Globally Sparse Probabilistic PCA

Pierre-Alexandre Mattei
MAP5, UMR CNRS 8145
Université Paris Descartes
Sorbonne Paris Cité

Charles Bouveyron
MAP5, UMR CNRS 8145
Université Paris Descartes
Sorbonne Paris Cité

Pierre Latouche
SAMM, EA 4543
Université Paris 1
Panthéon-Sorbonne

Abstract

With the flourishing development of high-dimensional data, sparse versions of principal component analysis (PCA) have imposed themselves as simple, yet powerful ways of selecting relevant features in an unsupervised manner. However, when several sparse principal components are computed, the interpretation of the selected variables may be difficult since each axis has its own sparsity pattern and has to be interpreted separately. To overcome this drawback, we propose a Bayesian procedure that allows to obtain several sparse components with the same sparsity pattern. To this end, using Roweis' probabilistic interpretation of PCA and an isotropic Gaussian prior on the loading matrix, we provide the first exact computation of the marginal likelihood of a Bayesian PCA model. In order to avoid the drawbacks of discrete model selection, we propose a simple relaxation of our framework which allows to find a path of models using a variational expectation-maximization algorithm. The exact marginal likelihood can eventually be maximized over this path, relying on Occam's razor to select the relevant variables. Since the sparsity pattern is common to all components, we call this approach globally sparse probabilistic PCA (GSPPCA). Its usefulness is illustrated on synthetic data sets and on several real unsupervised feature selection problems.

1 INTRODUCTION

From the children test results of the seminal paper of Hotelling (1933) to the challenging analysis of microarray data (Ringnér, 2008), principal component analysis (PCA) has become one of the most popular tools for data-preprocessing and dimension-reduction. The original procedure consists in projecting the data onto a "principal" subspace spanned by the leading eigenvectors of the sample covariance matrix. It was later shown that this subspace could also be retrieved from the maximum-likelihood estimator of a parameter, in a particular factor analysis model called probabilisitic PCA (PPCA) (Roweis, 1998; Tipping and Bishop, 1999). This probabilistic framework led to diverse Bayesian analysis of PCA (Bishop, 1999a; Minka, 2000; Nakajima et al., 2011).

1.1 Local and global sparsity

A drawback of PCA is that the principal components are linear combinations of every single original variable, and can be difficult to interpret. To tackle this issue, several procedures have been designed to project the data onto subspaces generated by sparse vectors while retaining as much variance as possible. Many of them were based on convex or partially convex relaxations of cardinality-constrained PCA problems (Jolliffe et al., 2003; Zou et al., 2006; d'Aspremont et al., 2008), or on using sparsity-inducing prior distributions on the projection matrix (Archambeau and Bach, 2009; Guan and Dy, 2009).

However, when several principal components are computed, these various techniques do not enforce them to have to same sparsity pattern, and each component has to be interpreted individually. While individual interpretation is particularly natural in several cases – when PCA serves visualization, for example –, it is not adapted to situations where the practitioner aims at *globally* selecting which features are relevant. In these situations, a simple and popular approach has been to consider that the relevant variables correspond

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51. Copyright 2016 by the authors.

to the sparsity pattern of the first principal component (Zou et al., 2006; Zhang et al., 2012). However, this procedure has its limits, and several important aspects of the data may lie in the next principal components. For example, in the colon cancer data set studied by d'Aspremont et al. (2008), the most relevant genes were the ones selected by the *second* principal component. Another motivation for global sparsity is the fact that, in many real-life situations, the sparsity pattern of the axes computed by a sparse PCA algorithm are extremely close . This is for example the case of the three axes of the template attacks application considered by Archambeau and Bach (2009). Another interesting feature of global sparsity is the fact that, once the common sparsity pattern has been determined, performing PCA on the relevant variables yields orthogonal and uncorrelated principal components – contrarily to most sparse PCA procedures.

In this paper, we present a Bayesian approach that allows to project the data onto a *globally sparse subspace* (i.e a subspace spanned by vectors with the same sparsity pattern) while preserving a large part of the variance. To this end, we use Roweis' noiseless PPCA model (Roweis, 1998) together with an isotropic Gaussian prior on the projection matrix and a binary vector that segregates relevant from irrelevant variables. While past Bayesian PCA frameworks had to rely on variational (Archambeau and Bach, 2009; Bishop, 1999b; Guan and Dy, 2009) or Laplace (Bishop, 1999a; Minka, 2000) methods to approximate the marginal likelihood, we derive here a closed-form expression for the evidence based on the multivariate Bessel distribution. In order to be able to avoid the drawbacks of discrete model selection and to treat high-dimensional data, we also present a simple relaxation of our model by replacing the binary vector with a continuous one. Inference in this relaxed model can be performed using a variational expectation-maximization (VEM) algorithm. Such a procedure allows to find a path of models. The exact evidence is eventually maximized over this path, relying on Occam's razor (MacKay, 2003, chap. 28) to select the relevant variables.

1.2 Related work

Since the seminal papers of Jolliffe (1972, 1973), several methods have been designed to discard features in PCA (see e.g Brusco (2014) for a recent review). However, these techniques were designed to eliminate redundant, rather that irrelevant variables, and are based on combinatorial algorithms that are not really suitable for high-dimensional problems.

Another natural approach to global sparsity is ℓ_1-based regularization, which has imposed itself as one of the most versatile and efficient approaches to sparse statistical learning (Hastie et al., 2015). In a context of *structured* sparse PCA, Jenatton et al. (2009) proposed to recast sparse PCA as a penalized matrix factorization problem and suggested that limiting the number of sparsity patterns allowed within the principal vectors could improve the feature extraction quality – particularly in face recognition problems. Using the $\ell_1 - \ell_2$ norm, they derived an algorithm (hereafter referred as SSPCA) that allows to compute d sparse components with exactly $m \leq d$ sparsity patterns. However, they only considered cases where m is larger than 2 and therefore did not focus on global sparsity. Other similar approaches based on structured composite norms have been conducted by Masaeli et al. (2010) and Khan et al. (2015).

1.3 Notation

Vectors and matrices are denoted by bold cases. Given a vector $\mathbf{x} \in \mathbb{R}^p$, we define its support as $\text{Supp}(\mathbf{x}) = \{i \in \{1, ..., p\}, x_i \neq 0\}$, and its ℓ_0-pseudonorm as $||\mathbf{x}||_0 = \#\text{Supp}(\mathbf{x})$, where x_i denotes the i-th coordinate of $\mathbf{x}$. Given a vector $\mathbf{x} \in \mathbb{R}^n$, we denote $\text{diag}(\mathbf{x})$ the $n \times n$ matrix with diagonal $\mathbf{x}$ and 0 outside the diagonal. The identity matrix of dimension n is denoted by $\mathbf{I}_n$. Given a binary vector $\mathbf{z} \in \{0, 1\}^p$, we denote $\bar{\mathbf{z}}$ the binary vector of $\{0, 1\}^p$ whose support is exactly the complement of $\text{Supp}(\mathbf{z})$. Given a binary vector $\mathbf{z} \in \{0, 1\}^p$ and a matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$, we denote $\mathbf{A}_\mathbf{z}$ the extracted matrix of $\mathbf{A}$ where only the columns corresponding to the nonzero indexes of $\mathbf{z}$ have been kept. Given a mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and a positive definite covariance matrix $\mathbf{S} \in \mathbb{R}n \times n$, the density of the normal distribution is denoted $\mathcal{N}(\,\cdot\,; \boldsymbol{\mu}, \mathbf{S})$. Given a real order ν, we respectively denote by J_ν and K_ν the Bessel function of the first kind and the modified Bessel function of the third kind (Abramowitz and Stegun, 1965, chap. 10 and 11). The complete gamma function is denoted by Γ.

2 BAYESIAN VARIABLE SELECTION FOR PCA

We assume that a centered i.i.d. sample $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^p$ is observed. We wish to project it onto a d-dimensional subspace while retaining as much variance as possible. All the observations are stored in the $n \times p$ matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$.

2.1 Probabilistic PCA

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

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

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

Tipping and Bishop (1999) proved that the principal components of $\mathbf{X}$ could be retrieved using the maximum likelihood (ML) estimator $\mathbf{W}_{\text{ML}}$ of $\mathbf{W}$. Indeed, if $\mathbf{A}$ is a $p \times d$ 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, we have

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

where $\mathbf{R}$ is an arbitrary orthogonal matrix. Since (2) is true for all $\sigma > 0$, the limit noiseless setting $\sigma \to 0$ also allows to recover the principal components, while getting rid of the noise term. This convenient framework was first studied by Roweis (1998) and has proven to be useful in several contexts (Sigg and Buhmann, 2008; Ilin and Raiko, 2010).

2.2 Sparsity, variable selection and marginal likelihood

In a classical (locally) sparse PCA context, the loading matrix $\mathbf{W}$ would be expected to contain few nonzero coefficients. However, to reach global sparsity, *several entire lines* of $\mathbf{W}$ have to be further constrained to be null. In this work, we will handle variable selection using a binary vector $\mathbf{v} \in \{0,1\}^p$ whose nonzero entries correspond to relevant variables. We denote $q = ||\mathbf{v}||_0$. In the PPCA framework, this would lead to the following model for each observation:

$$\mathbf{x} = \mathbf{V}\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon} \tag{3}$$

where $\mathbf{V} = \text{diag}(\mathbf{v})$. Notice that, since the lines of $\mathbf{VW}$, corresponding the zero entries of $\mathbf{v}$ are null, the principal subspace will be generated by a basis of vectors who share the sparsity pattern of $\mathbf{v}$. Such spaces spanned by a family of vectors sharing the same sparsity pattern will be called *globally sparse subspaces.* This definition of global sparsity is closely related to the notion of *row sparsity* introduced by Vu and Lei (2013).

We assume that the coefficients of the matrix $\mathbf{W}$ are endowed with Gaussian priors $w_{ij} \sim \mathcal{N}(0, \alpha)$, for all i, j. Following the empirical Bayes framework leads to seeking the parameters $\mathbf{v}$, α and σ that maximizes the *marginal likelihood* or *evidence*

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

of the data. In previous Bayesian PCA models, the marginal likelihood was never derived because too hard to compute in practice or even intractable. Here, the evidence of the model can be expressed analytically as a univariate integral using the anisotropy of the prior on $\mathbf{W}$.

Theorem 1. *The density of* $\mathbf{x}$ *is given by*

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

A proof of this theorem is given in Appendix A. While reducing the dimension of the integration domain to one appears to be a valuable improvement, the integral of (4), albeit univariate, is delicate to compute. This is partly due to the fact that high-order Bessel functions are difficult to evaluate precisely and can have fast oscillations.

To obtain a closed-form expression of the marginal likelihood, we consider the following modification of model (3). For the relevant variables, we use the noiseless PPCA model of Roweis (1998), and we assume that the irrelevant variables are generated by a Gaussian white noise. More specifically, we write

$$\mathbf{x} = \mathbf{V}\mathbf{W}\mathbf{y} + \bar{\mathbf{V}}\boldsymbol{\varepsilon}_\mathbf{1} + \mathbf{V}\boldsymbol{\varepsilon}_\mathbf{2} \tag{5}$$

where $\boldsymbol{\varepsilon}_\mathbf{1} \sim \mathcal{N}(0, \sigma_1^2 \mathbf{I}_p)$ is the noise of the inactive variables and $\boldsymbol{\varepsilon}_\mathbf{2} \sim \mathcal{N}(0, \sigma_2^2 \mathbf{I}_p)$ is the noise of the active variables, having in mind that we aim at investigating the noiseless limit $\sigma_2 \to 0$.

In this particular case, the evidence has a closed form expression.

Theorem 2. *In the noiseless limit* $\sigma_2 \to 0$, $\mathbf{x}$ *converges in probability to a random variable* $\tilde{\mathbf{x}}$ *whose density is*

$$p(\tilde{\mathbf{x}}|\mathbf{v}, \alpha, \sigma_1) = C e^{\frac{-||\tilde{\mathbf{x}}_{\bar{\mathbf{v}}}||_2^2}{2\sigma_1^2}} ||\tilde{\mathbf{x}}_\mathbf{v}||_2^{\frac{d-q}{2}} K_{\frac{p-d}{2}} \left(\frac{||\tilde{\mathbf{x}}_\mathbf{v}||_2}{\sqrt{\alpha}} \right), \tag{6}$$

where

$$C = \frac{\alpha^{-\frac{p+d}{4}} (2\pi)^{-p/2} 2^{1-d/2}}{\Gamma\left(\frac{d}{2}\right) \sigma_1^{p-q}}.$$

A proof of this theorem is given in Appendix B. Interestingly, the distribution of the active variables exactly corresponds to a particular case of the multivariate Bessel distribution introduced by Fang et al. (1990, Def. 2.5).

If we assume that $\mathbf{v}$ is known, (6) allows us to efficiently compute the marginal log-likelihood $\sum_{i=1}^n \log \mathbb{P}(\tilde{\mathbf{x}} = \mathbf{x}_i|\mathbf{v}, \alpha, \sigma_1)$ and to optimize it with respect to the parameters α and σ_1. Indeed, α can be found using univariate gradient ascent and σ by computing the standard error of the variables which were not selected by $\mathbf{v}$. Regarding σ_1, another option is to use the maximum likelihood estimator from (3) which is simply the mean of the $p - d$ smallest eigenvalues of the sample covariance matrix (Tipping and Bishop, 1999). We found this choice to be the most effective in the experiments that we carried out.

2.3 Continuous relaxation

In spite of the results of the previous subsection, maximizing the evidence, even in the noiseless case, is particularly difficult (because of the discreteness of $\mathbf{v}$ which can take 2^p possible values). We therefore consider a simple continuous relaxation of the problem by replacing $\mathbf{v}$ by a continuous vector $\mathbf{u} \in [0,1]^p$. This relaxation is close to the one considered by Latouche et al. (2016) in a sparse linear regression framework. Denoting $\mathbf{U} = \text{diag}(\mathbf{u})$, this relaxed model can be written as

$$\mathbf{x} = \mathbf{U}\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon}. \tag{7}$$

We denote $\boldsymbol{\theta} = (\mathbf{u}, \alpha, \sigma)$ the vector of parameters. In order to maximize the evidence $p(\mathbf{X}|\boldsymbol{\theta})$, we adopt a variational approach. We view $\mathbf{y}_1, ...\mathbf{y}_n$ and $\mathbf{W}$ as latent variables.

Given a (variational) distribution q over the space of latent variables, the variational free energy is given by

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

where H denotes the differential entropy, and is an upper bound to the negative log-evidence:

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

To minimize $\mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta})$, the following mean-field approximation is made on the variational distribution:

$$q(\mathbf{Y}, \mathbf{W}) = q(\mathbf{Y})q(\mathbf{W}). \tag{9}$$

With this factorization, the variational posterior distribution q^*, which minimizes the free energy, can be derived. Note that two factorizations arise naturally. This will conveniently keep the size of the covariance matrices lower than d.

Proposition 1. *The variational posterior distribution of the latent variables which minimizes the free energy is given by*

$$q^*(\mathbf{Y}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{y}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \tag{10}$$

and

$$q^*(\mathbf{W}) = \prod_{k=1}^{p} \mathcal{N}(\mathbf{w}_k|\boldsymbol{m}_k, \mathbf{S}_k) \tag{11}$$

where

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

$$\boldsymbol{\Sigma} = \left(\mathbf{I}_d + \frac{1}{\sigma^2}\mathbf{M}^T\mathbf{U}^2\mathbf{M} + \frac{1}{\sigma^2}\sum_{k=1}^{p} u_k^2\mathbf{S}_k\right)^{-1},$$

$$\mathbf{S}_k = \left(\frac{1}{\alpha}\mathbf{I}_d + \frac{nu_k^2}{\sigma^2}\boldsymbol{\Sigma} + \frac{u_k^2}{\sigma^2}\boldsymbol{\mathcal{M}}^T\boldsymbol{\mathcal{M}}\right)^{-1},$$

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

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

A proof of this proposition, as well as a detailed computation of the free energy, are provided as supplementary material. Both rely on standard results in variational mean-field approximations (Bishop, 2006, chap. 10). The four equations of Proposition 1 will constitute the E-step of the variational expectation-maximization (VEM) algorithm used to maximized the evidence. Maximizing the free energy then leads to the following M-step updates:

$$\alpha^* = \frac{1}{dp}\sum_{k=1}^{p} \text{Tr}(\mathbf{S}_k + \mathbf{m}_k\mathbf{m}_k^T), \tag{12}$$

$$\sigma^* = \frac{\text{Tr}(\mathbf{X}\mathbf{X}^T + \mathbf{X}\mathbf{U}\mathbf{M}\boldsymbol{\mathcal{M}})}{np} + \frac{1}{np}\sum_{i=1}^{n}\sum_{k=1}^{p} u_k^2 \text{Tr}[(\boldsymbol{\Sigma} + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T)(\boldsymbol{S}_k + \boldsymbol{m}_i\boldsymbol{m}_i^T)], \tag{13}$$

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

$$u_k^* = \text{argmin}_{u \in [0,1]} \frac{u^2}{2\sigma^2}\sum_{i=1}^{n} \text{Tr}[(\boldsymbol{\Sigma} + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T)(\boldsymbol{S}_k + \boldsymbol{m}_i\boldsymbol{m}_i^T)] - u\sum_{i=1}^{n} x_{i,k}\mathbf{m}_k^T\boldsymbol{\mu}_i. \tag{14}$$

Note that the objective function of the optimization problem (14) is simply an univariate polynomial.

2.4 Final estimation

Once the VEM algorithm has converged, we still need to transform the continuous vector $\mathbf{u}$ into a binary one. To do so, the following simple procedure is considered:

- a family of p nested models is built using the order of the coefficients of $\mathbf{u}$ as a way of ranking the variables

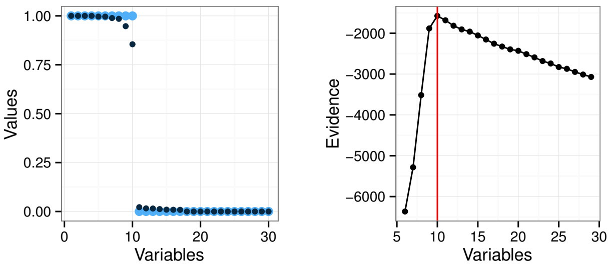


Figure 1: Variable selection with GSPPCA on the introductory example.

- 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.

Once the model is estimated, the globally sparse principal components of $\mathbf{X}$ can be computed by simply performing PCA on $\mathbf{X}_{\hat{\mathbf{v}}}$. This type of post-processing is similar to the *variational renormalization* introduced by Moghaddam et al. (2005) and extended by Journée et al. (2010).

3 NUMERICAL SIMULATIONS

3.1 An introductory example

We consider here a simple introductory example which aims at highlighting the main features of the proposed combination between the relaxed VEM algorithm and the closed-form expression of the marginal likelihood. For this experiment, $n = 50$ observations are simulated according to (3) with $p = 30$, $d = 5$ and $q = 10$. Each coefficient of $\mathbf{W}$ is drawn at random according to a standard Gaussian distribution.Fig. 1 presents the results of GSPPCA on this toy data set. The left panel presents in dark blue the coefficients of the estimated $\hat{\mathbf{u}}$ obtained after running the VEM algorithm (sorted in decreasing order) and the corresponding true values of $\mathbf{v}$ (pale blue points) used in the simulations. The right panel shows the values of evidence computed on the family of models inferred by the order of the coefficients of $\hat{\mathbf{u}}$. On this simple example, $\hat{\mathbf{u}}$ captures the true ranking of the variables and the model with the largest evidence is actually the true one.

3.2 Range of the noiseless assumption

In all the experiments that we carried out, since the noiseless PPCA model is not a true generative p-dimensional model (the random variable $\tilde{\mathbf{x}}$ belongs to a strict subspace of $\mathbb{R}^p$), we chose not to use it to generate data in our experiments. We rather chose the more realistic and natural model (3). Since this model includes a nonzero noise, it is important to know the limits of the noiseless assumption. We therefore simulated two scenarios according to (3):

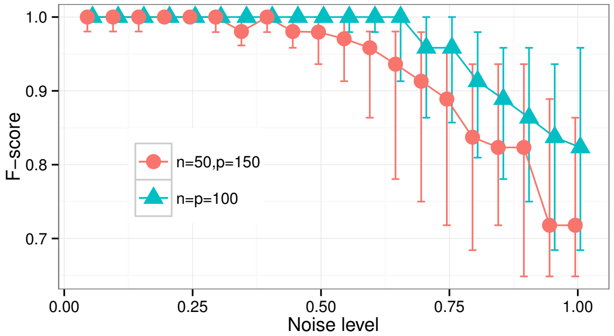


Figure 2: Median, first and third quartiles of the F-score for the experiment of subsection 3.2, based on 100 runs

- a first one with $n = 50$ and $p = 150$
- a second one with $n = p = 100$.

In both scenarios, $d = 10$, each coefficient of $\mathbf{W}$ is drawn according to a standard Gaussian distribution and a grid of different noise levels from $\sigma = 0.05$ to $\sigma = 1$ are considered. To evaluate the quality of the variable selection, we computed the F-score between $\hat{\mathbf{v}}$ and $\mathbf{v}$ on 100 runs. We recall that the F-score is the harmonic mean of precision and recall, and is closer to 1 when the selection is faithful. Unsurprisingly, when the noise rises, the quality of the variable selection diminishes. However, GSPPCA appears to be quite robust to noise, even when the data are not generated according to the underlying noiseless model.

3.3 Model selection

We compared the model selection accuracies of GSPPCA and SSPCA (Jenatton et al., 2009). Regarding SSPCA, we used the Matlab code available at the main author's webpage and chose the tuning parameter using 5-fold cross-validation on the reconstruction error.

Table 1: F-score for the experiment of subsection 3.3, based on 50 runs

	$n = p/2$	$n = p$	$n = 2p$
SSPCA-CV	0.944 ± 0.061	0.985 ± 0.022	1 ± 0
GSPPCA	0.97 ± 0.071	0.985 ± 0.34	1 ± 0

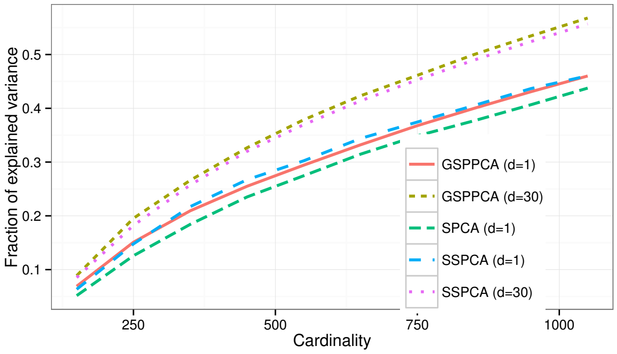
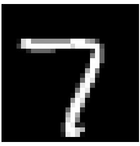
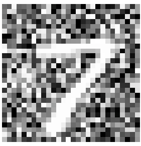
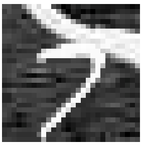
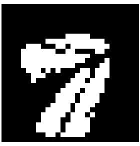
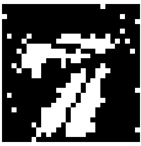
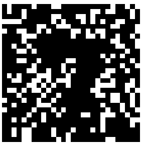
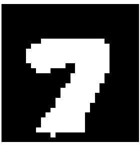
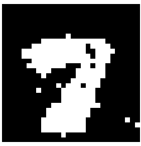
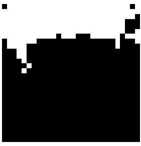
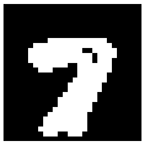
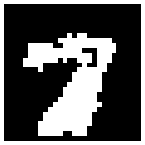
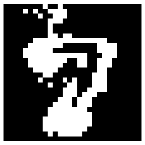


Figure 3: Percentage of variance explained by the data projected onto a 30-dimensional globally sparse subspace

Table 2: Variable selection of SPCA and GSPPCA for the three datasets of Larochelle et al. (2007), selected variables are in white

	mnist-basic	*mnist-back-rand*	*mnist-back-image*
Sample			
SPCA			
SSPCA			
GSPPCA			

We constrained the algorithm in order to obtain globally sparse solutions. Using the framework of the previous experiments, we chose $p = 100$, $d = 10$, $\sigma = 0.6$ and considered three cases regarding n. The F-scores obtained after 50 runs are stored in Table 1. We can see that the performances of the two algorithms are very similar and accurate.

3.4 Global versus local

In this subsection, we illustrate on real data sets how using GSPPCA instead of computing the leading sparse principal component for model selection can lead to selecting "better" variables – i.e variables that retain more variance or are more interpretable.

Explained variance. First, we consider the leukemia data set introduced by Golub et al. (1999) consisting in expression levels of $p = 3051$ genes for $n = 38$ leukemia patients. Given a cardinality q, we used five methods to select relevant genes:

- we computed the first q-sparse principal component using SPCA (Zou et al., 2006), GSPPCA and SSPCA with $d = 1$
- we computed the support of the globally q-sparse subspace of dimension $d = 30$ using GSPPCA and SSPCA.

For each method, we projected the data onto a 30-dimensional globally q-sparse subspace using the sparsity pattern found by the algorithm and computed the percentage of explained variance using the criterion introduced by Shen and Huang (2008) – for each method, we applied the post-processing technique of Moghaddam et al. (2005). The results are plotted on Fig. 3. While the performances of the three *local* methods are mostly similar, the variables selected by GSPPCA and SSPCA ($d = 30$) retain much more variance, and may consequently be of superior interest. Since the data is eventually projected onto a globally sparse subspace, it is not surprising that the global methods outperform the local ones. However, we can notice that this this improvement is highly significative. This means that achieving global sparsity by selecting the variables according to a single axis will lead to very suboptimal choices. It is worth noticing that, among the two global methods, GSPPCA consistently outperforms SSPCA.

Interpretability. Inspired by Hastie et al. (2015, section 8.2.3.1), we consider the problem of learning which features are relevant on three data sets of handwritten digits. We consider $n = 500$ gray-scale images (with $p = 758$ pixels) of handwritten sevens from three data sets introduced by Larochelle et al. (2007):

- *mnist-basic* which is simply a subsample of sevens from the original MNIST data set,
- *mnist-back-rand* in which random backgrounds were inserted in the images. Each pixel value of the background was generated uniformly between 0 and 255,
- *mnist-back-image* in which random patches extracted from a set of 20 grey-scale natural images were used as backgrounds for the sevens.

On these three data sets, we apply SPCA (with $d = 1$), SSPCA and GSPPCA (both with $d = 100$) in order to select $q = 200$ relevant pixels. On *mnist-basic*, even if SPCA's result is a little bit more erratic than the two others, all selections are interpretable and we can easily recognize a seven. On *mnist-back-rand* however, while the two globally sparse selections are still consistent, SPCA's pixels are more scattered and it is harder to recognize the shape of a seven. Eventually, on *mnist-back-image*, GSPPCA's selection is less smooth but a seven can still be recognized, whereas SPCA appears to randomly select pixels *almost everywhere but near the mean seven.* SSPCA seems to notice that the zone occupied by the upper bars of the sevens is of interest, but its selection does not appear interpretable. It is worth noticing that, on this last data set, applying GSPPCA with a small value of d would also lead to poorly interpretable pixels.

4 CONCLUSION

Unsupervised feature selection is an hazy and exciting problem. It becomes particularly difficult and ill-posed when no specific learning task (such as clustering) is driving it. We have proposed in this paper a new method for unsupervised feature selection based on the idea that the data may lie close to a subspace of moderate dimension spanned by a basis with a shared sparsity pattern. On several real data sets, this approach outperforms a popular method which consists in finding the sparsity pattern of the single leading principal vector of the data. These results suggest that, on many real-life high-dimensional data sets, an important part of the information cannot be captured by one-dimensional subspace approximations.

While building our framework, we derived the first closed-form expression of the marginal likelihood of a Bayesian PCA model, using the noiseless model of Roweis (1998). A thorough study of the consequences and applications of this result will be the subject of future work. It would be interesting to see if more complex priors can be used and to what extend our expression can lead to a simultaneous estimation of the sparsity level and the dimension of the latent space. Indeed, intrinsic dimension estimation, which was beyond the scope of this paper, has an enduring relationship with probabilistic versions of PCA (Minka, 2000; Bouveyron et al., 2011; Nakajima et al., 2015) and would be an interesting direction.

Acknowledgements

The authors would like to thank Florentin Damiens for helpful discussions about Bessel functions.

Appendix A. Proof of Theorem 1

Proof. Let us first consider the case where all variables are active and assume that $\mathbf{v} = (1, 1, ..., 1)$. Therefore, $\mathbf{V} = \mathbf{I}_p$ and the considered model reduces to probabilistic PCA. In this framework, we will derive the density of $\mathbf{x}$ by computing the Fourier transform of its characteristic function.

In order to compute the characteristic function of $\mathbf{x}$, we first decompose the latent vector $\mathbf{y}$ in the canonical base:

$$\mathbf{y} = y_1\mathbf{e_1} + ... + y_d\mathbf{e_d}$$

where $(\mathbf{e_i})_{i \geq d}$ is the canonical base of $\mathbb{R}^d$. We can now write the vector $\mathbf{Wy}$ as a sum of of d i.i.d variables

$$\mathbf{Wy} = y_1\mathbf{We_1} + ... + y_d\mathbf{We_d}.$$

Its characteristic function will consequently be

$$\varphi_{\mathbf{Wy}} = (\varphi_{y_1\mathbf{We_1}})^d.$$

Now, for all $\mathbf{u} \in \mathbb{R}^d$, we have

$$\varphi_{y_1\mathbf{We_1}}(\mathbf{u}) = \mathbb{E}[\exp(iy_1\mathbf{e_1}^T\mathbf{W}^T\mathbf{u})] \tag{15}$$

$$= \mathbb{E}\left[\exp\left(iy_1\sum_{k=1}^{p} w_{k1}u_k\right)\right] \tag{16}$$

but, since $w_{st} \sim \mathcal{N}(0, \alpha)$ for all s, t, we will have

$$\frac{1}{\sqrt{\alpha}||\mathbf{u}||_2}\sum_{k=1}^{p} w_{k1}u_k \sim \mathcal{N}(0, 1)$$

thus, since $\mathbf{y}$ and $\mathbf{W}$ are independent, the law of $(\sqrt{\alpha}||\mathbf{u}||_2)^{-1}y_1\sum_{k=1}^{p} w_{k1}u_k$ will be the one of a product of two standard Gaussian random variables, whose density is $1/\pi K_0(|.|)$ (Wishart and Bartlett, 1932). Therefore, we find that

$$\varphi_{y_1\mathbf{We_1}}(\mathbf{u}) = \frac{1}{\pi}\int_{-\infty}^{+\infty} K_0(|t|)e^{i\sqrt{\alpha}||\mathbf{u}||_2 t}dt$$

$$= \frac{2}{\pi}\int_{0}^{+\infty} K_0(t)\cos(\sqrt{\alpha}||\mathbf{u}||_2 t)dt$$

is simply the cosine Fourier transform of a univariate Bessel function. Using a formula in Abramowitz and Stegun (1965, p. 486), we eventually find that

$$\varphi_{y_1\mathbf{W}}(\mathbf{u}) = \frac{1}{\sqrt{1 + \alpha||\mathbf{u}||_2^2}}$$

which leads to

$$\varphi_{\mathbf{Wy}}(\mathbf{u}) = \frac{1}{(1 + \alpha||\mathbf{u}||_2^2)^{d/2}}.$$

Finally, since the noise term and $\mathbf{Wy}$ are independent, the characteristic function of $\mathbf{x}$ will be

$$\varphi_{\mathbf{x}}(\mathbf{u}) = \varphi_{\mathbf{Wy}}(\mathbf{u})\varphi_{\boldsymbol{\varepsilon}}(\mathbf{u}) = \frac{e^{-\sigma^2||\mathbf{u}||_2^2}}{(1 + \alpha||\mathbf{u}||_2^2)^{d/2}}.$$

The density of $\mathbf{x}$ is then given by the Fourier transform of its characteristic function:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^p} \int_{\mathbb{R}^p} \varphi_{\mathbf{x}}(\mathbf{u}) e^{i\mathbf{x}^T\mathbf{u}} d\mathbf{u}$$

but, since $\varphi_{\mathbf{x}}(\mathbf{u})$ is a radial function (i.e a function that only depends on the norm of its argument), its Fourier transform can be expressed as a univariate integral (Schaback and Wu, 1996) and we can write

$$p(\mathbf{x}) = \frac{||\mathbf{x}||_2^{1-p/2}}{(2\pi)^{p/2}} \int_0^{+\infty} \frac{u^{p/2} e^{-\sigma^2 u^2}}{(1+\alpha u^2)^{d/2}} J_{p/2-1}(u||\mathbf{x}||_2) du \tag{17}$$

which is the desired form for the case with no inactive variable.

In the general case, $\mathbf{v}$ is not necessarily equal to $(1, 1, ..., 1)$ but we can notice that, since $\mathbf{x}_{\mathbf{v}}$ and $\mathbf{x}_{\bar{\mathbf{v}}}$ are independent, we can write $p(\mathbf{x}) = p(\mathbf{x}_{\bar{\mathbf{v}}})p(\mathbf{x}_{\mathbf{v}}) = (\sqrt{2\pi}\sigma)^{-(p-q)} e^{\frac{-||\mathbf{x}_{\bar{\mathbf{v}}}||_2^2}{2\sigma^2}} p(\mathbf{x}_{\mathbf{v}})$. Applying (17) to the vector of nonzero coefficients of $\mathbf{x}_{\mathbf{v}}$ allows us to compute $p(\mathbf{x}_{\mathbf{v}})$ and to eventually obtain the expression of the density given by the theorem. □

Appendix B. Proof of Theorem 2

Proof. Let us first consider the case where all variables are active and assume that $\mathbf{v} = (1, 1, ..., 1)$. In that particular case, the noiseless limit reduces to Roweis' probabilistic interpretation of PCA Roweis (1998). Using Lévy's continuity theorem, it is straightforward to see that $\boldsymbol{\varepsilon_2}$ weakly converges to zero when σ_2 vanishes. Since zero is a constant, this convergence also happens to be in probability (Van der Vaart, 2000, p. 10). The variable $\mathbf{x}$ therefore converges in probability to $\mathbf{W}\mathbf{y}$. The first computations of the proof of Theorem 1 assure that the characteristic function of $\mathbf{W}\mathbf{y}$ is, for all $\mathbf{u} \in \mathbb{R}^p$,

$$\varphi_{\mathbf{W}\mathbf{y}}(\mathbf{u}) = \frac{1}{(1+\alpha||\mathbf{u}||_2^2)^{d/2}}$$

which is the characteristic function of the multivariate Bessel distribution (Kotz et al., 2001, p. 257), whose density is exactly the one of (6) in the case with no inactive variable.

Similarly to the proof of Theorem 1, we can prove (6) in the general case when $\mathbf{v}$ is not necessarily equal to $(1, 1, ..., 1)$ by invoking the independence of $\mathbf{x}_{\mathbf{v}}$ and $\mathbf{x}_{\bar{\mathbf{v}}}$. □

References

M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions.* Dover Publications, 1965.

C. Archambeau and F. Bach. Sparse probabilistic projections. In *Advances in neural information processing systems*, pages 73–80, 2009.

C. M. Bishop. Bayesian PCA. *Advances in neural information processing systems*, pages 382–388, 1999a.

C. M. Bishop. Variational principal components. In *Proceedings of the Ninth International Conference on Artificial Neural Networks*, pages 509–514, 1999b.

C. M. Bishop. *Pattern recognition and machine learning.* Springer, 2006.

C. Bouveyron, G. Celeux, and S. Girard. Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA. *Pattern Recognition Letters*, 32(14):1706–1713, 2011.

M. J. Brusco. A comparison of simulated annealing algorithms for variable selection in principal component analysis and discriminant analysis. *Computational Statistics & Data Analysis*, 77:38–53, 2014.

A. d'Aspremont, F. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research*, 9:1269–1294, 2008.

K.-T. Fang, S. Kotz, and K. W. Ng. *Symmetric multivariate and related distributions.* Chapman and Hall, 1990.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, and M. A. Caligiuri. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286 (5439):531–537, 1999.

Y. Guan and J. G. Dy. Sparse probabilistic principal component analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 185–192, 2009.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations.* CRC Press, 2015.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research*, 11:1957–2000, 2010.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *International Conference on Artificial Intelligence and Statistics*, 2009.

I. T. Jolliffe. Discarding variables in a principal component analysis. I: Artificial data. *Applied statistics*, pages 160–173, 1972.

I. T. Jolliffe. Discarding variables in a principal component analysis. II: Real data. *Applied Statistics*, pages 21–31, 1973.

I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research*, 11:517–553, 2010.

Z. Khan, F. Shafait, and A. Mian. Joint group sparse pca for compressed hyperspectral imaging. *Image Processing, IEEE Transactions on*, 24(12):4934–4942, 2015.

S. Kotz, T. Kozubowski, and K. Podgorski. *The Laplace distribution and generalizations: a revisit with applications to communications, exonomics, engineering, and finance*. Number 183. Springer Science & Business Media, 2001.

H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, pages 473–480. ACM, 2007.

P. Latouche, P.-A. Mattei, C. Bouveyron, and J. Chiquet. Combining a relaxed EM algorithm with Occam's razor for Bayesian variable selection in high-dimensional regression. *Journal of Multivariate Analysis*, 146:177 – 190, 2016.

D. J. C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge university press, 2003.

M. Masaeli, Y. Yan, Y. Cui, G. Fung, and J. G. Dy. Convex principal feature selection. In *In SIAM International Conference on Data Mining*, pages 619–628, 2010.

T. P. Minka. Automatic choice of dimensionality for PCA. In *NIPS*, volume 13, pages 598–604, 2000.

B. Moghaddam, Y. Weiss, and S. Avidan. Spectral bounds for sparse PCA: Exact and greedy algorithms. In *Advances in neural information processing systems*, pages 915–922, 2005.

S. Nakajima, M. Sugiyama, and D. Babacan. On Bayesian PCA: Automatic dimensionality selection and analytic solution. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 497–504, 2011.

S. Nakajima, R. Tomioka, M. Sugiyama, and S. D. Babacan. Condition for perfect dimensionality recovery by variational bayesian PCA. *Journal of Machine Learning Research*, 16:3757–3811, 2015. URL `http://jmlr.org/papers/v16/nakajima15a.html`.

M. Ringnér. What is principal component analysis? *Nature biotechnology*, 26(3):303–304, 2008.

S. Roweis. EM algorithms for PCA and SPCA. *Advances in neural information processing systems*, pages 626–632, 1998.

R. Schaback and Z. Wu. Operators on radial functions. *Journal of computational and applied mathematics*, 73(1):257–270, 1996.

H. Shen and J. Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99(6):1015–1034, 2008.

C. D. Sigg and J. M. Buhmann. Expectation-maximization for sparse and non-negative PCA. In *Proceedings of the 25th international conference on Machine learning*, pages 960–967. ACM, 2008.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

V. Q. Vu and J. Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

J. Wishart and M. S. Bartlett. The distribution of second order moment statistics in a normal system. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28, 10 1932. doi: 10.1017/s0305004100010690.

Y. Zhang, A. d'Aspremont, and L. El Ghaoui. Sparse PCA: Convex relaxations, algorithms and applications. In *Handbook on Semidefinite, Conic and Polynomial Optimization*, pages 915–940. Springer, 2012.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.