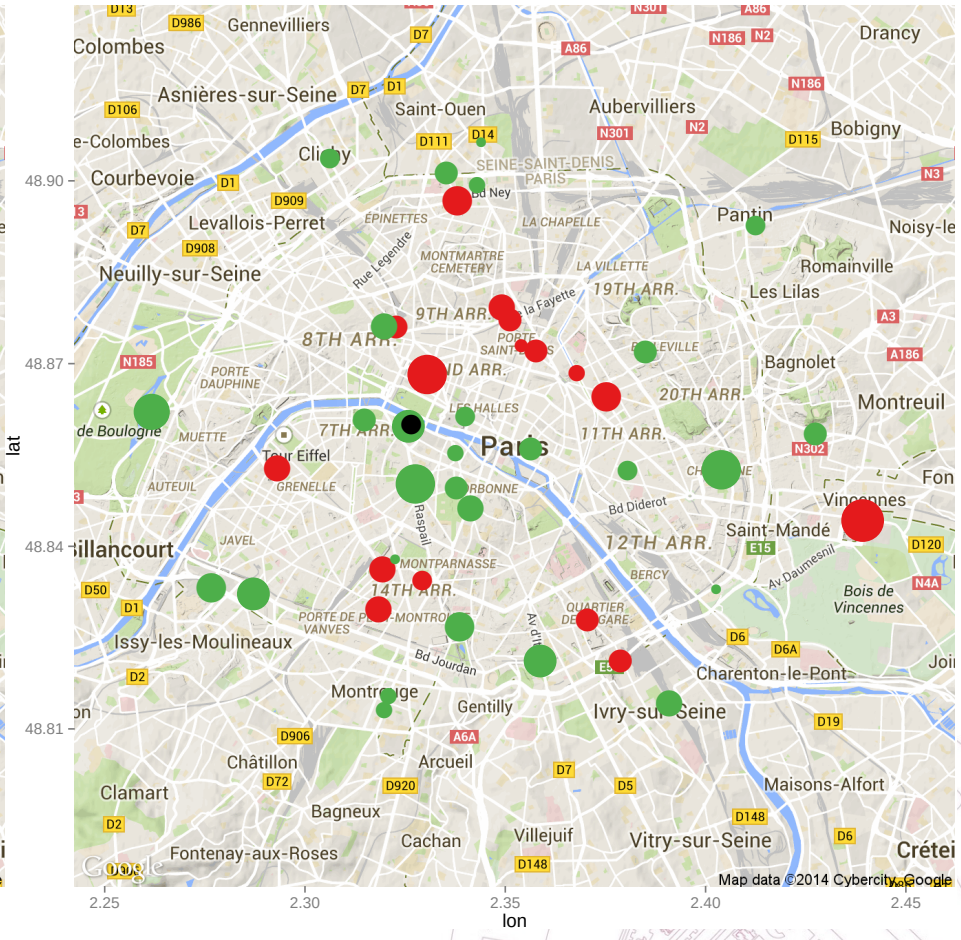
	Ridge	SSEP	Lasso	Adalasso	SpinyReg
MSE$\times 10^{-4}$	145.66	144.38	**132.08**	159.17	**127.36**
Selected variables	1158	1146	**167**	155	**45**



SpinyReg gives more interpretable results



Lasso (167 var.)



SpinyReg (45 var.)

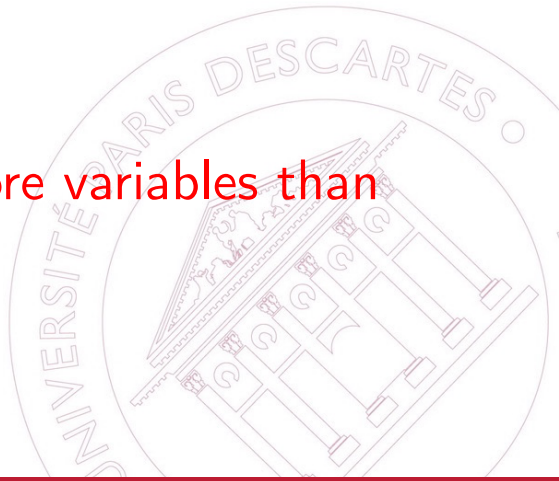
Principal component analysis aims at summarizing multivariate data

Goal: Summarize all p variables with $d \ll p$ scores.

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 !

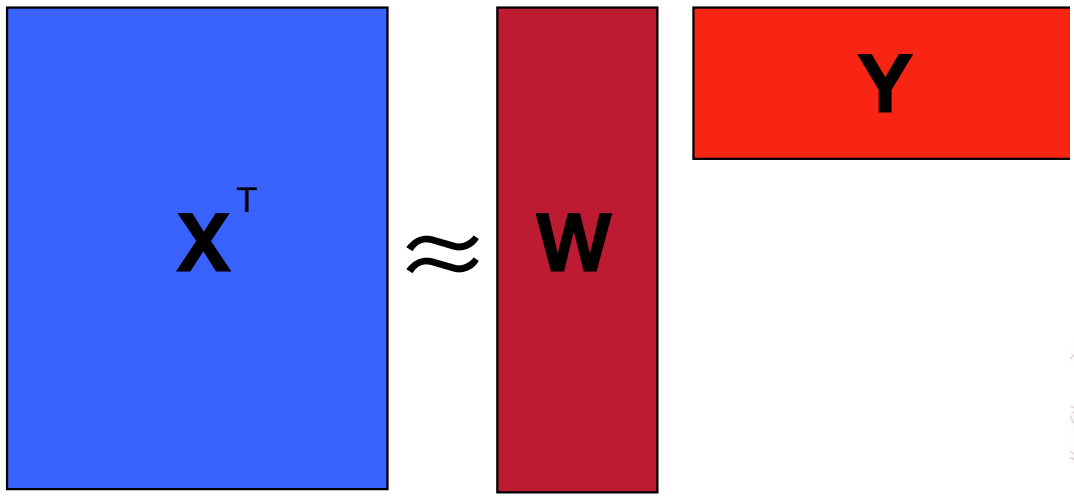


Principal component analysis

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:

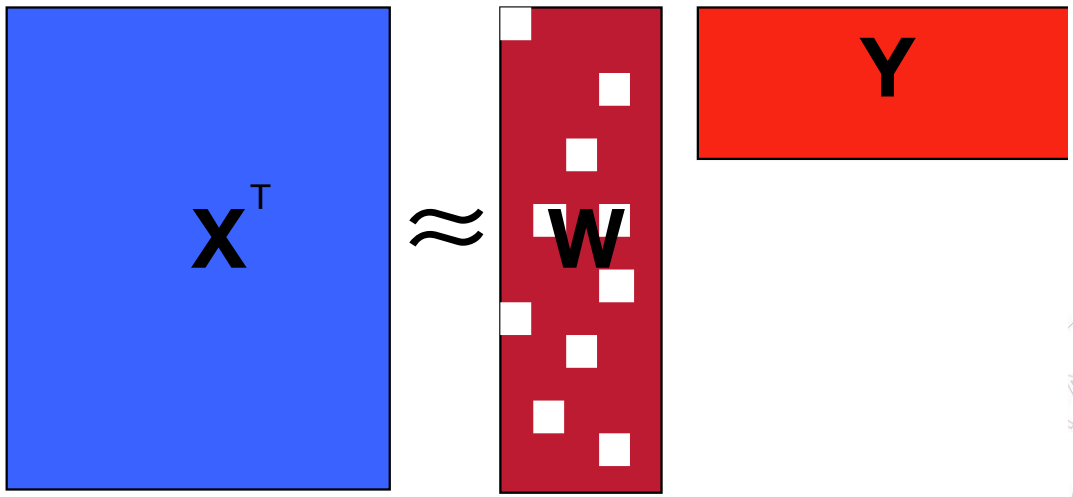


(Locally) Sparse Principal Component Analysis

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:

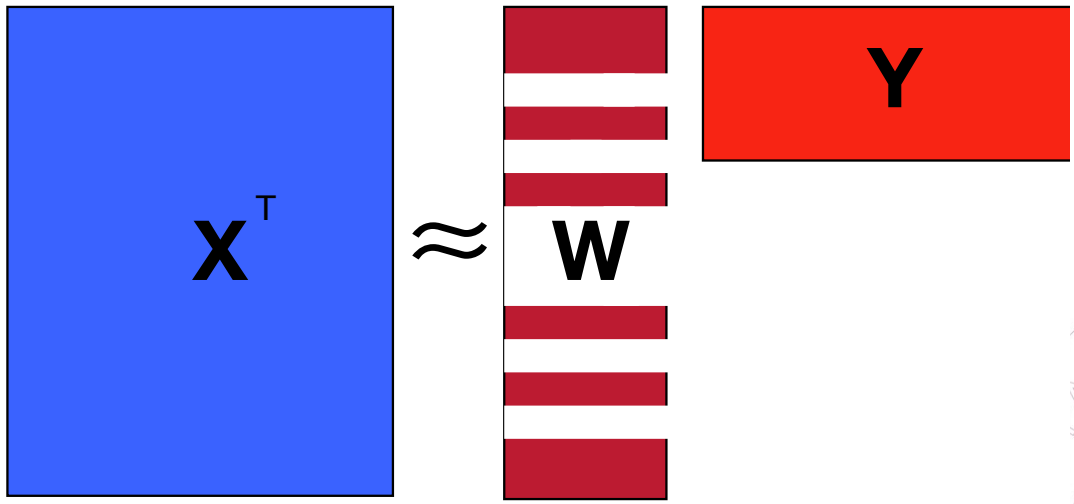


Globally Sparse Principal Component Analysis

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:



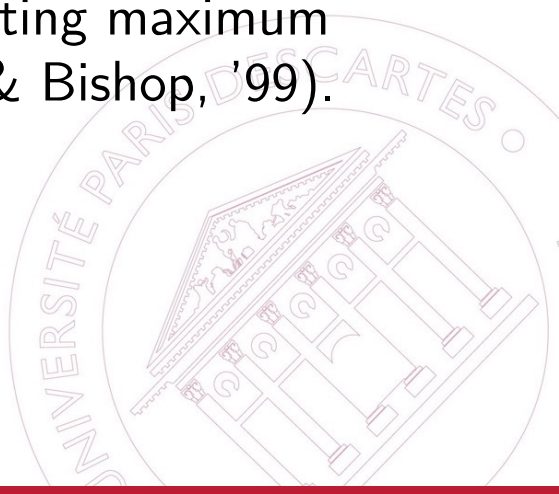
Probabilistic PCA: a generative model for PCA

PPCA assumes that each observation is driven by the following generative model:

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

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 sense that computing maximum likelihood estimates recovers the principal axes (Tipping & Bishop, '99).



Betting of global sparsity within PPCA

We consider the model

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

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:

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

The marginal likelihood appears to be intractable!

Theorem

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

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

This kind of integral is known to be difficult to compute (Ogata, '05). 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 ?

Our solution: 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:

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

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

Aparté: Multiplying a Gaussian matrix with a Gaussian vector

A random variable $\mathbf{z} \in \mathbb{R}^p$ is said to have a multivariate generalized asymmetric Laplace distribution with parameters $s > 0, \boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathcal{S}_p^+$ if its characteristic function is

$$\forall \mathbf{u} \in \mathbb{R}^p, \ \phi_{\mathrm{GAL}_p(\boldsymbol{\Sigma},\boldsymbol{\mu},s)}(\mathbf{u}) = \left(\frac{1}{1 + \frac{1}{2}\mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u} - i\boldsymbol{\mu}^T\mathbf{u}} \right)^s .$$

(Kotz, Kozubowski & Podgórski, '01)

Theorem

Let $\mathbf{W}$ *be a* $p \times d$ *random matrix with i.i.d. columns following a* $\mathcal{N}(0, \boldsymbol{\Sigma})$ *distribution and let* $\mathbf{y} \sim \mathcal{N}(0, \mathbf{I}_d)$ *independent from* $\mathbf{W}$*. Then*

$$\mathbf{W}\mathbf{y} \sim \mathrm{GAL}_p(2\boldsymbol{\Sigma}, 0, d/2).$$

Theorem

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

$$p(\tilde{\mathbf{x}}|\mathbf{v}, \alpha, \sigma_1) = \mathcal{N}(\tilde{\mathbf{x}}_{\mathbf{v}^\perp}|0, \sigma_1 \mathbf{I}_{p-q}) \mathrm{GAL}_q(\tilde{\mathbf{x}}_{\mathbf{v}}|2/\alpha^2 \mathbf{I}_q, 0, d/2).$$

This theorem allows us to exactly compute the noiseless empirical Bayes marginal log-likelihood defined as $\mathcal{L}(\mathbf{X}, \mathbf{v}, \alpha, \sigma_1) = \sum_{i=1}^n \log p(\tilde{\mathbf{x}}_i|\mathbf{v}, \alpha, \sigma_1)$. Up to unnecessary constants,

$$\mathcal{L}(\mathbf{X}, \mathbf{v}, \alpha, \sigma_1) = -\frac{||\mathbf{X}_{\mathbf{v}^\perp}||_{\mathrm{F}}^2}{2\sigma_1^2} - n(p-q)\log\sigma_1 + \frac{nq}{2}\log\alpha + \sum_{i=1}^n \left(\log K_{(q-d)/2}(\alpha||\mathbf{x}_{\mathbf{v}\,i}||_2) - q\log||\mathbf{x}_{\mathbf{v}\,i}||_2\right).$$

Automated hyperparameters choice

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 properties of Bessel functions insure that the objective function is univariate and concave !

Scalable inference through a 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

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

We follow a variational approach to minimize the free energy

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

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

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

The relaxation is reversed as before.

Comparisons with other methods

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

Table: F-score$\times$100 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	$\mathbf{86.8 \pm 7.06}$	$\mathbf{93.9 \pm 3.66}$	$\mathbf{97.2 \pm 2.55}$	$\mathbf{99.2 \pm 1.4}$	$\mathbf{100 \pm 0}$

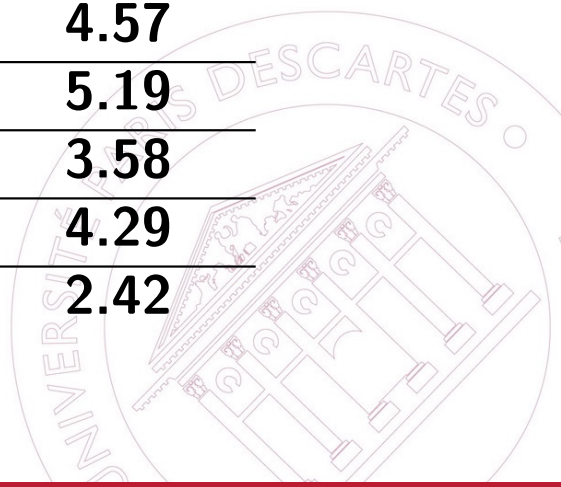
The local method (SPCA) is unable to select the relevant variables.
GSPPCA consistently outperforms the global $\ell_1 - \ell_2$ based method.

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

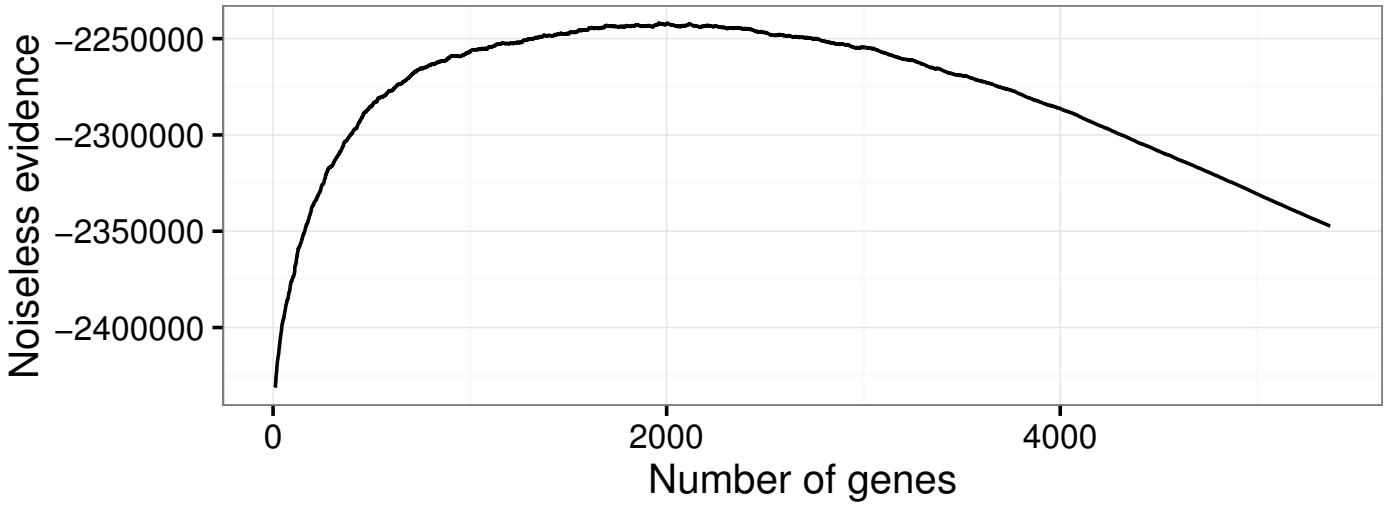
Microarray data from Wang et al. ('05) and Minn et al. ('07).
We can measure the biological significance using the pathway enrichment index (PEI) introduced by Teschendorff, Journée, Absil, Sepulchre & Caldas ('07).

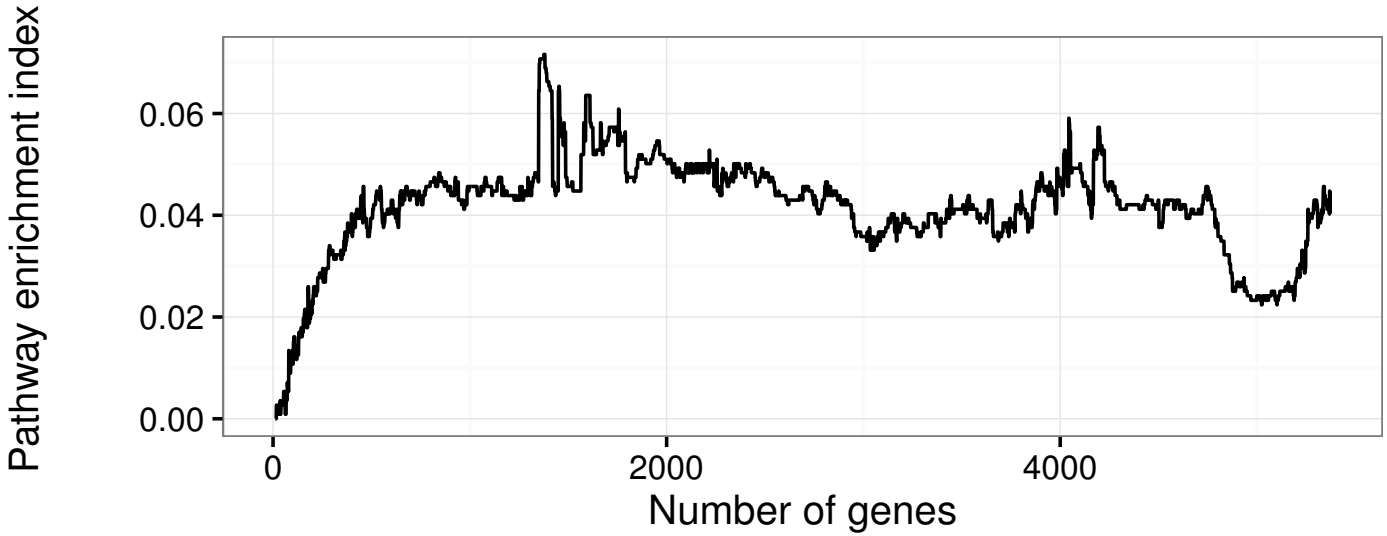
Table: PEI for several fixed cardinalities

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 ($n = 334$, $p = 5391$)





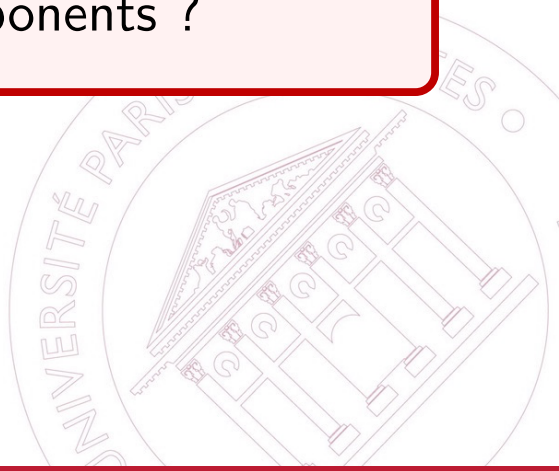
Principal component analysis

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 $\mathbf{X}^T\mathbf{X}$, called principal components.

How to choose the number d of principal components ?



Bayesian model selection for PPCA

Bayesian model selection provides an automatic way of choosing d.

Once we have a prior distribution $p(\mathcal{M}_d)p(\mathbf{W}, \sigma|\mathcal{M}_d)$, we can compute posterior probabilities of dimensions as

$$p(\mathcal{M}_d|\mathbf{X}) \propto p(\mathbf{X}|\mathcal{M}_d)p(\mathcal{M}_d),$$

where

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

is the marginal likelihood of the data.

Bayesian model selection for PPCA

Problem: the marginal likelihood is a challenging high-dimensional integral!

Usual solutions: variational (Archambeau & Bach, '08) or Laplace approximations (Minka '00, Hoyle '08, Sobczyk, Bogdan & Josse '17), MCMC sampling (Hoff '07).

Our approach: play once again with the PPCA model to obtain a closed form expression of the marginal likelihood.

Looking at the marginal distribution of the data

$$\mathbf{x}_i = \underbrace{\mathbf{W}\mathbf{y}_i}_{\text{Random matrix} \times \text{Gaussian vector}} + \underbrace{\varepsilon_i}_{\text{Gaussian vector with random variance}}$$

Is is possible to make these two terms follow the same kind of distribution?

The generalized Gauss-Laplace representation

Theorem (Kotz, Kozubowski & Podgórski, '01)

If $u \sim \text{Gamma}(s, 1)$ *and* $\mathbf{e} \sim \mathcal{N}(0, \mathbf{\Sigma})$ *is independent of* u*, we have*

$$\sqrt{u}\mathbf{e} \sim \text{GAL}_p(\mathbf{\Sigma}, 0, s).$$

Proposition

Let $s_1, s_2 > 0, \boldsymbol{\mu} \in \mathbb{R}^p$ *and* $\mathbf{\Sigma} \in \mathcal{S}_p^+$*. If* $\mathbf{z}_1 \sim \text{GAL}_p(\mathbf{\Sigma}, \boldsymbol{\mu}, s_1)$ *and* $\mathbf{z}_2 \sim \text{GAL}_p(\mathbf{\Sigma}, \boldsymbol{\mu}, s_2)$ *are independant random variables, then*

$$\mathbf{z}_1 + \mathbf{z}_2 \sim \text{GAL}_p(\mathbf{\Sigma}, \boldsymbol{\mu}, s_1 + s_2).$$

This can be combined with our result on the distribution of the product of a Gaussian matrix with a Gaussian vector!

Towards exact marginal likelihood with a normal-gamma prior

All of this motivates the following normal-gamma prior:

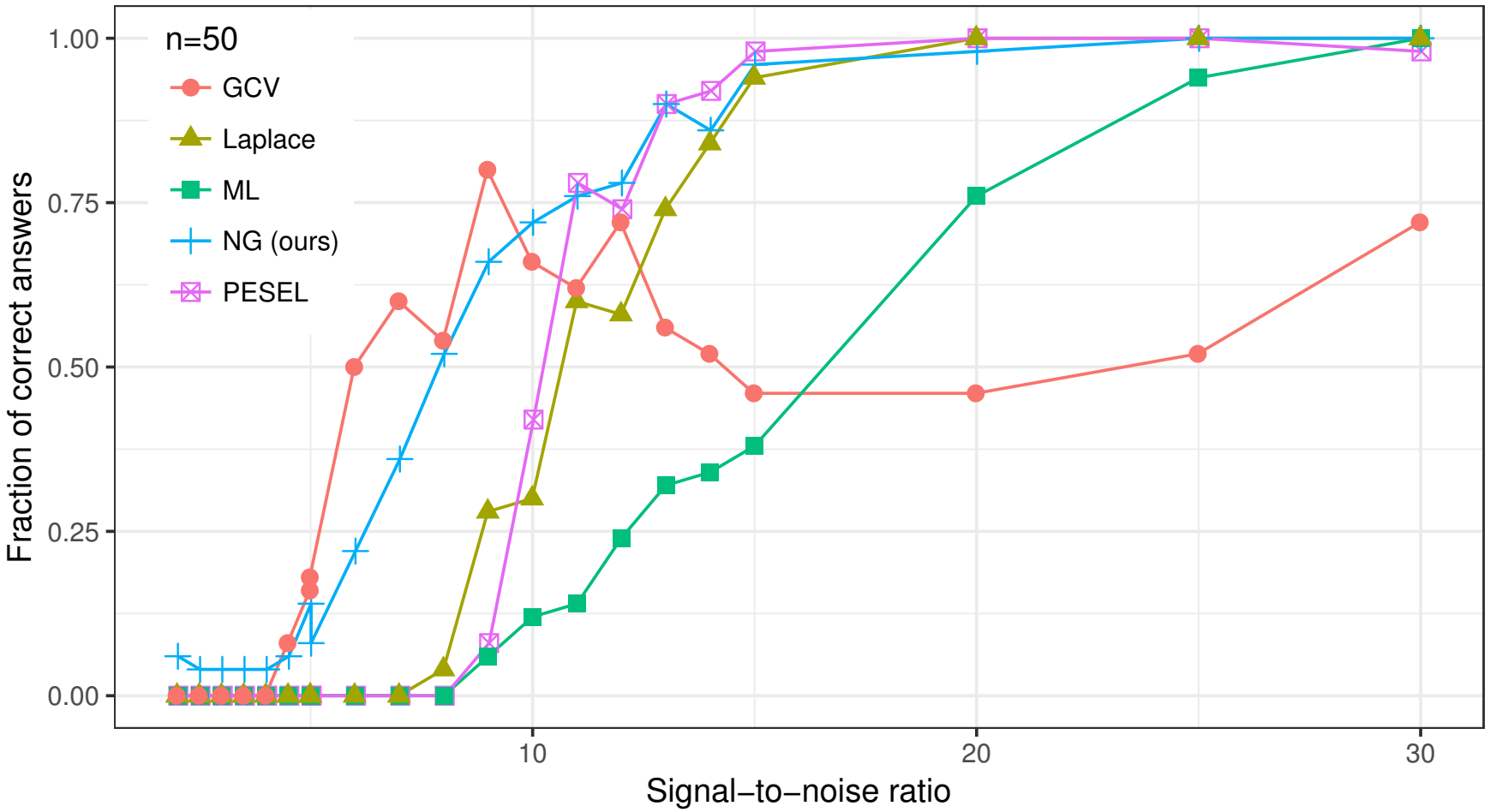
- Gaussian prior for the loading matrix $w_{jk} \sim \mathcal{N}(0, \phi^{-1})$ for $j \in \{1, ..., p\}$ and $k \in \{1, ..., d\}$ with some precision hyperparameter $\phi > 0$.
- Gamma prior for the noise variance $\sigma^2 \sim \text{Gamma}(a, b)$ with hyperparameters $a > 0$ and $b > 0$.

Theorem

Let $d \in \{1, ..., p\}$. Under the normal-gamma prior with $b = \phi/2$, the log-marginal likelihood of model $\mathcal{M}_d$ is given by

$$\begin{aligned}
\log p(\mathbf{X}|a, \phi, \mathcal{M}_d) &= \sum_{i=1}^{n} \log p(\mathbf{x}_i|a, \phi, \mathcal{M}_d) \\
&= -\frac{np}{2}\log(2\pi) - \frac{np}{2}\log(2\phi^{-1}) - n\log\Gamma(a + d/2) \\
&\quad + (a + \frac{d-p}{2})\sum_{i=1}^{n}\log(\frac{\sqrt{\phi}||\mathbf{x}_i||_2}{2}) \\
&\quad + \sum_{i=1}^{n}\log K_{a+(d-p)/2}(\sqrt{\phi}||\mathbf{x}_i||_2).
\end{aligned}$$

Simulations with $p = 50$, $n = 50$, $d = 20$



A retrospective

Main contributions of this thesis:

- Scaling up Bayesian variable selection through simple relaxations.
- Computing marginal likelihoods of Bayesian PCA models via a new theoretical result.

A retrospective: publications

- **Discussion on the Paper "A Bayesian Information Criterion for Singular Models" by Drton and Plummer**, *Journal of the Royal Statistical Society: Series B*, vol. 79, pp. 370–371 (2017)
- **Multiplying a Gaussian Matrix by a Gaussian Vector**, *Statistics & Probability Letters*, vol. 128, pp. 67–70 (2017)
- **Combining a Relaxed EM Algorithm with Occam's Razor for Bayesian Variable Selection in High-Dimensional Regression** (with Charles Bouveyron, Julien Chiquet, and Pierre Latouche), *Journal of Multivariate Analysis*, vol. 146, pp. 177–190 (2016)
- **Globally Sparse Probabilistic PCA** (with Charles Bouveyron and Pierre Latouche), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, vol. 51, pp. 976–984 (2016)

A retrospective: preprints

- **Exact Dimensionality Selection for Bayesian PCA** (with Charles Bouveyron and Pierre Latouche), Preprint HAL-01484099, Université Paris Descartes (2017).
- **Bayesian Variable Selection for Globally Sparse Probabilistic PCA** (with Charles Bouveyron and Pierre Latouche), Preprint HAL-01310409, Université Paris Descartes (2016).

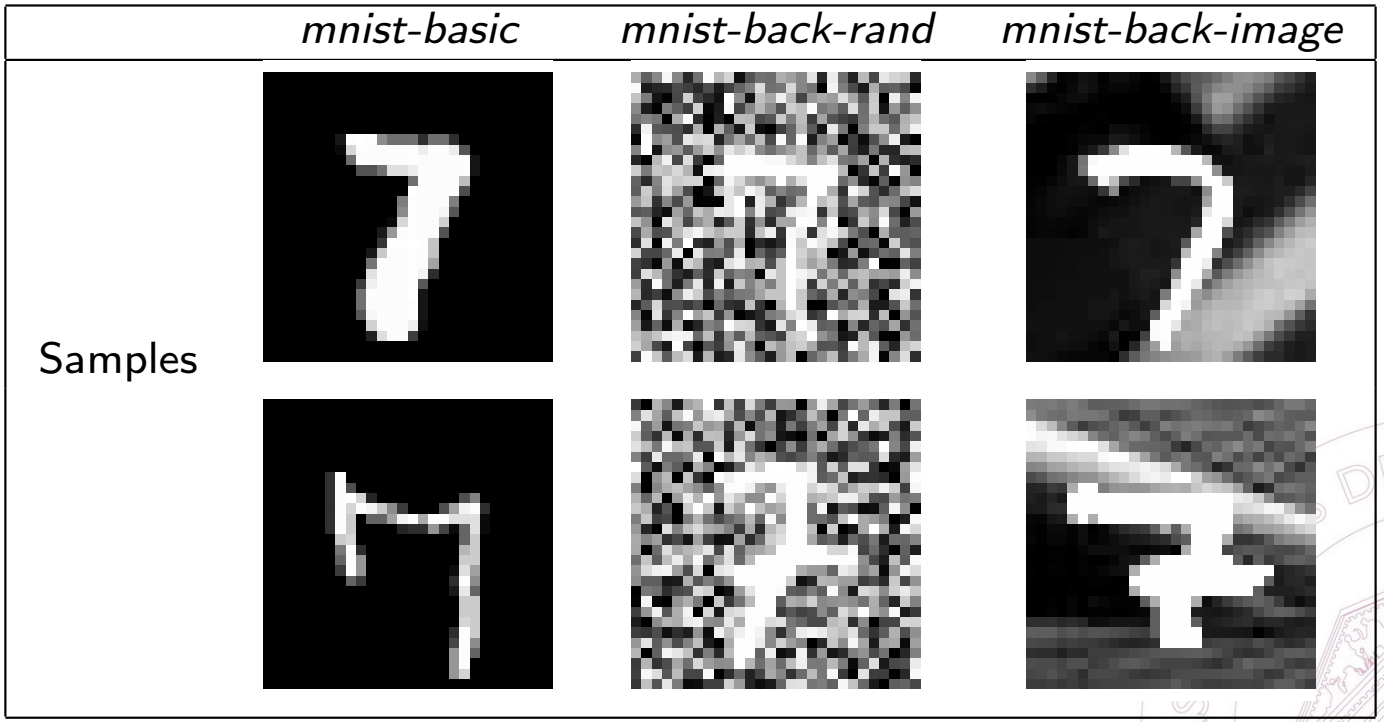
Some ongoing work

- Mixtures of Globally Sparse Probabilistic PCA for sparse and interpretable clustering and discriminant analysis
- Speeding up high-dimensional multiclass discriminant analysis via variable screening
- Deep adversarial clustering: learning deep representations for cluster analysis

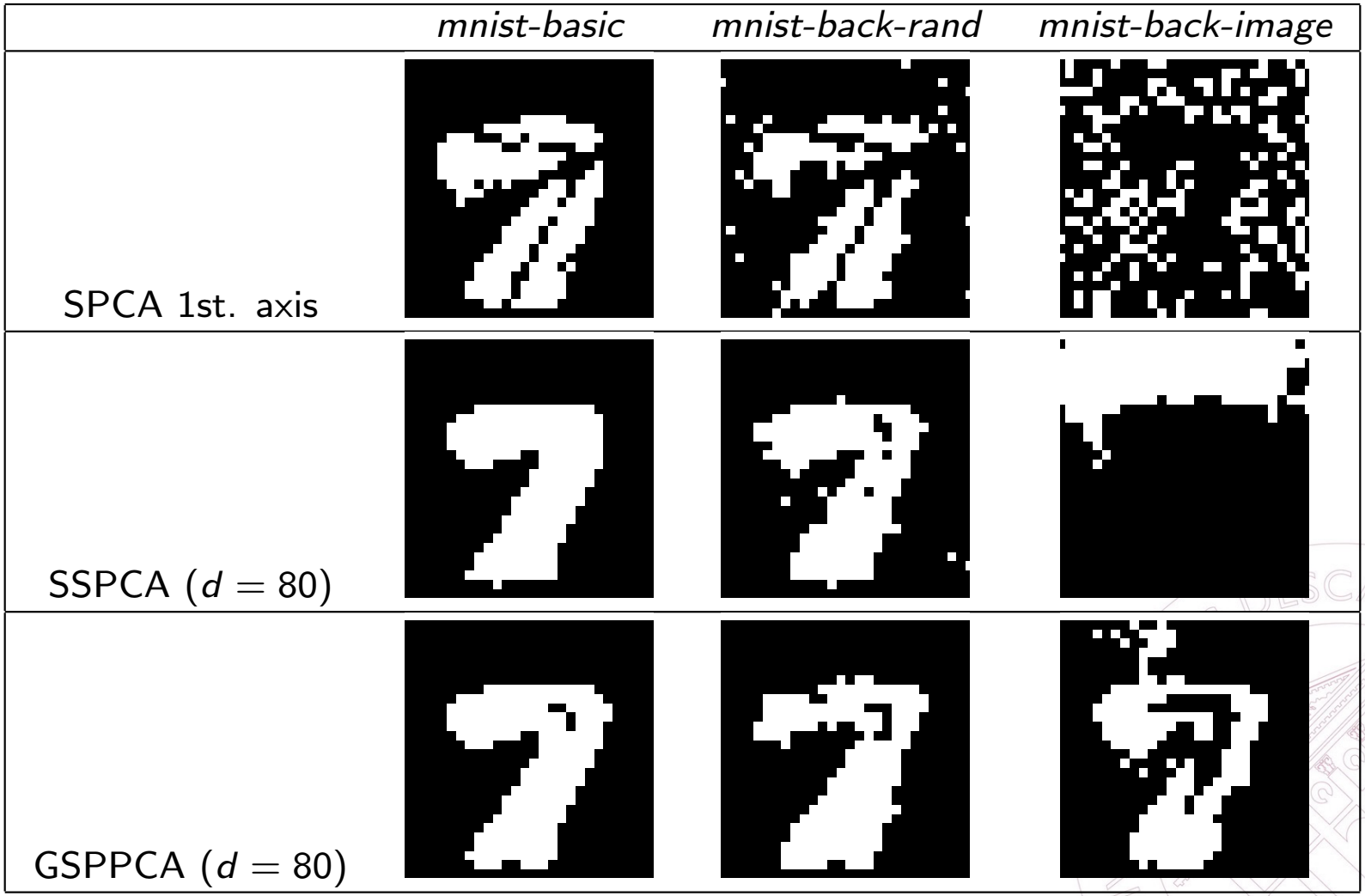

UNIVERSITÉ PARIS DESCARTES

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).



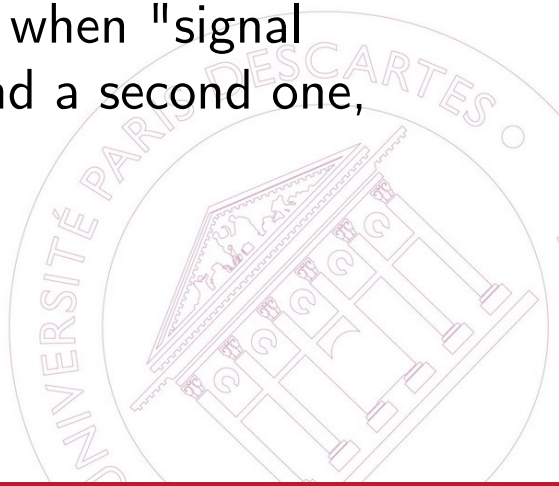
Global versus local - Variations on MNIST ($n = 500$, $p = 784$)

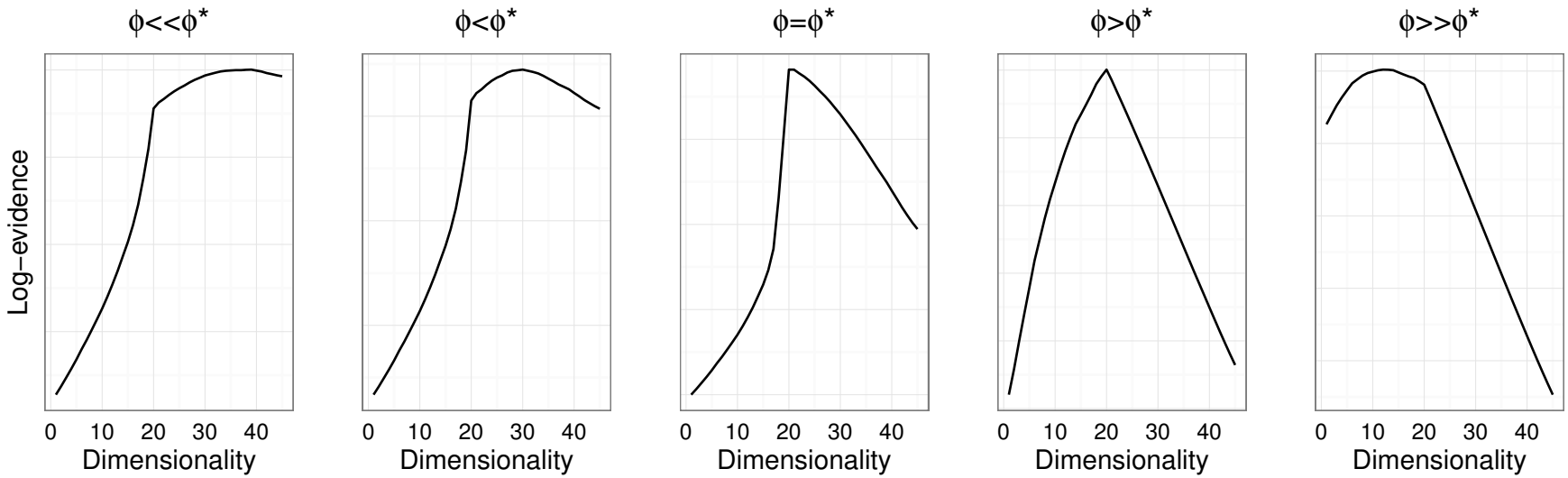


Heuristics for hyperparameter choice

For each dimension d, we choose a such that the prior of σ is roughly centered around an estimate of the noise variance.
Then, a single ϕ is chosen for all models by maximizing a heuristic criterion built in light of two statements:

- overestimation of d should be preferred to underestimation since loosing some information is much more damageable than having a representation not parsimonious enough,
- consequently, the marginal likelihood curve as a function of the dimension should have two distinct phases: a first one when "signal dimensions" are added (before the true value of d), and a second one, when "noise dimensions" are added.



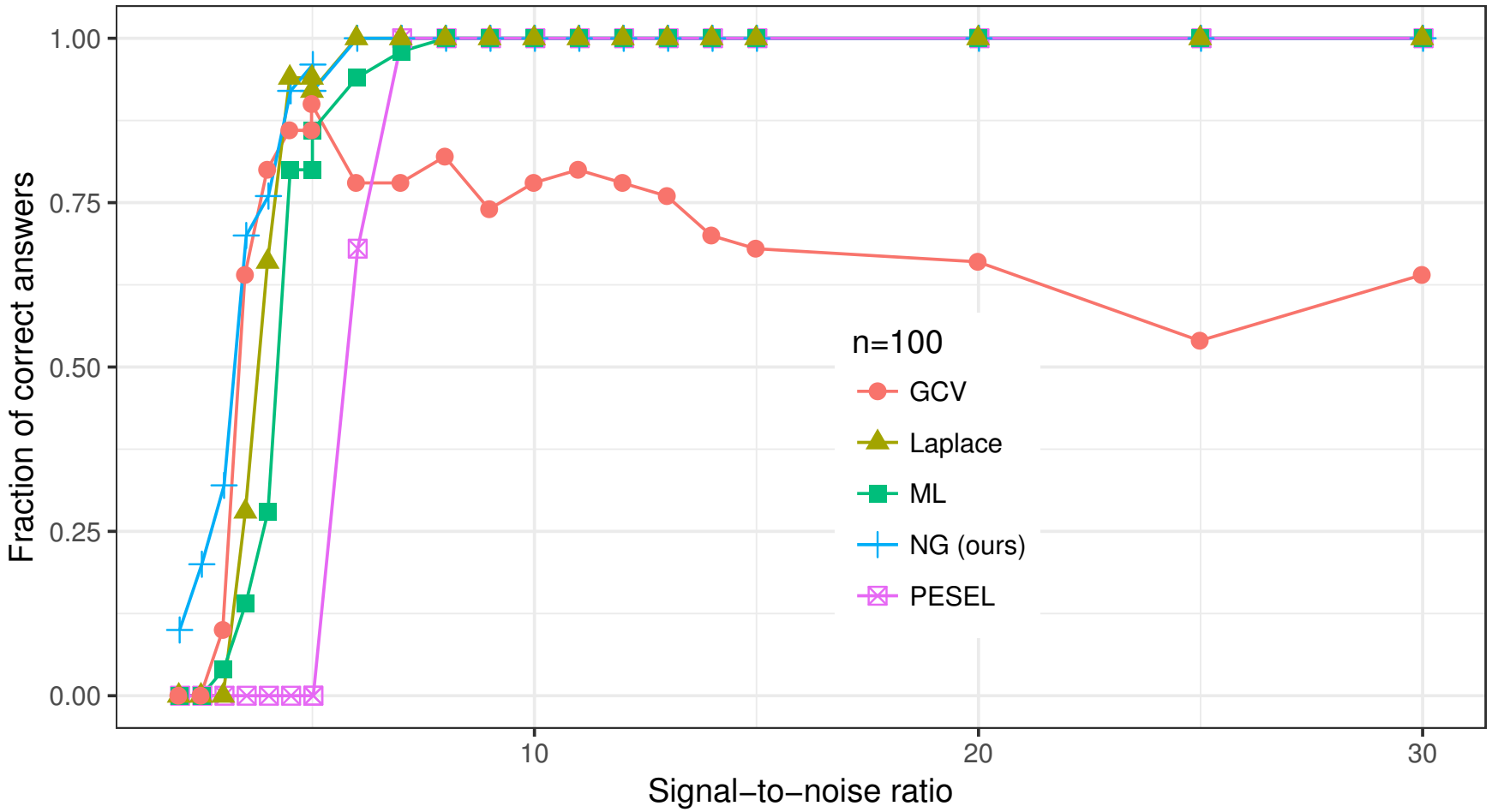


Data simulated according to a PPCA model, true d is 20.

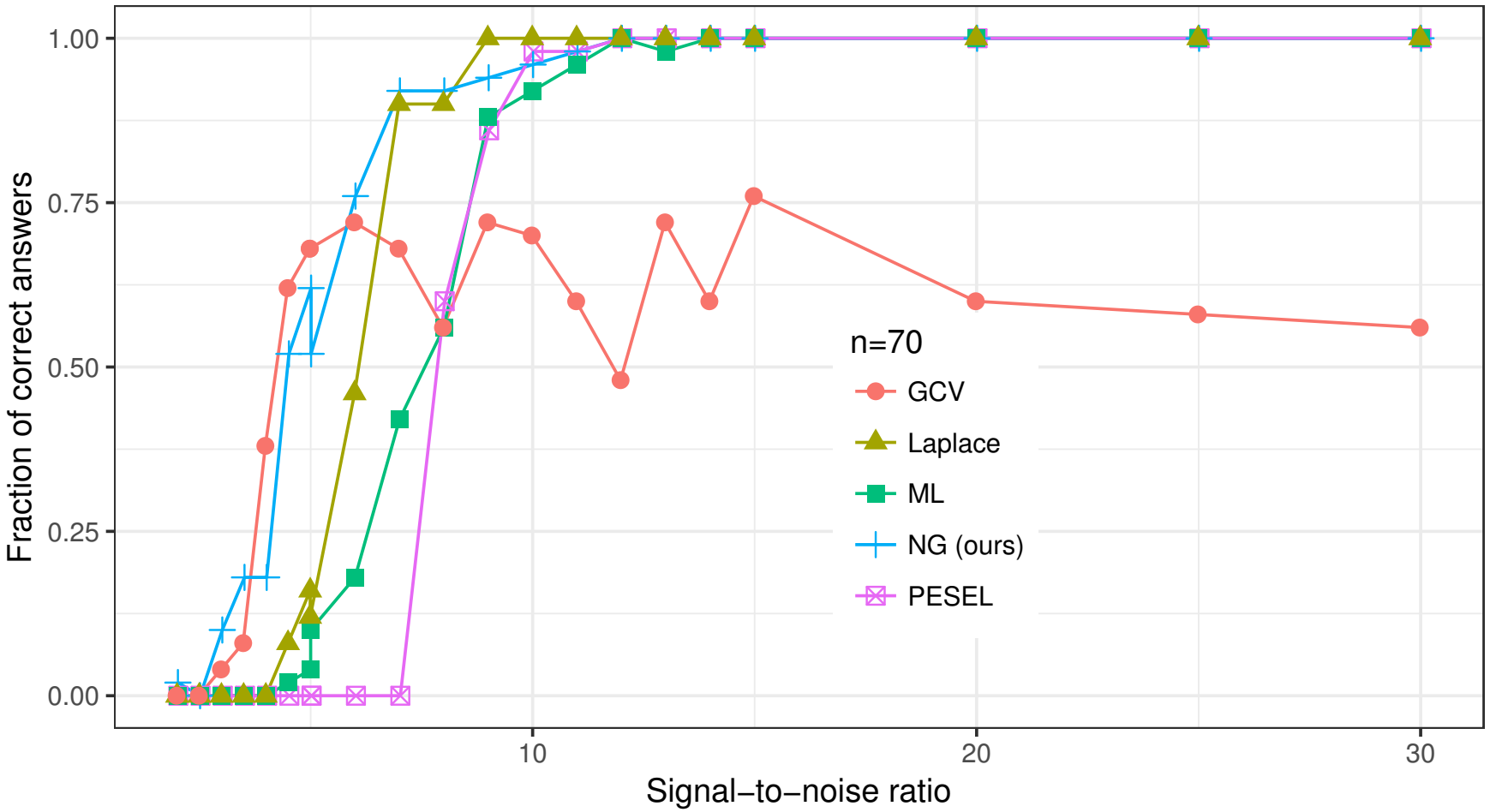
Simulation scheme

We simulate some data according to a PPCA model ($p = 50$, $d = 20$). The performance criterion is the percentage of correctly estimated dimensions for different sample sizes (50 replications for each case) of our method (NG) and competitors for different SNRs

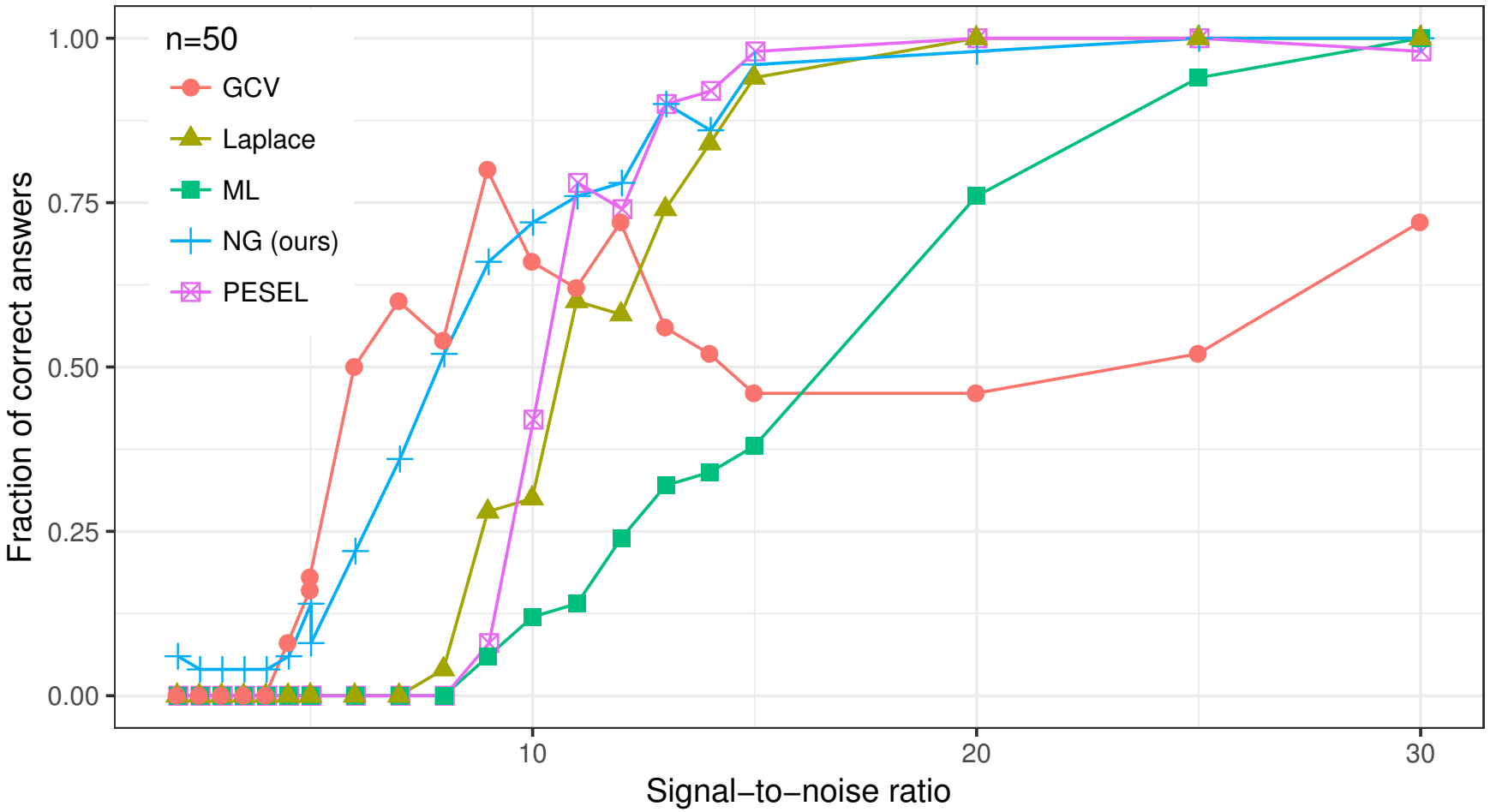
$p = 50,\ n = 100$


1.00
0.75
0.50
0.25
0.00
Fraction of correct answers
10
20
30
Signal−to−noise ratio
n=100
GCV
Laplace
ML
NG (ours)
PESEL

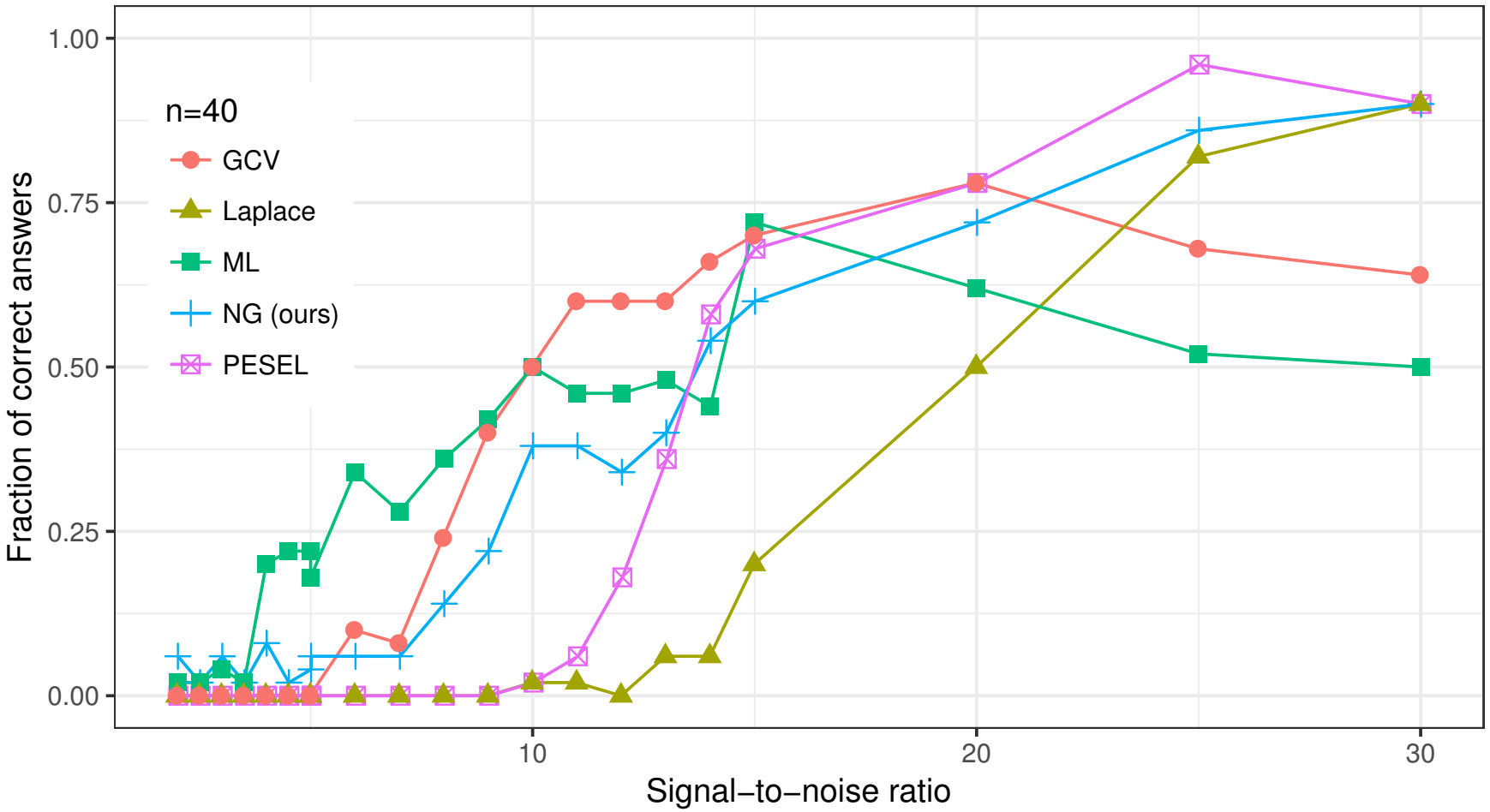
$p = 50,\ n = 70$


1.00
0.75
0.50
0.25
0.00
Fraction of correct answers
10
20
30
Signal−to−noise ratio
n=70
GCV
Laplace
ML
NG (ours)
PESEL

$p = 50,\ n = 50$


n=50
GCV
Laplace
ML
NG (ours)
PESEL
Fraction of correct answers
1.00
0.75
0.50
0.25
0.00
10
20
30
Signal−to−noise ratio

$p = 50,\ n = 40$


n=40
GCV
Laplace
ML
NG (ours)
PESEL
Fraction of correct answers
1.00
0.75
0.50
0.25
0.00
10
20
30
Signal−to−noise ratio