Electronic Journal of Statistics
Vol. 12 (2018) 3036–3070
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1450

Bayesian variable selection for globally sparse probabilistic PCA

Charles Bouveyron[1], **Pierre Latouche**[2] **and Pierre-Alexandre Mattei**[3]

[1] *Laboratoire J. A. Dieudonné, UMR CNRS 7531 & INRIA Epione, Sophia-Antipolis, Université Côte d'Azur*

[2] *Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes*

[3] *Department of computer science, IT University of Copenhagen*

Abstract: Sparse versions of principal component analysis (PCA) have imposed themselves as simple, yet powerful ways of selecting relevant features of high-dimensional data 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. This allows the practitioner to identify which original variables are most relevant to describe the data. 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. Moreover, in order to avoid the drawbacks of discrete model selection, a simple relaxation of this framework is presented. It allows to find a path of candidate 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 coming from signal processing and genomics. In particular, using unlabeled microarray data, GSPPCA is shown to infer biologically relevant subsets of genes. According to a metric based on pathway enrichment, it vastly surpasses in this context the performance of traditional sparse PCA algorithms. An R implementation of the GSPPCA algorithm is available at *http://github.com/pamattei/GSPPCA*.

Keywords and phrases: Principal components, sparsity, unsupervised learning, variable selection.

Received November 2016.

1. Introduction

From the children test results of the seminal paper of Hotelling (1933) to the challenging analysis of microarray data (Ringnér, 2008) and the recent successes of deep learning (Chan et al., 2015), 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, Sugiyama and Babacan, 2011).

1.1. Local and global sparsity

A potential drawback of PCA is that the principal components are linear combinations of every single original variable, and can therefore 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 – among these techniques are the popular ℓ_1-based SPCA algorithm of Zou, Hastie and Tibshirani (2006) or the semidefinite relaxation of d'Aspremont, Bach and El Ghaoui (2008). Another strategy is to use a sparsity-inducing prior distributions on the coefficients of the projection matrix (Archambeau and Bach, 2009; Guan and Dy, 2009; Khanna et al., 2015).

However, when several principal components are computed, these various techniques do not enforce them to have the same sparsity pattern (i.e. the same active variables), 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 to the sparsity pattern of the first principal component (Zou, Hastie and Tibshirani, 2006; Zhang, d'Aspremont and El Ghaoui, 2012). However, this procedure is limited, 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, Bach and El Ghaoui (2008), the most relevant genes were the ones selected not by the first but 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). In this setting, forcing these patterns to be equal will give the practitioner a precise idea of which variables are relevant. 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 – conversely to most sparse PCA procedures.

1.2. Related work

Since the seminal papers of Jolliffe (1972, 1973) and Robert and Escoufier (1976), 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.

A simple and scalable way of performing variable selection for PCA is to simply keep the features that have the largest marginal variance. In certain cases, this technique is theoretically sound, and was applied for instance to the analysis of electrocardiogram (ECG) data (Johnstone and Lu, 2009). Zhang and El Ghaoui (2011) also proved that it could be used as an efficient preprocessing technique to reduce the dimensionality of ultra-high dimensional problems before applying a traditional sparse PCA algorithm. However, this technique has two main drawbacks. First, it is not robust to simple transformations of the data since simply multiplying a variable by a constant may wrongfully select (or discard) it. An unfortunate consequence of this is the fact that this technique can not be applied to scaled data. Moreover, since it ignores non-marginal information, this technique will behave badly in the case of correlated features.

A more refined 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, Tibshirani and Wainwright, 2015). In a context of *structured* sparse PCA, Jenatton, Obozinski and Bach (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. They were followed by Khan, Shafait and Mian (2015) who, in a very close framework, argued that global sparsity (which they called *joint sparsity*) led to better representations of hyperspectral images. Other similar approaches based on structured composite norms have been conducted by Masaeli et al. (2010), Gu, Li and Han (2011) and Xiaoshuang et al. (2013). Ulfarsson and Solo (2008, 2011) used sparsity inducing penalties together with a PPCA model to enforce global sparsity. They proposed an algorithm called *sparse variable noisy PCA* (hereafter refered as svnPCA) and fixed the amount of penalization using the Bayesian information criterion (BIC) of Schwarz (1978).

Eventually, it is worth mentioning that global sparsity has also been investigated in other contexts, such as partial least squares regression (Liu et al., 2013) or electroencephalography (EEG) imaging (Wipf and Nagarajan, 2009; Gramfort et al., 2013).

1.3. Contributions and organization of the paper

We present in Section 2 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 the noiseless PPCA model introduced by 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 relied on variational (Bishop, 1999b; Archambeau and Bach, 2009; Guan and Dy, 2009) or Laplace (Bishop, 1999a; Minka, 2000; Sobczyk, Bogdan and Josse, 2017) 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 avoid the drawbacks of discrete model selection and to treat high-dimensional data, we also present a relaxation of our model by replacing the binary vector with a continuous one. Inference of 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.

We illustrate the behaviour of our algorithm and compare it to other methods in Section 3. In particular, we show that Bayesian model selection empirically outperforms ℓ_1-ℓ_2-based regularization on a series of tasks.

Sections 4 and 5 are devoted to two applications showcasing the features of our method. The first one concerns signal denoising with wavelets, and shows how global sparsity can surpass traditional sparse PCA algorithms within this context. The second one treats about unsupervised gene selection. Given an (unlabeled) microarray data matrix, we show how GSPPCA can select biologically relevant subsets of genes. Interestingly, we exhibit an important correlation between our exact marginal likelihood expression and a criterion of biological relevance based on pathway enrichment.

Note that this paper is an extended version of previous work (Mattei, Bouveyron and Latouche, 2016) published in the Proceedings of the 19$^{\text{th}}$ Conference on Artificial Intelligence and Statistics.

2. Bayesian variable selection for PCA

Let us assume that a centered i.i.d. sample $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^p$ is observed which one wishes to project 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

The PPCA model 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 $p \times d$ parameter matrix called the *loading matrix* and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$ is a Gaussian noise term.

This model is a particular instance of factor analysis and was first introduced by Lawley (1953). Following Theobald (1975), Tipping and Bishop (1999) confirmed that this generative model is equivalent to PCA in the sense that the principal components of $\mathbf{X}$ can be retrieved using the maximum likelihood (ML) estimator $\mathbf{W}_{\mathrm{ML}}$ of $\mathbf{W}$. Indeed, 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, we have

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

where $\mathbf{R}$ is an arbitrary orthogonal matrix.

Several Bayesian treatments of this model have been conducted by using different priors on the loading matrix. However, the marginal likelihood of these models appeared to be untractable. To tackle this issue, several computational techniques were considered. The automatic relevance determination (ARD) prior was used together with Laplace (Bishop, 1999a) or variational (Bishop, 1999b; Archambeau and Bach, 2009) approximations. Minka (2000) introduced more complex conjugate priors to perform Bayesian model selection on the dimension d of the latent space using the Laplace approximation. Combined with variational inference, several sparsity inducing priors such as the Laplace (Guan and Dy, 2009), the generalized hyperbolic (Archambeau and Bach, 2009) or the spike-and-slab (Lázaro-Gredilla and Titsias, 2011) prior were also chosen for $\mathbf{W}$.

In this work, we aim at avoiding these approximations. Our approach is to investigate in which cases the marginal likelihood can be analytically computed. To this end, we will use the fact that, within the PPCA model (1), the limit noiseless setting $\sigma \to 0$ also allows to recover the principal components. This convenient framework was first studied by Roweis (1998) and has proven to be useful in several situations. The noiseless PPCA model was used for instance to facilitate inference in the presence of missing data (Yu et al., 2010; Ilin and Raiko, 2010). More importantly in our context, it was successfully used by Sigg and Buhmann (2008) to enforce sparsity within an ℓ_1-penalized PPCA framework – which means that getting rid of the noise term is likely to be compatible with variable selection.

2.2. A general framework for globally sparse PPCA

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 rows* of $\mathbf{W}$ have to be further constrained to be null. In this work, we handle variable selection using a binary vector $\mathbf{v} \in \{0,1\}^p$ whose nonzero entries correspond to relevant variables. For technical purposes, we also denote by $\bar{\mathbf{v}}$ the binary vector of $\{0,1\}^p$ whose support is exactly the complement of $\mathrm{Supp}(\mathbf{v})$. We denote $q = ||\mathbf{v}||_0$ the number of relevant variables. In the PPCA framework, this leads 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 the rows of $\mathbf{VW}$, corresponding to the zero entries of $\mathbf{v}$, are null. Therefore, the principal subspace will be generated by a basis of vectors which shares 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* of Vu and Lei (2013).

We further assume that the coefficients of the matrix $\mathbf{W}$ are endowed with the Gaussian priors $w_{ij} \sim \mathcal{N}(0, 1/\alpha^2)$, for all i, j. Following the parametric empirical Bayes framework (Kass and Steffey, 1989) leads to seeking the parameters $\mathbf{v}$, α and σ that maximizes the *marginal likelihood* or *evidence*

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

In previous Bayesian PCA models, the marginal likelihood was never derived because it was too difficult to compute in practice or even intractable. Here, specifically, the evidence of the model can be expressed analytically as a univariate integral using the isotropy of the prior on $\mathbf{W}$. In the following, $\mathbf{x_v}$ denotes the subvector of $\mathbf{x}$ where only the variables corresponding to the nonzero indexes of $\mathbf{v}$ are kept. Given a real order ν, we denote repectively by J_ν and K_ν the Bessel function of the first kind and the modified Bessel function of the second kind (Abramowitz and Stegun, 1965, Chap. 10 and 11).

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

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

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 Equation (4), albeit univariate, falls within the category of Hankel-like integrals known to be particularly delicate to compute, even numerically. This is due to the fact that the integrand has singularities near the real axis (Ogata, 2005). To overcome this limitation, we investigate in the following subsection the use of the noiseless PPCA model to obtain a tractable expression.

2.3. A closed-form evidence for globally sparse noiseless PPCA

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, and we assume that the irrelevant variables are generated by a Gaussian white noise. More specifically, we write

$$\mathbf{x} = \mathbf{VWy} + \bar{\mathbf{V}}\boldsymbol{\varepsilon_1} + \mathbf{V}\boldsymbol{\varepsilon_2}, \tag{5}$$

where $\bar{\mathbf{V}} = \text{diag}(\bar{\mathbf{v}})$, $\boldsymbol{\varepsilon_1} \sim \mathcal{N}(0, \sigma_1^2 \mathbf{I}_p)$ is the noise of the inactive variables and $\boldsymbol{\varepsilon_2} \sim \mathcal{N}(0, \sigma_2^2 \mathbf{I}_p)$ is the noise of the active variables, having in mind that we aim at investigating the noiseless limit $\sigma_2 \to 0$. We will see that, with this particular formulation of the problem, the evidence has a closed form expression which involves the multivariate Bessel distribution, introduced by Fang, Kotz and Ng (1990, Def. 2.5).

Definition 1. *A random vector is said to have a* ***symmetric multivariate Bessel distribution*** *with parameters $\beta > 0$ and $\nu > -k/2$ if its density is*

$$\forall \mathbf{z} \in \mathbb{R}^k,\ \text{Bessel}(\mathbf{z}|\beta, \nu) = \frac{2^{-k-\nu+1}\beta^{-k-\nu}}{\Gamma(\nu + k/2)\pi^{k/2}} ||\mathbf{z}||_2^{\nu} K_\nu(||\mathbf{z}||_2/\beta).$$

Note that the modified Bessel function of the second kind K_ν involved in the Bessel density can be delicate to compute as soon as its order or its argument is large. This issue can be tackled using asymptotic expansions based on Debye polynomials (Abramowitz and Stegun, 1965, Formula 9.8.7).

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*

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

This theorem (proved in Appendix B) allows us to efficiently compute the noiseless marginal log-likelihood defined as

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

It is worth noticing that Ando (2009) also obtained a closed-form expression for the marginal likelihood in the related, but different, context of factor analysis. More specifically, he considered heavy-tailed factors and a inverse Wishart prior for the (unconstrained) noise covariance matrix. Regarding hyper-parameter tuning, if we assume that $\mathbf{v}$ is known, the regularization parameter α can be optimized efficiently using univariate gradient ascent. In fact, as stated by next proposition (proved in Appendix C), the marginal log-likelihood is even a strictly concave function of α.

Proposition 1. *The function $\alpha \mapsto \mathcal{L}(\mathbf{X}, \mathbf{v}, \alpha, \sigma_1)$ is strictly concave on $\mathbb{R}_+^*$.*

The unique optimal value $\hat{\alpha}$ can therefore be found easily using univariate convex programming.

The noise variance σ_1 can be estimated using (6) by computing the standard error of the variables which were not selected by $\mathbf{v}$. However, since model (3) is a particular instance of PPCA, it is possible to use any regular PPCA noise variance estimator. A discussion on which estimator to choose is provided in Subsection 2.7.

2.4. High-dimensional inference through a 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 $\mathbf{U} = \mathrm{diag}(\mathbf{u})$, this relaxed model can be written as

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

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 (Bishop, 2006, Chap. 10). 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

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

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

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

To minimize $\mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta})$, similarly to Bishop (1999b) and Archambeau and Bach (2009), 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, a variational expectation-maximization (VEM) algorithm can be derived. For the E-step, the variational posterior distribution q^*, which minimizes the free energy, is computed.

Proposition 2. *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

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

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

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

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

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

It is worth noticing that two factorizations arise naturally. The four equations of Proposition (2) (proved in Appendix D) will constitute the E-step of the VEM algorithm used to minimized the free energy.

We can now compute the negative free energy which will be maximized during the M-step.

Proposition 3. *Up to unnecessary additive constants, the negative free energy is given by*

$$\begin{aligned} -\mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta}) = \frac{n}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{k=1}^{p}\ln|\mathbf{S}_k| - np\ln\sigma + dp\ln\alpha - \frac{1}{2\sigma^2}\mathrm{Tr}(\mathbf{X}^T\mathbf{X}) \\ - \frac{1}{2\sigma^2}\sum_{k=1}^{p} u_k^2\mathrm{Tr}[(n\boldsymbol{\Sigma} + \boldsymbol{\mathcal{M}}^T\boldsymbol{\mathcal{M}})(\boldsymbol{S}_k + \mathbf{m}_k\mathbf{m}_k^T)] + \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{U}\mathbf{M}\boldsymbol{\mu}_i \\ + \sum_{k=1}^{p} -\frac{\alpha^2}{2}\mathrm{Tr}(\mathbf{S}_k + \mathbf{m}_k\mathbf{m}_k^T) - \frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}(\boldsymbol{\Sigma} + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T). \end{aligned} \tag{12}$$

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

$$\alpha^* = \left(\frac{1}{dp}\sum_{k=1}^{p}\mathrm{Tr}(\mathbf{S}_k + \mathbf{m}_k\mathbf{m}_k^T)\right)^{-1/2}, \tag{13}$$

$$\sigma^* = \sqrt{\frac{\mathrm{Tr}(\mathbf{X}\mathbf{X}^T - 2\mathbf{X}\mathbf{U}\mathbf{M}\boldsymbol{\mathcal{M}})}{np} + \frac{1}{np}\sum_{i=1}^{n}\sum_{k=1}^{p} u_k^2\mathrm{Tr}[(\boldsymbol{\Sigma} + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T)(\boldsymbol{S}_k + \mathbf{m}_i\mathbf{m}_i^T)]}, \tag{14}$$

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

$$u_k^* = \mathrm{argmin}_{u\in[0,1]}\frac{u^2}{2}\sum_{i=1}^{n}\mathrm{Tr}[(\boldsymbol{\Sigma} + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T)(\boldsymbol{S}_k + \mathbf{m}_i\mathbf{m}_i^T)] - u\sum_{i=1}^{n} x_{i,k}\mathbf{m}_k^T\boldsymbol{\mu}_i. \tag{15}$$

Note that the objective function of the optimization problem (15) is simply a quadratic polynomial with positive leading coefficient. Denoting

$$\xi_k = \frac{\sum_{i=1}^{n} x_{i,k}\mathbf{m}_k^T\boldsymbol{\mu}_i}{\sum_{i=1}^{n}\mathrm{Tr}[(\boldsymbol{\Sigma} + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T)(\boldsymbol{S}_k + \mathbf{m}_i\mathbf{m}_i^T)]}, \tag{16}$$

the solution can be written as

$$u_k^* = \min\{\max\{\xi_k, 0\}, 1\}. \tag{17}$$

2.5. The GSPPCA algorithm

Once the VEM algorithm has converged, the continuous vector $\mathbf{u}$ still needs to be transformed into a binary one. To do so, we rely on a technique close to the one introduced by Latouche et al. (2016) in a sparse linear regression framework. Specifically, the following simple procedure (summarized in Algorithm 1) 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. Specifically, for each $k \leq p$, the k-th element of this family is the binary vector $\mathbf{v}^{(k)}$ such that the k top coefficients of $\mathbf{u}$ are set to 1 and the others to 0.
- the marginal likelihood $\mathcal{L}$ of the noiseless model (computed using the formula of Theorem 2) is then maximized over this family of models.
- the model $\mathbf{v}$ 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 $\mathbf{X}_{\mathbf{v}}$. This type of post-processing is similar to the *variational renormalization* introduced by Moghaddam, Weiss and Avidan (2005). In the case of local sparsity, variational renormalization can be achieved using an alternating maximization scheme (Journée et al., 2010). However, the global sparsity structure greatly simplifies this procedure by reducing it to performing PCA on the relevant variables.

Algorithm 1: GSPPCA algorithm for unsupervised variable selection.

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

// VEM algorithm to infer the path of models
initialize $\mathbf{u}, \alpha, \sigma, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_n, \mathbf{m}_1, ..., \mathbf{m}_p, \mathbf{S}_1, ..., \mathbf{S}_p$ and $\boldsymbol{\Sigma}$;
repeat
  E-step from Proposition 2;
  M-step from equations (13),(14),(17);
until *convergence of the variational free energy*;

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

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

2.6. Links with other sparsity-inducing Bayesian procedures

Spike-and-slab models Model (3) may be rewritten $\mathbf{x} = \tilde{\mathbf{W}}\mathbf{y} + \boldsymbol{\varepsilon}$ where $\tilde{\mathbf{W}} = \mathbf{V}\mathbf{W}$. The prior distribution for the parameter $\tilde{\mathbf{W}}$ is similar to the spike-and-slab prior introduced by Mitchell and Beauchamp (1988) in a linear regression framework. Indeed, each coefficient $\tilde{w}_{ij}$ follows *a priori* either a Dirac

distribution with mass at zero (if $v_i = 0$) which is usually called the *spike* or a Gaussian distribution with variance $1/\alpha^2$ (if $v_i = 1$) which is usually called the *slab*. However, contrary to standard spike-and-slab models which would assume a product of Bernoulli prior distributions over $\mathbf{v}$, we see $\mathbf{v}$ here as a deterministic parameter to be inferred from the data. It is worth noticing that spike-and-slab priors have already been applied to locally sparse PCA by Lázaro-Gredilla and Titsias (2011) and Mohamed, Heller and Ghahramani (2012).

Automatic relevance determination Introduced in the context of feedforward neural networks (MacKay, 1994; Neal, 1996), automatic relevance determination (ARD) is a popular empirical Bayes procedure to induce sparsity. ARD was applied to Bayesian PCA models together with VEM algorithms in order to obtain automatic dimensionality selection (Bishop, 1999b) of local sparsity (Archambeau and Bach, 2009). In order to obtain global sparsity, ARD may be built using Model (1) together with Gaussian priors $\mathbf{w}_i \sim \mathcal{N}(0, a_i\mathbf{I}_d)$ for $i \in \{1, ..., p\}$. Similarly to Tipping (2001), maximizing the marginal likelihood would discard irrelevant variables by leading several variance parameters a_i to vanish. Interestingly, this model is somehow related to the relaxed GSPPCA model. Indeed the relaxed model (7) assumes that the i-th line of the loading matrix $\mathbf{UW}$ follows *a priori* a $\mathcal{N}(0, u_i^2/\alpha^2\mathbf{I}_d)$ distribution. The relaxed model will consequently inherit the good properties of ARD – listed for example by Wipf, Rao and Nagarajan (2011). However, similarly to Latouche et al. (2016), using the exact marginal likelihood to eventually obtain a sparse solution will avoid many classical drawbacks of ARD. First, as pointed out by Wipf and Nagarajan (2008), convergences of EM algorithms are extremely slow in the case of the ARD models. However, with our approach, since we only need the *ordering* of the coefficients of $\mathbf{u}$, we do not have to wait for the complete convergence of this parameter. In practice, in all the experiments that we carried out, we only had to perform less than a few hundreds of iterations of the algorithm to obtain convergence of the free energy in order to perform variable selection. It is worth mentioning that the fact that the objective function converges faster than the parameters of the model is a quite general property of EM algorithms (Xu and Jordan, 1996). Our procedure also avoids the lack of flexibility of ARD by computing posterior probabilities of models rather than simply giving an estimate of the best sparse model. Combined with a greedy technique similar to Occam's window (Madigan and Raftery, 1994), this feature could allow for example to perform Bayesian model averaging, which is not possible with ARD. Eventually, in the context of Bayesian PCA, ARD models such as the ones of Bishop (1999a,b) or Archambeau and Bach (2009) have to rely on approximations of the marginal likelihood while we use an exact expression.

2.7. Computational and practical considerations

Standardization Several approches to large-scale sparse PCA are based on techniques that rely on computing and ranking the marginal variances of all

variables (e.g. Johnstone and Lu, 2009; Zhang and El Ghaoui, 2011), and are therefore inefficient for standardized data. While sharing the scalability of these marginal methods (due to its linear complexity), GSPPCA is readily available for standardized data. Whether or not standardization is appropriate for PCA is a delicate issue that should be dealt with on a problem-specific basis. Notably, when all variables are on the same scale, using standardization is not customary as it may destroy some relevant information. For a detailed overview of the problem of scaling for PCA, see Bro and Smilde (2003, Sec. 3).

Intrinsic dimension estimation Since model (3) is a particular instance of PPCA, any intrinsic dimension estimator for PCA can be applied to estimate beforehand the intrinsic dimension d (see e.g. Sobczyk, Bogdan and Josse, 2017, for a recent overview of existing estimators). Although the problem of finding d is of critical importance, we assume in this work that a reasonable choice of dimension has already been made by the practitioner. While it could be tempting to use the exact noiseless marginal likelihood to select d, the close relationship existing between the noise level and d in PPCA (Tipping and Bishop, 1999; Nakajima, Sugiyama and Babacan, 2011) suggests that loosing the noise information is likely to be prejudicial for intrinsic dimension estimation.

Estimation of the noise variance As mentioned in Subsection 2.3, the standard error σ_1 of irrelevant predictors can be estimated using any regular PPCA estimator. Specifically, three important estimators are considered: the ML estimator (Tipping and Bishop, 1999), its unbiased correction (Passemier, Li and Yao, 2017), or simply the median of the variances of all features (Johnstone and Lu, 2009). Since the ML estimator is known to be biased in the high-dimensional regime, it is usually preferable to use its bias-corrected version. Both of these estimators can also be computed using the singular value decomposition (SVD) of $\mathbf{X}$. Note that since the median estimator does not need to perform this decomposition, it is therefore more suitable for large-scale inference. Note that when dealing with scaled data, the method of Johnstone and Lu (2009) reduces to taking $\sigma_1 = 1$.

Initialization strategies for the VEM algorithm Regarding the initialization of the relaxed model parameter $\mathbf{u}$, we chose to initialize all its coefficients to one. This allows to avoid premature vanishing of these coefficients which is a common drawback of ARD-like techniques (Wipf and Nagarajan, 2008). A random alternative to this strategy would be to draw the initial coefficients of $\mathbf{u}$ from a distribution with mean close to one, such as a a Beta$(9, 1)$ distribution, for instance. The noise standard error can be simply initialized using any classical PPCA noise estimator. Similarly to Latouche et al. (2016), the slab precision parameter α controls the sparsity of the VEM solution and a too small initial value is likely to lead to a too sparse solution such as the useless local optimum $\mathbf{u} = 0$. Following Biernacki, Celeux and Govaert (2003), we chose to perform short VEM runs (with less than 5 iterations) on a small grid

(typically $\alpha \in \{0.1, 1, 10\}$) and to select the value of α that led to the lowest free energy. The posterior means of the PCA loadings $\mathbf{m}_1, ..., \mathbf{m}_p$ and of the corresponding scores $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_n$ can be initialized using the singular vectors of $\mathbf{X}$. If the size of the data forbids to perform this SVD, using random Gaussian coefficients as starting points does not significantly alter the results. Finally, the initial values chosen for the posterior covariance matrices are $\boldsymbol{\Sigma} = \mathbf{I}_d$ and $\mathbf{S}_1 = ... = \mathbf{S}_p = \alpha^{-2}\mathbf{I}_d$.

Computational cost of VEM iterations Thanks to the factorizations that arised naturally during variational inference, the cost of each VEM iteration is of order $O(pnd^3)$ which is linear *both in sample size and dimensionality* and therefore particularly suitable for high-dimensional inference.

Large scale inference In the GSPPCA algorithm, SVD is used twice. Indeed, the top d singular vectors can be used to initiate the VEM algorithm and the $p - d$ smallest singular values can be used to estimate the noise variance (both as a VEM starting point for σ and as an estimator for σ_1). This can be done efficiently using a *truncated SVD algorithm*. We chose specifically the R interface (Qiu and Mei, 2016) of the Spectra[1] C++ library. However, for very large scale problems, even a fast truncated SVD algorithm appears computationally prohibitive. To tackle this issue, we offer two alternatives. First, the posterior parameters initialized using the eigenvectors can be initialized using random standard Gaussian coefficients. Moreover, following Johnstone and Lu (2009), the noise variance can be estimated using the median of the variable variances. This leads to a "SVD-free" version of the GSPPCA algorithm suitable for very large scale problems.

Stopping criterion In all of our experiments, the stopping criterion of the VEM algorithm is when the relative change of free energy gets below a fixed tolerance. More sophisticated criteria – such as one based on Aitken's acceleration (McLachlan and Krishnan, 2008, Sec. 4.9) – could be used, but since we do not need a very precise estimate of $\mathbf{u}$ (only the ordering of the coefficients is relevant), it is not mandatory.

Initialization strategy for the gradient ascents The evaluation of the exact noiseless marginal likelihoods requires to solve p univariate convex optimization problems. While this procedure can be easily parrallelized, choosing poor initial starting points may necessitate to perform a high number of gradient steps. For a given model $\mathbf{v}$, an efficient way to initialize α is to use the method of moments as a crude estimate of maximum marginal likelihood. The fact that the prior variance of all relevant variables is equal to d/α^2 leads to the choice $\alpha = \sqrt{dnq}/||\mathbf{X}_\mathbf{v}||_\mathrm{F}$. We noticed that this method of moments estimate also lead to good results as an initialization strategy for the VEM algoritm.

[1] http://yixuan.cos.name/spectra/index.html.

Model selection speedup A simple way to reduce the number of gradient ascents is to rely on the links between our relaxed model and ARD. Specifically, we can discard *before* the model selection step all the variables corresponding to the subset $\{i \in \{1, ..., p\} | u_i = 0\}$ where $\mathbf{u}$ is the relaxed model parameter obtained after convergence of the VEM algorithm. When $\mathbf{u}$ is sparse, this will bring about a substantial speedup. Notice that, since ARD is known to converge slowly, $\mathbf{u}$ is unlikely to be sparse enough and the model selection step is still necessary.

Evaluation of Bessel functions We used the R package `Bessel` (Mäechler, 2013), which allows to compute Bessel functions using either the subroutines of Amos (1986) or accurate approximations based on Debye polynomials (Abramowitz and Stegun, 1965, Formula 9.8.7), which give usually more stable results.

3. Numerical simulations

This section aims at highlighting the specific features and abilities of the proposed GSPPCA approach on simulated and real data sets.

3.1. An introductory example

We consider here a simple introductory example to illustrate the proposed combination between a 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 and the noise variance is equal to 0.1. Figure 1 presents the results of GSPPCA on this toy data set. The left panel presents in dark blue the coefficients of the estimated $\mathbf{u}$ 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 $\mathbf{u}$. On this simple example, $\mathbf{u}$ 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.

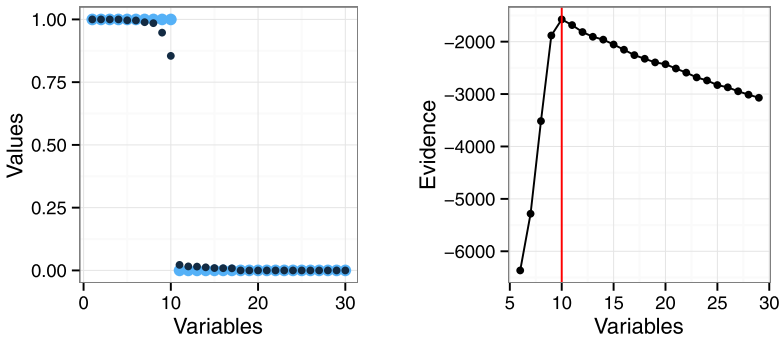


FIG 1. *Variable selection with GSPPCA on the introductory example ($n = 50$, $p = 30$, $q = 10$). Left: values of* $\mathbf{u}$ *after convergence of the VEM algorithm (dark blue) and true values of the binary sparsity pattern* $\mathbf{v}$ *(pale blue). The VEM algorithm correctly ranks the variables.* Right: *values of the exact (log-)evidence computed over the path of models derived by the VEM algorithm. The true model ($q = 10$) is recovered.*

We therefore simulated two scenarios according to Model (3): a first one with $n = 40$ observations and a second one with $n = 200$. In both scenarios, $p = 200$, $d = 10$, $q = 20$, and each coefficient of $\mathbf{W}$ is drawn according to a standard Gaussian distribution. The sparsity pattern chosen is simply

$$\mathbf{v} = (\overbrace{1, ..., 1}^{\text{20 times}}, \overbrace{0, ..., 0}^{\text{180 times}})^T. \tag{18}$$

In this simple simulation scheme, the signal-to-noise ratio (SNR) may be defined as $\text{SNR} = \frac{1}{p\sigma^2}\mathbb{E}_{\mathbf{W}}[(\mathbf{VW})^T\mathbf{VW}]p\sigma^2 = \frac{dq}{p\sigma^2}$. We chose a linear grid of 20 SNR ranging from 0.1 (most difficult scenario) to 3 (easiest scenario) and generated 100 datasets for each noise level. 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 SNR gets close to zero, the quality of the variable selection diminishes. However, GSPPCA appears to be quite robust to noise, even though the data are not generated according to the underlying noiseless model. Indeed, even in the case where $n = 40$, we observe an almost perfect recovery as long as $\text{SNR} > 0.5$.

3.3. Model selection

In this subsection, we compare the model selection accuracies of three global methods – GSPPCA, ARD and SSPCA (Jenatton, Obozinski and Bach, 2009) – and a local one – SPCA (Zou, Hastie and Tibshirani, 2006).

Simulation setup While the simple simulation setup of Subsection 3.2 conveniently allowed to compute the SNR in closed formed in order to assess the

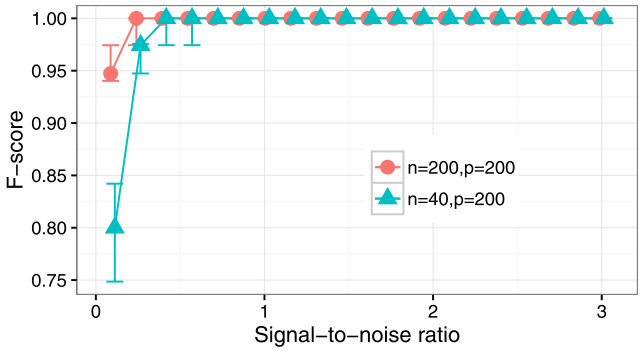


FIG 2. *Median, first and third quartiles of the F-score for different signal-to-noise ratios, based on 100 runs. A high F-score indicates good support recovery.*

range of the noiseless assumption, we introduce here a more realistic scheme by considering a finer correlation structure as well as a non-Gaussian noise. Specifically, first we generate n i.i.d observations $(\mathbf{z}_1, ..., \mathbf{z}_n)$ following multivariate normal distribution $\mathcal{N}(0, \mathbf{R})$ where $\mathbf{R} = \text{diag}(\mathbf{R}_1, ..., \mathbf{R}_4)$ is a 4-blocks diagonal matrix where R_ℓ is such that $r_{\ell ii} = 0.3$ and $r_{\ell ij} = \rho$ for $i, j = 1, \ldots, p/4$ and $i \neq j$. Then, a globally sparse PCA model is obtained as followed. First, PPCA is performed on the sample $(\mathbf{z}_1, ..., \mathbf{z}_n)$, which leads to a non-sparse ML estimate $\mathbf{W}_{\text{ML}}$ for the loading matrix. Then, given a sparsity pattern $\mathbf{v} \in \{0, 1\}^p$ and denoting $\mathbf{V} = \text{diag}(\mathbf{v})$ as before, the loading matrix matrix is "globally sparsified" by considering $\mathbf{VW}_{\text{ML}}$. The final observations are eventually generated according to the non-noiseless model

$$\forall i \leq n, \quad \mathbf{x}_i = \mathbf{VW}_{\text{ML}}\mathbf{y}_i + \boldsymbol{\varepsilon}. \tag{19}$$

The simple sparsity pattern (18) is kept and the vectors $\mathbf{y_1}, ..., \mathbf{y}_n$ are standard Gaussian as in regular PPCA. Regarding the noise term $\boldsymbol{\varepsilon}$, we consider two scenarios. A first one with Gaussian noise and a second one with Laplacian noise, both centered with unit variance. We choose $p = 200$, $d = 10$, $q = 20$ and consider five cases for the sample size: $n = p/5$, $p/4$, $n = \lfloor p/3 \rfloor$, $n = p/2$ and $n = p$. More classical $n > p$ cases are not presented here since regular PCA is known to perform well in this context and variable selection thus may not be of great use (Johnstone and Lu, 2009). Each experiment was repeated 50 times.

Implementation and model selection criteria Regarding ARD, we adapted the VEM algorithm for GSPPCA by replacing the condition $\mathbf{u} \in [0, 1]^p$ by $\mathbf{u} \in \mathbb{R}_+^p$, and by waiting for the convergence of this parameter. 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. We constrained the algorithm in order to obtain globally sparse solutions. For SPCA, we used the `elasticnet` R package and an *ad-hoc* method

TABLE 1
F-score×100 for the model selection experiment of subsection 3.3 with Gaussian noise (mean and standard deviation for 50 replications). A high F-score indicates good support recovery.

	$n = p/5$	$n = p/4$	$n = \lfloor p/3 \rfloor$	$n = p/2$	$n = p$
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
ARD	62.9 ± 5.68	62 ± 7.21	57.8 ± 6.85	58.3 ± 5.4	57.7 ± 5.91
GSPPCA	$\mathbf{87.8 \pm 6.56}$	$\mathbf{92 \pm 3.63}$	$\mathbf{96.8 \pm 2.34}$	$\mathbf{99.2 \pm 1.4}$	$\mathbf{100 \pm 0}$

TABLE 2
F-score×100 for the model selection experiment of subsection 3.3 with Laplacian noise (mean and standard deviation for 50 replications). A high F-score indicates good support recovery.

	$n = p/5$	$n = p/4$	$n = \lfloor p/3 \rfloor$	$n = p/2$	$n = p$
SPCA	20.8 ± 0.6	21.3 ± 0.6	21.6 ± 0.8	21.8 ± 0.6	25.3 ± 1.7
SSPCA	60.6 ± 22.4	63.9 ± 25.2	82.7 ± 18.1	$\mathbf{94.2 \pm 10.2}$	97.4 ± 9.5
ARD	47.0 ± 5.41	47.6 ± 4.87	48.5 ± 5.07	50.1 ± 5.17	55.0 ± 5.45
GSPPCA	$\mathbf{66.4 \pm 8.2}$	$\mathbf{72.6 \pm 9}$	$\mathbf{79.5 \pm 8.6}$	89.4 ± 5.1	$\mathbf{99.2 \pm 1.4}$

by selecting enough variables to explain 99% of the total variance. We also tried to apply another globally sparse algorithm, vsnPCA-ℓ_0 from Ulfarsson and Solo (2011). However, their use of the Bayesian information criterion (BIC) led to selecting very few variables. This is not very surprising: since BIC is an asymptotic sparsity criterion, it is thus likely to perform poorly when p is larger than n.

Results Tables 1 and 2 reports the mean and standard error of the F-score for the experiments described is this subsection. The three globally sparse methods vastly outperform SPCA, which is unable to identify the particular structure of the data. It appears that ARD selects too many variables, but usually retains the good ordering of the variables. This emphasizes the necessity of our approach. When p is larger than $n/2$, both SSPCA and GSPPCA perform very well, GSPPCA being slightly better in the Gaussian noise case. It is not surprising to see SSPCA adapt efficiently to Laplacian noise because cross-validation is a model-free technique and is more likely to outperform model-based techniques when the data is not generated according to the model distribution. However, when n is smaller than $p/2$, GSPPCA significantly outperforms SSPCA in both noise scenarios. This reminds the fact that, in many $p \gg n$ situations, Bayesian model selection empirically outperforms ℓ_1-based methods (Celeux et al., 2012; Latouche et al., 2016).

3.4. Global versus local

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

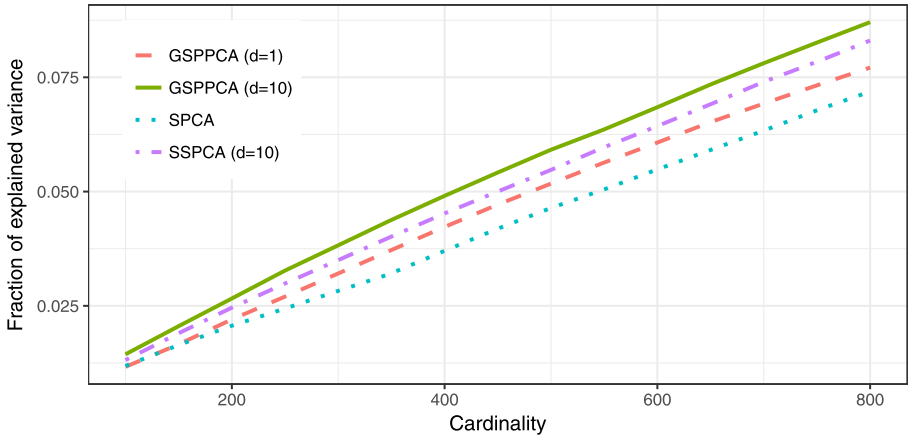


FIG 3. *Percentage of variance explained by projecting the microarray data onto a 10-dimensional globally sparse subspace.*

Explained variance We consider the data base from the `breastCancerVDX` R package (Schroeder et al., 2011), consisting in expression levels of $p = 5391$ genes for $n = 344$ breast cancer patients. More details regarding this data set – including the preprocessing technique used – are given in Appendix F. Given a cardinality q, we applied four methods to select relevant genes:

- we computed the first q-sparse principal component using SPCA (Zou, Hastie and Tibshirani, 2006) and GSPPCA with $d = 1$
- we computed the support of the globally q-sparse subspace of dimension $d = 10$ using GSPPCA and SSPCA.

For each method, we projected the data onto a 10-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, Weiss and Avidan (2005). The results are plotted on Figure 3. GSPPCA with $d = 1$ outperforms its local competitor SPCA by a significant margin, which means that the VEM algorithm finds more relevant genes than the ℓ_1 approach of Zou, Hastie and Tibshirani (2006) – this is consistent with the experiments of Archambeau and Bach (2009). Both global methods explain consistently more variance than local ones. This fact is not surprising since the data is indeed projected onto a globally sparse subspace, but the significance of this variance gap highlights the fact that different dimensions lead to very different sparsity patterns. This means that projecting the data onto a single sparse axis is likely to lead to an important information loss (this fact is confirmed in Section 5). The variables selected by GSPPCA retain significantly more variance than the ones selected by SSPCA, and may consequently be of superior interest.

Interpretability Inspired by Hastie, Tibshirani and Wainwright (2015, Sec. 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.

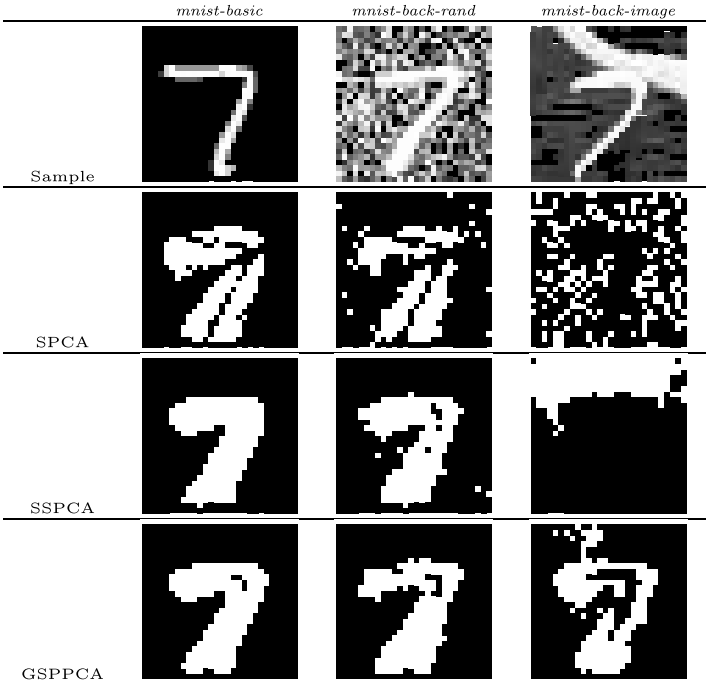
4. Application to signal denoising

In this section, we focus on a first possible application of GSPPCA for signal denoising through the sparsification of a wavelet decomposition. PCA is indeed a popular way to denoise multivariate signals (Aminghafari, Cheze and Poggi, 2006; Johnstone and Lu, 2009). To illustrate the potential interest of GSPPCA in this context, we consider hereafter two simulation scenarios, each using a specific form of signal and wavelet. The simulation scenarios are as follows:

- Scenario A: it consists in a square wave signal with 6 states of different lengths. The observed signal is sampled with a time step of 5×10^{-3} with an additional Gaussian noise with zero mean and 0.2 standard deviation. The Haar wavelet is used here for signal reconstruction.
- Scenario B: the original signal is here a mixture of 4 Gaussian densities. The observed signal is also sampled with a time step of 5×10^{-3} with an additional Gaussian noise with zero mean and 0.2 standard deviation. The Daubechies D8 wavelet is used here for signal reconstruction.

Figure 4 presents the original signals and observed signals for scenarios A and B. In both cases, $n = 100$ signals were sampled during the training phase and decomposed as $p = 175$ wavelet coefficients. For signal denoising, GSPPCA is

TABLE 3
Variable selection of SPCA and GSPPCA for the three datasets of Larochelle et al. (2007), selected variables are in white.



applied on the $n \times p$ wavelet coefficient matrix to extract $d = 10$ globally sparse principal axes. Then, a new sampled signal is projected on those extracted principal axes and back-projected in the original wavelet domain. It is worth mentioning that the estimated value for $q = \|\mathbf{v}\|_0$ is 17 on scenario A and 15 on scenario B.

As an illustration, we plotted on Figure 4 the denoising results for newly sampled signals A and B with GSPPCA. We used the same projection-reconstruction protocol for PCA, thresholded PCA (PCA loading smaller than 1×10^{-3} are set to 0) and SPCA (λ is chosen such that 99% of the PCA projected variance is conserved). Denoising results obtained with those methods are also supplied on Figure 4. First, on both signal A and B, PCA achieves a very satisfying denoising and thus confirms his validity in this context. One can also show that a simple thresholding of the PCA loadings allows a clear denoising improvement and turns out to be competitive with the one performed by SPCA. The SPCA result is here somehow disappointing due to the fact that the sparsity is not global and most wavelet levels stay active in the final reconstruction. Finally, the global sparsity of GSPPCA retains only a few wavelet levels and achieves here the best reconstruction in both scenarios.

Finally, Table 4 presents the reconstruction error (sum of squared errors) averaged on 50 test signal reconstructions, on the two simulation scenarios.

TABLE 4
Reconstruction error (sum of squared errors) for wavelet signal denoising on the two simulation scenarios (results are averaged on 50 signal reconstructions). Standard deviations are also provided.

Scenario	Wavelet	PCA	tPCA	SPCA	GSPPCA
A	9.516±0.819	2.719±0.439	2.484±0.372	2.480±0.371	**2.283±0.344**
B	8.156±0.725	1.390±0.351	1.253±0.343	1.406±0.354	**1.193±0.337**

The results confirms the observations made on Figure 4. GSPPCA achieves particularly good performances on both scenarios and thus imposes itself as a competitive tool for signal denoising. Moreover, the GSPPCA reconstruction uses fewer wavelet levels and is therefore visually smoother.

5. Application to unsupervised gene selection

Considering again the breast cancer data set previously studied in Section 3, we address here the issue of the biological significance of the selected genes. To this end, we will use the *pathway enrichment index* (PEI) introduced by Teschendorff et al. (2007) and used in a sparse PCA framework by (Journée et al., 2010).

5.1. Pathway enrichment as a measure of biological significance

In this subsection, we briefly review how the PEI can be computed in order to evaluate the quality of a given subset of genes. For more details on the PEI, see Teschendorff et al. (2007) or Journée (2009), and on hypergeometric tests and enrichment, see Rivals et al. (2007).

Suppose that using a microarray data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ where each variable corresponds to a gene, an algorithm infers a subset $\mathbf{s} \subset \{1, ..., p\}$ of genes. A way to assess its biological significance is to compare $\mathbf{s}$ to many other subsets *which are known to be biologically relevant*. In this case, the biologically relevant subsets are defined by *biological pathways*, and are therefore groups of genes involved in series of biochemical reactions linked to a certain biological function. Let us denote these known subsets $\mathbf{b}_1, ..., \mathbf{b}_N \subset \{1, ..., p\}$. For our breast cancer experiment, we use the $N = 1116$ pathways from the Reactome database (Fabregat et al., 2016) included in the R package `reactomePA` (Yu and He, 2016). For $k \leq N$, the *enrichment* of $\mathbf{s}$ in the k-th pathway of this list is the statistical significance of its overlap with $\mathbf{b}_k$, evaluated using the *hypergeometric test*. More specifically, for each $k \leq N$, the null hypothesis of this test is that the genes in $\mathbf{s}$ are chosen uniformly at random from the total gene population. Under this hypothesis, the test statistic $\#(\mathbf{s} \cap \mathbf{b}_k)$ follows a hypergeometric distribution and a p-value can be computed to assess the statistical significance of the overlap. Because we are conducting one test for each pathway considered, these p-value are then adjusted using the Benjamini-Hochberg procedure to control the false discovery rate (Benjamini and Hochberg, 1995). The subset

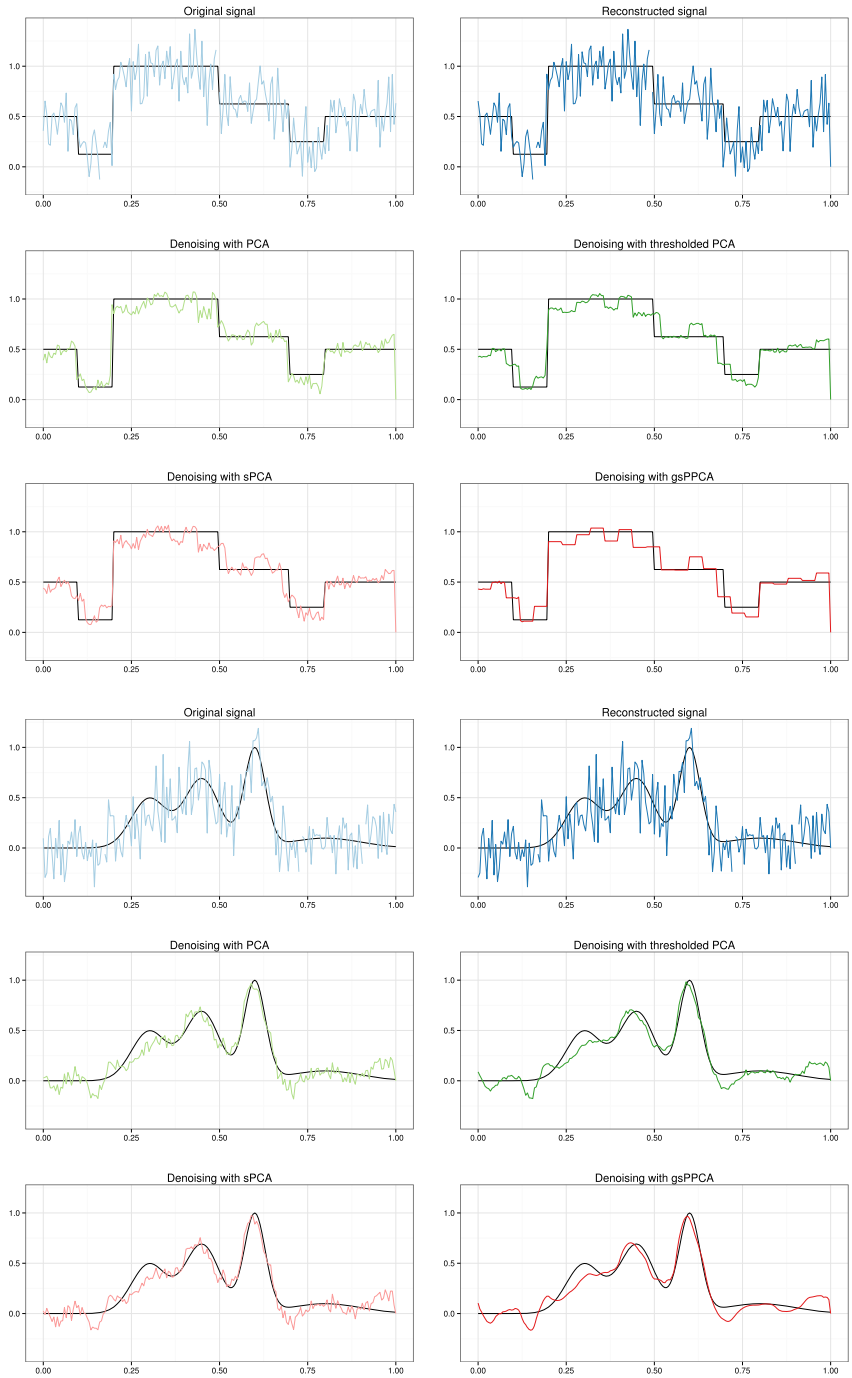


FIG 4. *Denoising results for signals A (top) and B (bottom) with PCA, thresholded PCA, SPCA and GSPPCA.*

TABLE 5
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**

s is eventually declared enriched for a certain pathway if the adjusted p-value of the corresponding hypergeometric test is lower than 0.01. The PEI is finally defined as the percentage of enriched pathways in the Reactome family.

5.2. Results

We compare in Table 5 the PEI obtained by GSPPCA with $d = 10$, SPCA and thresholded PCA for several fixed cardinalities. Similarly to Zou, Hastie and Tibshirani (2006), the two local methods are computing a single sparse axis. As in Journée et al. (2010) SPCA appears to give slightly better results than thresholded PCA. GSPPCA significantly outperforms the two other methods. This means that the genes selected by GSPPCA are consistently more associated with the Reactome pathways, and are therefore more interpretable. This highlights the fact that projecting the data onto a globally sparse subspace of dimension higher than one leads to significantly more interpretable and biologically plausible results. Regarding the estimation of the sparsity level, choosing the one that explains 99% of the variance led SPCA to selecting 4466 genes, which is difficult to interpret. For thresholded PCA, we selected the sparsity level using a criterion proposed by Teschendorff et al. (2007). Even though it led to the sparsest solution, its PEI was very small. Regarding GSPPCA, the noiseless marginal log-likelihood and the PEI of the corresponding models are plotted on Figure 5. We can see that the marginal likelihood peak corresponds to highly interpretable genes: more than 5% of the biological pathways in the Reactome family have a significant overlap with the genes selected by GSPPCA. Furthermore, models with a lower marginal likelihood have generally a lower PEI. To a certain extend, this shows that our marginal likelihood expression can stand as an indicator of biological significance.

6. 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 con-

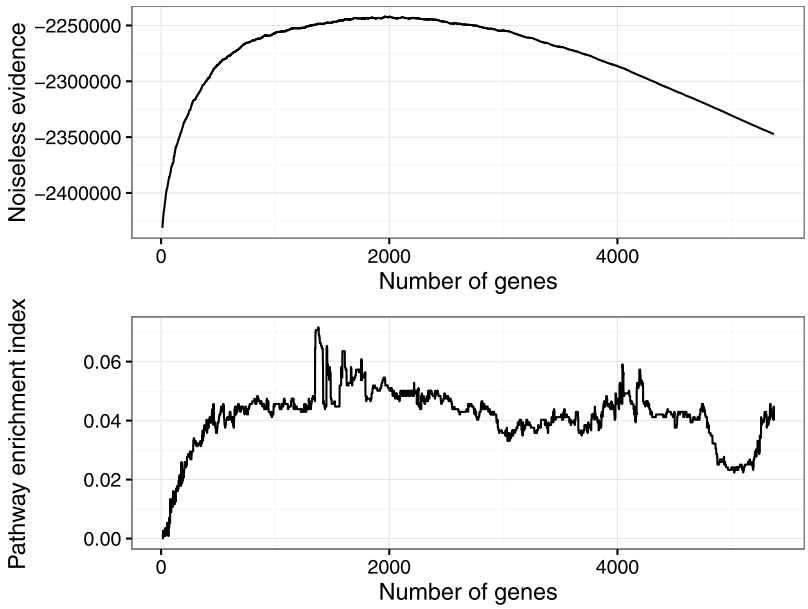


FIG 5. *Marginal likelihood and PEI for the gene selection problem.*

sists 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.

The scalability of our approach comes from the linear complexity of the VEM algorithm of the relaxed model. Such continuous relaxations of discrete Bayesian model selection problems have been empirically very successful in the past, particularly for PCA (Bishop, 1999a,b; Archambeau and Bach, 2009), which was an important motivation for our approach. However, the theoretical foundations of these relaxations remain unclear – although some work has been done regarding linear regression (Wipf, Rao and Nagarajan, 2011). Moreover, the fact that the noise term is separated in the exact model, and unified in the relaxed one, might bring about different behaviours for difficult data sets. A theoretical investigation of the links will be the subject of future work.

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). Regarding 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, Celeux and Girard, 2011; Nakajima et al., 2015) and would be an interesting direction.

Acknowledgements

The authors would like to thank Magnus Ulfarsson for providing svnPCA software and Florentin Damiens for helpful discussions on Bessel functions. Part of this work was done while PAM was visiting University College Dublin, funded by the Fondation Sciences Mathématiques de Paris (FSMP).

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

$$\varphi_{y_1\mathbf{We_1}}(\mathbf{u}) = \mathbb{E}[\exp(iy_1{\mathbf{e_1}}^T\mathbf{W}^T\mathbf{u})] \tag{20}$$

$$= \mathbb{E}\left[\exp\left(iy_1\sum_{k=1}^{p} w_{k1}u_k\right)\right], \tag{21}$$

but, since $w_{st} \sim \mathcal{N}(0, \alpha^{-2})$ for all s, t, we will have

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

thus, since $\mathbf{y}$ and $\mathbf{W}$ are independent, the law of $(\alpha/||\mathbf{u}||_2)y_1\sum_{k=1}^{p} w_{k1}u_k$ 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

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

$$= \frac{2}{\pi}\int_{0}^{+\infty} K_0(t)\cos(\alpha^{-1}||\mathbf{u}||_2 t)dt,$$

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

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

which leads to

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

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

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

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

$$p(\mathbf{x}) = \frac{||\mathbf{x}||_2^{1-p/2}}{(2\pi)^{p/2}}\int_0^{+\infty}\frac{u^{p/2}e^{-\sigma^2u^2}}{(1+u^2/\alpha^2)^{d/2}}J_{p/2-1}(u||\mathbf{x}||_2)du, \tag{22}$$

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 $p(\mathbf{x}) = p(\mathbf{x}_{\bar{\mathbf{v}}})p(\mathbf{x}_{\mathbf{v}})$. Applying (22) to $\mathbf{x}_{\mathbf{v}}$ 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

We begin by proving the following lemma, which links the distribution of the product between a Gaussian matrix and a Gaussian vector with the Bessel distribution. This result may be of independent interest. While this paper was under review, we proved a more general result about the distribution of the product of a Gaussian matrix with a Gaussian vector (Mattei, 2017).

Lemma 1. *Let* $\mathbf{A}$ *be a* $q \times d$ *random matrix such that* $a_{ij} \sim \mathcal{N}(0, s^2)$ *with* $s > 0$ *for all* i, j *and let* $\mathbf{b} \sim \mathcal{N}0, \mathbf{I}_d)$. *Then* $\mathbf{A}\mathbf{b}$ *follows a Bessel distribution with parameters* s *and* $(d-q)/2$.

Proof. Using the decomposition arguments from the proof of Theorem 1, the characteristic function of $\mathbf{A}\mathbf{b}$ is, for all $\mathbf{u} \in \mathbb{R}^k$,

$$\varphi_{\mathbf{A}\mathbf{b}}(\mathbf{u}) = \frac{1}{(1+s^2||\mathbf{u}||_2^2)^{d/2}},$$

which is exactly the characteristic function of the symmetric multivariate Bessel distribution Fang, Kotz and Ng (1990, Def. 2.5). □

We can now prove Theorem 2.

Proof. Let us first consider the case where all variables are active and assume that $\mathbf{v} = (1, 1, ..., 1)$. Using Lévy's continuity theorem, $\boldsymbol{\varepsilon_2}$ 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 $\mathbf{Wy}$, which follows a Bessel$(1/\alpha, (d-q)/2)$ distribution according to our lemma.

In the general case when $\mathbf{v}$ is not necessarily equal to $(1, 1, ..., 1)$ we can prove (6) by invoking the independence between $\mathbf{x_v}$ and $\mathbf{x_{\bar{v}}}$, similarly to the proof of Theorem 1. □

Appendix C: Proof of Proposition 1

Proof. Since a sum of concave functions is concave, it is sufficient to prove that the function $g : \alpha \mapsto p(\tilde{\mathbf{x}}|\mathbf{v}, \alpha, \sigma_1)$ is strictly concave. Up to unnecessary additive constants, we have for all $\alpha > 0$,

$$g(\alpha) = d \log \alpha + \log \left((\alpha ||\tilde{\mathbf{x}}_\mathbf{v}||_2)^{\frac{q-d}{2}} K_{\frac{q-d}{2}} (||\tilde{\mathbf{x}}_\mathbf{v}||_2 \alpha) \right).$$

Using standard results about Bessel functions derivatives (Abramowitz and Stegun, 1965, p. 376), it can be shown that

$$g'(u) = \frac{d}{\alpha} - ||\tilde{\mathbf{x}}_\mathbf{v}||_2 h(u),$$

where the h is the ratio

$$h(\alpha) = \frac{K_{\frac{q-d}{2}-1} (||\tilde{\mathbf{x}}_\mathbf{v}||_2 \alpha)}{K_{\frac{q-d}{2}} (||\tilde{\mathbf{x}}_\mathbf{v}||_2 \alpha)}.$$

As proven independently by Lorch (1967) and Hartman and Watson (1974), since $q - d \geq 0$, h is a increasing function on $\mathbb{R}_+^*$. Therefore g' is stricly decreasing and g is strictly concave. □

Appendix D: Proof of Proposition 2

Proof. Variational distribution of the latent vectors. Using a standard result in variational mean-field approximations (Bishop, 2006, Chap. 10), we can write

$$\ln q^*(\mathbf{y}) = \mathbb{E}_{q(\mathbf{W})}[\ln p(\mathbf{X}, \mathbf{Y}, \mathbf{W}|\boldsymbol{\theta})]$$

which leads to the factorization $q^*(\mathbf{y}) = \prod_{i \leq n} q^*(\mathbf{y}_i)$. Then, for each $i \leq n$, we can write, up to unnecessary additive constants,

$$\ln q^*(\mathbf{y}_i) = \mathbb{E}_{q(\mathbf{W})}[\ln p(\mathbf{x}_i, \mathbf{y}_i, \mathbf{W}|\boldsymbol{\theta})] = \mathbb{E}_{q(\mathbf{W})} \left[\frac{-1}{2\sigma^2} ||\mathbf{x}_i - \mathbf{UW}\mathbf{y}_i||_2^2 \right] - \frac{1}{2} ||\mathbf{y}_i||_2^2,$$

thus

$$\ln q^*(\mathbf{y}_i) = \frac{-1}{2\sigma^2}\mathbf{y_i}^T\mathbb{E}_{q(\mathbf{W})}[\mathbf{W}^T\mathbf{U}^2\mathbf{W}]\mathbf{y}_i + \frac{1}{\sigma^2}\mathbf{y}_i^T\mathbb{E}_{q(\mathbf{W})}[\mathbf{W}]^T\mathbf{U}\mathbf{x}_i - \frac{1}{2}||\mathbf{y}_i||_2^2,$$

which leads to the desired form.

Variational distribution of the loading matrix. Similarly, up to unnecessary additive constants,

$$\ln q^*(\mathbf{W}) = \frac{-1}{2\sigma^2}\sum_{i=1}^{n}\mathbb{E}_{q(\mathbf{y}_i)}[||\mathbf{x}_i - \mathbf{U}\mathbf{W}\mathbf{y}_i||_2^2] - \frac{\alpha^2}{2}\sum_{i=1}^{p}||\mathbf{w}_i||_2^2,$$

which leads to

$$\ln q^*(\mathbf{W}) = \sum_{i=1}^{n}\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{p}u_j^2\mathbf{w}_j^T\mathbb{E}_{q(\mathbf{y}_i)}[\mathbf{y}_i\mathbf{y}_i^T]\mathbf{w}_j + \frac{1}{\sigma^2}\sum_{j=1}^{p}x_{i,j}u_j\mathbf{w}_j^T\mathbb{E}_{q(\mathbf{y}_i)}[\mathbf{y}_i]\right) - \frac{\alpha^2}{2}\sum_{i=1}^{p}||\mathbf{w}_i||_2^2,$$

and

$$\ln q^*(\mathbf{W}) = \sum_{i=1}^{p}\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{p}u_j^2\mathbf{w}_j^T\mathbb{E}_{q(\mathbf{y}_i)}[\mathbf{y}_i\mathbf{y}_i^T]\mathbf{w}_j + \frac{1}{\sigma^2}\sum_{j=1}^{p}x_{i,j}u_j\mathbf{w}_j^T\mathbb{E}_{q(\mathbf{y}_i)}[\mathbf{y}_i]\right) - \frac{\alpha^2}{2}\sum_{i=1}^{p}||\mathbf{w}_i||_2^2,$$

leading to the factorization $q^*(\mathbf{W}) = \prod_{j\leq p} q^*(\mathbf{w}_i)$ and to the desired expression. □

Appendix E: Proof of Proposition 3

Proof. By definition, we have

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

therefore

$$\begin{aligned}-\mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta}) = &-np\ln\sigma - \frac{1}{2\sigma^2}\mathrm{Tr}(\mathbf{X}^T\mathbf{X}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\mathbb{E}_q[\mathbf{y}_i\mathbf{W}^T\mathbf{U}^2\mathbf{W}\mathbf{y}_i]\\ &+ \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{U}\mathbf{M}\boldsymbol{\mu}_i + \sum_{k=1}^{p}\left(d\ln\alpha - \frac{\alpha^2}{2}\mathbb{E}_q[\mathbf{w}_k^T\mathbf{w}_k]\right)\\ &- \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}_q[\mathbf{y}_i^T\mathbf{y}_i] + \frac{n}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{k=1}^{p}\ln|\mathbf{S}_k|,\end{aligned}$$

and computing the expectations leads to

$$
\begin{aligned}
-\mathcal{F}_q(\mathbf{X}|\boldsymbol{\theta}) = & -np\ln\sigma + dp\ln\alpha - \frac{1}{2\sigma^2}\mathrm{Tr}(\mathbf{X}^T\mathbf{X}) \\
& - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{k=1}^{p} u_k^2 \mathrm{Tr}[(\boldsymbol{\Sigma} + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T)(\boldsymbol{S}_k + \mathbf{m}_k\mathbf{m}_k^T)] \\
& + \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i^T\mathbf{U}\mathbf{M}\boldsymbol{\mu}_i + \sum_{k=1}^{p} -\frac{\alpha^2}{2}\mathrm{Tr}(\mathbf{S}_k + \mathbf{m}_k\mathbf{m}_k^T) \\
& - \frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}(\boldsymbol{\Sigma} + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T) + \frac{n}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{k=1}^{p}\ln|\mathbf{S}_k|, \qquad (23)
\end{aligned}
$$

which allows us to conclude. □

Appendix F: Details about the breast cancer data set

The microarray data set used in this paper is included in the `breastCancerVDX` R package (Schroeder et al., 2011) and contains the gene expression data published by Wang et al. (2005) and Minn et al. (2007). It contains expression levels of 22283 probes for 344 patients. In order to be able to provide an interpretation of feature selection, we reduced the data from probe-level to gene-level using the following procedure:

- first, the probes with no gene identifier were discarded,
- then, the data was aggregated to gene-level using the `collapseRows` program of Miller et al. (2011) with default settings (specifically, this means choosing the probe with the largest mean value for each gene),
- among the genes obtained, only the genes listed in the Reactome database (Fabregat et al., 2016) were kept in order to eventually perform pathway enrichment,
- finally, the data was centered but not standardized.

The resulting data matrix contains 5391 variables (genes) and 344 observations (patients).

References

Abramowitz, M. and Stegun, I. (1965). *Handbook of Mathematical Functions*. Dover Publications. MR0208797

Aminghafari, M., Cheze, N. and Poggi, J. M. (2006). Multivariate denoising using wavelets and principal component analysis. *Computational Statistics & Data Analysis* **50** 2381–2398. MR2225575

Amos, D. E. (1986). Algorithm 644: A portable package for Bessel functions of a complex argument and nonnegative order. *ACM Transactions on Mathematical Software* **12** 265–273. MR0889069

Ando, T. (2009). Bayesian factor analysis with fat-tailed factors and its exact marginal likelihood. *Journal of Multivariate Analysis* **100** 1717–1726. MR2535382

Archambeau, C. and Bach, F. (2009). Sparse probabilistic projections. In *Advances in Neural Information Processing Systems* 73–80.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57** 289–300. MR1325392

Biernacki, C., Celeux, G. and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis* **41** 561–575. MR1968069

Bishop, C. M. (1999a). Bayesian PCA. In *Advances in Neural Information Processing Systems* 382–388.

Bishop, C. M. (1999b). Variational Principal Components. In *Proceedings of the Ninth International Conference on Artificial Neural Networks* 509–514.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. MR2247587

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

Bro, R. and Smilde, A. K. (2003). Centering and scaling in component analysis. *Journal of Chemometrics* **17** 16–33.

Brusco, M. J. (2014). A comparison of simulated annealing algorithms for variable selection in principal component analysis and discriminant analysis. *Computational Statistics & Data Analysis* **77** 38–53. MR3210047

Celeux, G., El Anbari, M., Marin, J. M. and Robert, C. P. (2012). Regularization in regression: comparing Bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis* **7** 477–502. MR2934959

Chan, T. H., Jia, K., Gao, S., Lu, J., Zeng, Z. and Ma, Y. (2015). PCANet: A Simple Deep Learning Baseline for Image Classification? *IEEE Transactions on Image Processing* **24** 5017–5032. MR3406099

d'Aspremont, A., Bach, F. and El Ghaoui, L. (2008). Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research* **9** 1269–1294. MR2426043

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., Matthews, L., May, B., Milacic, M., Rothfels, K., Shamovsky, V., Webber, M., Weiser, J., Williams, M., Wu, G., Stein, L., Hermjakob, H. and D'Eustachio, P. (2016). The Reactome pathway Knowledgebase. *Nucleic Acids Research* **44** D481–D487.

Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric multivariate and related distributions*. Chapman and Hall. MR1071174

Gramfort, A., Strohmeier, D., Haueisen, J., Hämäläinen, M. S. and Kowalski, M. (2013). Time-frequency mixed-norm estimates: Sparse

M/EEG imaging with non-stationary source activations. *NeuroImage* **70** 410–422.

Gu, Q., Li, Z. and Han, J. (2011). Joint feature selection and subspace learning. In *Proceedings of the International Joint Conference on Artificial Intelligence* **22** 1294-1299.

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

Hartman, P. and Watson, G. S. (1974). "Normal" distribution functions on spheres and the modified Bessel functions. *The Annals of Probability* 593–607. MR0370687

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press. MR3616141

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24** 417.

Ilin, A. and Raiko, T. (2010). Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research* **11** 1957–2000. MR2678019

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

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104**. MR2751448

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I: Artificial data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **21** 160–173. MR0311034

Jolliffe, I. T. (1973). Discarding variables in a principal component analysis. II: Real data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **22** 21–31. MR0311034

Journée, M. (2009). Geometric algorithms for component analysis with a view to gene expression data analysis. PhD thesis, Université de Liège.

Journée, M., Nesterov, Y., Richtárik, P. and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research* **11** 517–553. MR2600619

Kass, R. E. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association* **84** 717–726. MR1132587

Khan, Z., Shafait, F. and Mian, A. (2015). Joint Group Sparse PCA for Compressed Hyperspectral Imaging. *IEEE Transactions on Image Processing* **24** 4934–4942. MR3402484

Khanna, R., Ghosh, J., Poldrack, R. and Koyejo, O. (2015). Sparse Submodular Probabilistic PCA. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics* 453–461.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J. and Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many

factors of variation. In *Proceedings of the 24th international conference on Machine learning* 473–480. ACM.

Latouche, P., Mattei, P. A., Bouveyron, C. and Chiquet, J. (2016). Combining a Relaxed EM Algorithm with Occam's Razor for Bayesian Variable Selection in High-Dimensional Regression. *Journal of Multivariate Analysis* **146** 177–190. MR3477658

Lawley, D. N. (1953). A modified method of estimation in factor analysis and some large sample results. *Proceedings of the Uppsala Symposium on Psychological Factor Analysis, Uppsala, Sweden* 35–42. MR0062392

Lázaro-Gredilla, M. and Titsias, M. K. (2011). Spike and slab variational inference for multi-task and multiple kernel learning. In *Advances in Neural Information Processing Systems* 2339–2347.

Liu, T. Y., Trinchera, L., Tenenhaus, A., Wei, D. and Hero, A. O. (2013). Globally sparse PLS regression. In *New Perspectives in Partial Least Squares and Related Methods* 117–127. Springer. MR3122897

Lorch, L. (1967). Inequalities for some Whittaker functions. *Archivum Mathematicum* **3** 1–9. MR0223611

MacKay, D. J. C. (1994). Bayesian methods for backpropagation networks. In *Models of Neural Networks III* 211–254. Springer.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press. MR2012999

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89** 1535–1546.

Mächler, M. (2013). `Bessel`: Bessel – Bessel Functions Computations and Approximations R package version 0.5-5.

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

Mattei, P. A. (2017). Multiplying a Gaussian matrix by a Gaussian vector. *Statistics & Probability Letters* **128** 67–70. MR3656377

Mattei, P. A., Bouveyron, C. and Latouche, P. (2016). Globally Sparse Probabilistic PCA. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* 976–984.

McLachlan, G. and Krishnan, T. (2008). *The EM Algorithm and Extensions. Second Edition.* John Wiley & Sons, New York. MR2392878

Miller, J. A., Cai, C., Langfelder, P., Geschwind, D. H., Kurian, S. M., Salomon, D. R. and Horvath, S. (2011). Strategies for aggregating gene expression data: the collapseRows R function. *BMC bioinformatics* **12** 1.

Minka, T. P. (2000). Automatic choice of dimensionality for PCA. In *Advances in Neural Information Processing Systems* 598–604.

Minn, A. J., Gupta, G. P., Padua, D., Bos, P., Nguyen, D. X., Nuyten, D., Kreike, B., Zhang, Y., Wang, Y., Ishwaran, H., Foekens, J. A., van de Vijver, M. and Massagué, J. (2007). Lung metastasis genes couple breast tumor size and metastatic spread. *Proceed-*

ings of the National Academy of Sciences **104** 6740–6745.

Mitchell, T. and Beauchamp, J. (1988). Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association* **83** 1023-1036. MR0997578

Moghaddam, B., Weiss, Y. and Avidan, S. (2005). Spectral bounds for sparse PCA: Exact and greedy algorithms. In *Advances in Neural Information Processing Systems* 915–922.

Mohamed, S., Heller, K. and Ghahramani, Z. (2012). Bayesian and L1 approaches for sparse unsupervised learning. In *Proceedings of the 29th International Conference on Machine Learning* 751–758.

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

Nakajima, S., Tomioka, R., Sugiyama, M. and Babacan, S. D. (2015). Condition for Perfect Dimensionality Recovery by Variational Bayesian PCA. *Journal of Machine Learning Research* **16** 3757-3811. MR3450552

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Ogata, H. (2005). A numerical integration formula based on the Bessel functions. *Publications of the Research Institute for Mathematical Sciences* **41** 949–970. MR2198133

Passemier, D., Li, Z. and Yao, J. (2017). On estimation of the noise variance in high dimensional probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 51–67. MR3597964

Qiu, Y. and Mei, J. (2016). RSpectra: Solvers for Large Scale Eigenvalue and SVD Problems R package version 0.12-0.

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

Rivals, I., Personnaz, L., Taing, L. and Potier, M. C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **23** 401–407.

Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **25** 257–265. MR0440801

Roweis, S. (1998). EM algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems* 626–632.

Schaback, R. and Wu, Z. (1996). Operators on radial functions. *Journal of Computational and Applied Mathematics* **73** 257–270. MR1424880

Schroeder, M., Haibe-Kains, B., Culhane, A., Sotiriou, C., Bontempi, G. and Quackenbush, J. (2011). breastCancerVDX: Gene expression datasets published by Wang et al. [2005] and Minn et al. [2007] (VDX) R package version 1.8.0.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464. MR0468014

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via

regularized low rank matrix approximation. *Journal of Multivariate Analysis* **99** 1015–1034. MR2419336

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

Sobczyk, P., Bogdan, M. and Josse, J. (2017). Bayesian dimensionality reduction with PCA using penalized semi-integrated likelihood. *Journal of Computational and Graphical Statistics* **26** 826–839. MR3765347

Teschendorff, A. E., Journée, M., Absil, P. A., Sepulchre, R. and Caldas, C. (2007). Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Computational Biology* **3** e161. MR2369371

Theobald, C. M. (1975). An Inequality with Application to Multivariate Analysis. *Biometrika* **62** 461-466. MR0391391

Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research* **1** 211–244. MR1875838

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61** 611–622. MR1707864

Ulfarsson, M. O. and Solo, V. (2008). Sparse variable PCA using geodesic steepest descent. *IEEE Transactions on Signal Processing* **56** 5823–5832. MR2518261

Ulfarsson, M. O. and Solo, V. (2011). Vector l0 sparse variable PCA. *IEEE Transactions on Signal Processing* **59** 1949–1958. MR2816474

Van der Vaart, A. W. (2000). *Asymptotic statistics* **3**. Cambridge University Press. MR1652247

Vu, V. Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics* **41** 2905–2947. MR3161452

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M. J. J., Atkins, D. and Foekens, J. A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet* **365** 671–679.

Wipf, D. and Nagarajan, S. (2008). A new view of automatic relevance determination. In *Advances in Neural Information Processing Systems* 1625–1632.

Wipf, D. and Nagarajan, S. (2009). A unified Bayesian framework for MEG/EEG source imaging. *NeuroImage* **44** 947–966.

Wipf, D. P., Rao, B. D. and Nagarajan, S. (2011). Latent variable Bayesian models for promoting sparsity. *IEEE Transactions on Information Theory* **57** 6236–6255. MR2857970

Wishart, J. and Bartlett, M. S. (1932). The distribution of second order moment statistics in a normal system. *Mathematical Proceedings of the Cambridge Philosophical Society* **28**.

Xiaoshuang, S., Zhihui, L., Zhenhua, G., Minghua, W., Cairong, Z. and Heng, K. (2013). Sparse Principal Component Analysis via Joint $L_{2,1}$-Norm

Penalty. In *AI 2013: Advances in Artificial Intelligence* 148–159. Springer. MR3160924

Xu, L. and Jordan, M. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation* **8** 129–151.

Yu, G. and He, Q. Y. (2016). ReactomePA: an R/Bioconductor package for Reactome pathway analysis and visualization. *Molecular BioSystems*.

Yu, L., Snapp, R. R., Ruiz, T. and Radermacher, M. (2010). Probabilistic principal component analysis with expectation maximization (PPCA-EM) facilitates volume classification and estimates the missing data. *Journal of Structural Biology* **171** 18–30.

Zhang, Y., d'Aspremont, A. and El Ghaoui, L. (2012). Sparse PCA: Convex relaxations, algorithms and applications. In *Handbook on Semidefinite, Conic and Polynomial Optimization* 915–940. Springer. MR2894674

Zhang, Y. and El Ghaoui, L. (2011). Large-scale sparse principal component analysis with application to text data. In *Advances in Neural Information Processing Systems* 532–539.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics* **15** 265–286. MR2252527