Bayesian nonparametric Principal Component Analysis

Clément Elvira, Pierre Chainais, Nicolas Dobigeon

To cite this version:

Clément Elvira, Pierre Chainais, Nicolas Dobigeon. Bayesian nonparametric Principal Component Analysis. 2017. hal-01687236

HAL Id: hal-01687236
https://hal.science/hal-01687236

Preprint submitted on 23 Jan 2018

Bayesian nonparametric Principal Component Analysis

Clément Elvira · Pierre Chainais · Nicolas Dobigeon

Abstract Principal component analysis (PCA) is very popular to perform dimension reduction. The selection of the number of significant components is essential but often based on some practical heuristics depending on the application. Only few works have proposed a probabilistic approach able to infer the number of significant components. To this purpose, this paper introduces a Bayesian nonparametric principal component analysis (BNP-PCA). The proposed model projects observations onto a random orthogonal basis which is assigned a prior distribution defined on the Stiefel manifold. The prior on factor scores involves an Indian buffet process to model the uncertainty related to the number of components. The parameters of interest as well as the nuisance parameters are finally inferred within a fully Bayesian framework via Monte Carlo sampling. A study of the (in-)consistence of the marginal maximum a posteriori estimator of the latent dimension is carried out. A new estimator of the subspace dimension is proposed. Moreover, for sake of statistical significance, a Kolmogorov-Smirnov test based on the posterior distribution of the principal components is used to refine this estimate. The behaviour of the algorithm is first studied on various synthetic examples. Finally, the proposed BNP dimension reduction approach is shown to be easily yet efficiently coupled with clustering or latent factor models within a unique framework.

Keywords

Bayesian nonparametrics, dimension reduction, distribution on the Stiefel manifold, Indian buffet process.

C. Elvira · P. Chainais
Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France
E-mail: {clement.elvira, pierre.chainais}@centralelille.fr

N. Dobigeon
University of Toulouse, IRIT/INP-ENSEEIHT, CNRS, 2 rue Charles Camichel, BP 7122, 31071 Toulouse cedex 7, France
E-mail: nicolas.dobigeon@enseeiht.fr

1 Introduction

Dimension reduction (DR) is an ubiquitous preprocessing step in signal processing and statistical data analysis. It aims at finding a lower dimensional subspace explaining a set of data while minimizing the resulting loss of information. Related interests are numerous, e.g., reducing the impact of noise, data storage, computational time.

Principal component analysis (PCA) permits DR by projecting observations onto a subset of orthonormal vectors. It provides an elegant solution to DR by looking for a K-dimensional representation of a dataset $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ with $\mathbf{y}_n \in \mathbb{R}^D$ in an orthonormal basis, referred to as principal components. Given K, the K-dimensional subspace spanned by these principal components is supposed to minimize the quadratic reconstruction error of the dataset, see (Jolliffe 1986) for a comprehensive review of PCA. According to one of its standard formulations, PCA can be interpreted as the search of an orthonormal basis $\mathbf{P}$ of $\mathbb{R}^D$ such that all matrices formed by the first K columns of $\mathbf{P}$ and denoted $\mathbf{P}_{:,1:K}$ ensures

$$\forall K \in \{1, \ldots, D\}, \quad \mathbf{P}_{:,1:K} = \underset{\mathbf{U} \in \mathcal{S}_D^K}{\operatorname{argmax}} \, \mathbf{U}^T \mathbf{Y} \mathbf{Y}^T \mathbf{U} \tag{1}$$

where $\mathcal{S}_D^K$ is the Stiefel manifold, i.e., the set of $D \times K$ orthonormal matrices.

However, Eq. (1) does not provide tools to assert the relevance of the selected principal components in expectation over the data distribution. To fill this gap, Tipping and Bishop (1999b) have shown that PCA can be interpreted as a maximum likelihood estimator of latent factors following the linear model

$$\forall n \in \{1, \ldots, N\}, \quad \mathbf{y}_n = \mathbf{W}\mathbf{x}_n + \boldsymbol{\varepsilon}_n \tag{2}$$

where $\mathbf{y}_n$ is the observation vector, $\mathbf{W}$ is the matrix of latent factors assumed to be Gaussian, $\mathbf{x}_n$ is the associated vector of coefficients and ε_n is an isotropic Gaussian noise. If the coefficients $\mathbf{x}_n$ are assumed Gaussian, they can be analytically marginalized out thanks to a natural conjugacy property. The resulting marginalized likelihood function $\mathrm{p}(\mathbf{y}|\mathbf{W}, \boldsymbol{\varepsilon}_n)$ can be expressed in terms of the empirical covariance matrix $\mathbf{Y}^T\mathbf{Y}$ and the hermitian matrix $\mathbf{W}^T\mathbf{W}$. Although no orthogonality constraint is imposed on the latent factors, the resulting marginal maximum likelihood estimator is precisely provided by the singular value decomposition (SVD) of the noise-corrected observation vector: the SVD produces a set of orthogonal vectors. The subspace can then be recovered using an expectation-maximization (EM) algorithm. One of the main advantages of this so-called probabilistic PCA (PPCA) lies in its ability to deal with non-conventional datasets. For instance, such an approach allows PCA to be conducted while facing missing data or non linearities (Tipping and Bishop 1999a,b). Several works have pursued these seminal contributions, e.g., to investigate these non linearities more deeply (Bolton et al 2003; Lawrence 2005; Lian 2009) or the robustness of PPCA with respect to the presence of corrupted data or outliers (Archambeau et al 2008; Schmitt and Vakili 2016).

Several studies have addressed the issue of determining the relevant latent dimension of the data, K here. The PPCA along with its variational approximation proposed by Bishop (1999a,b) automatically prunes directions associated with low variances, in the spirit of automatic relevance determination (MacKay

1995). Another strategy considers the latent dimension K as a random variable within a hierarchical model of the form $f(\mathbf{W}|K)f(K)$ and uses the SVD decomposition of $\mathbf{W}$. However, explicit expressions of the associated estimators are difficult to derive. To bypass this issue, Minka (2000) and Smídl and Quinn (2007) have proposed Laplace and variational approximations of the resulting posteriors, respectively. Solutions approximated by Monte Carlo sampling are even harder to derive since the size of the parameter space varies with K. Zhang et al (2004) have proposed to use reversible jump Markov chain Monte Carlo (RJ-MCMC) algorithms (Green 1995) to build a Markov chain able to explore spaces of varying dimensions. Despite satisfying results, this method is computationally very expensive.

Bayesian nonparametric (BNP) inference has been a growing topic over the past fifteen years, see for instance the review by Müller and Mitra (2013). Capitalizing on these recent advances of the BNP literature, this work proposes to use the Indian buffet process (IBP) as a BNP prior to deal with the considered subspace inference problem. More precisely, the basis of the relevant subspace and associated representation coefficients are incorporated into a single Bayesian framework called Bayesian nonparametric principal component analysis or BNP-PCA. A preliminary version of this work was presented at ICASSP 2017 (Elvira et al 2017). Following the approach by Besson et al (2011), the prior distribution of the principal components is a uniform distribution over the Stiefel manifold. Then, the IBP permits to model the observations by a combination of a potentially infinite number of latent factors. Inheriting from intrinsic properties of BNP, the IBP naturally penalizes the complexity of the model (i.e., the number K of relevant factors), which is a desired behaviour for dimension reduction. In addition, while the IBP still permits to infer subspaces of potentially infinite dimension, the orthogonality constraint imposed to the latent factors enforces their number K to be at most D: orthogonality has some regularization effect as well. The posterior of interest is then sampled using an efficient MCMC algorithm which does not require reversible jumps.

Compared to alternative approaches, in particular those relying on RJ-MCMC sampling, the adopted strategy conveys significant advantages. First, although RJ-MCMC is a powerful and generic tool, its implementation needs the definition of bijections between parameter spaces of different sizes. As a consequence, Jacobian matrices contribute to the probability of jumping between spaces of different dimensions. These Jacobian terms are often both analytically and computationally expensive. Within a BNP framework, there are no such Jacobian terms. Monte Carlo sampling of BNP models implicitly realizes trans-dimensional moves since the IBP prior is a distribution on infinite binary matrices. Combined to the conjugacy properties of the IBP, such a formulation permits more efficient Monte Carlo sampling. Then, the use of the IBP and its induced sparsity alleviates the overestimation of the latent dimension coupled with a subsequent pruning strategy followed by other crude approaches. The proposed model also opens the door to a theoretical analysis of the consistency of estimators. Finally, the method is flexible enough to be coupled with standard machine learning (e.g., classification) and signal processing (e.g., signal decomposition) tasks.

This paper is organized as follows. Section 2 recalls notions on directional statistics and the IBP. Section 3 describes the proposed hierarchical Bayesian model for BNP-PCA. Section 4 describes the MCMC inference scheme. Section 5

Symbol	Description
N, n	number of observations, with index
D, d	dimension of observations with index
K, k	number of latent factors, with index
$\mathcal{P}(\alpha)$	Poisson distribution with parameter α
$\mathcal{S}_D^K$	set of $D \times K$ matrices $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{P} = \mathbb{I}_K$
$\mathcal{O}_D$	The orthogonal group
etr	exp tr
${}_i\mathrm{F}_j$	Confluent hypergeometric function
$\gamma(a, b)$	Lower incomplete Gamma function
$\langle \cdot, \cdot \rangle$	Euclidean scalar product

Table 1 List of symbols

defines several estimators and gathers theoretical results on their properties, in particular their (in-)consistency. Section 6 illustrates the performance of the proposed method on numerical examples. Concluding remarks are finally reported in Section 8. Note that all notations are gathered in Table 1.

2 Preliminaries

2.1 Distribution on the Stiefel Manifold

The set of $D \times K$ real matrices $\mathbf{P}$ which verify the relation $\mathbf{P}^T\mathbf{P} = \mathbb{I}_K$ is called the Stiefel manifold and is denoted $\mathcal{S}_D^K$. Note that when $K = D$, The Stiefel manifold $\mathcal{S}_D^D$ corresponds to the orthogonal group $\mathcal{O}_D$. The Stiefel manifold is compact with finite volume

$$\mathrm{vol}\left(\mathcal{S}_D^K\right) = \frac{2^K \pi^{\frac{DK}{2}}}{\pi^{\frac{1}{4}K(K-1)} \prod_{i=1}^{D} \Gamma\left(\frac{D}{2} - \frac{i-1}{2}\right)}. \tag{3}$$

Hence, the uniform distribution $\mathcal{U}_{\mathcal{S}_D^K}$ on the Stiefel manifold is defined by the density with respect to the Lebesgue measure given by

$$\mathrm{p_U}\left(\mathbf{P}\right) = \frac{1}{\mathrm{vol}(\mathcal{S}_D^K)} \mathbb{1}_{\mathcal{S}_D^K}(\mathbf{P}). \tag{4}$$

Over the numerous distributions defined on the Stiefel manifold, two of them play a key role in the proposed Bayesian model, namely the *matrix von Mises-Fisher* and the *matrix Bingham* distributions. Their densities with respect to the Haar measure on the Stiefel Manifold have the following form

$$\mathrm{p_{vMF}}\left(\mathbf{P}|\mathbf{C}\right) = {}_0\mathrm{F}_1^{-1}\left(\emptyset, \frac{D}{2}, \mathbf{C}^T\mathbf{C}\right) \mathrm{etr}\left(\mathbf{C}^T\mathbf{P}\right) \tag{5}$$

$$\mathrm{p_B}\left(\mathbf{P}|\mathbf{B}\right) = {}_1\mathrm{F}_1^{-1}\left(\frac{D}{2}, \frac{K}{2}, \mathbf{B}\right) \mathrm{etr}\left(\mathbf{P}^T\mathbf{B}\mathbf{P}\right) \tag{6}$$

where $\mathbf{C}$ is a $D \times K$ matrix, $\mathbf{B}$ is a $D \times D$ symmetric matrix and $\mathrm{etr}(\cdot)$ stands for the exponential of the trace of the corresponding matrix. The two special functions ${}_0\mathrm{F}_1$ and ${}_0\mathrm{F}_0$ are two confluent hypergeometric functions of matrix arguments (Herz 1955).

2.2 Nonparametric sparse promoting prior

The Indian buffet process (IBP), introduced by Griffiths and Ghahramani (2011), defines a distribution over binary matrices with a fixed number N of columns but a potentially infinite number of rows denoted by K. The IBP can be understood with the following culinary metaphor. Let consider a buffet with an infinite number of available dishes. The first customer chooses $K_1 \sim \mathcal{P}(\alpha)$ dishes. The nth customer selects the kth dish among those already selected with probability $\frac{m_k}{n}$ (where m_k is the number of times dish k has been previously chosen) and tries $K_n \sim \mathcal{P}(\frac{\alpha}{n})$ new dishes. Let $\mathbf{Z}$ the binary matrix defined by $z_{k,n} = 1$ if the nth customer has chosen the kth dish, and zero otherwise. The probability of any realization of $\mathbf{Z}$ is called the *exchangeable feature probability function* by Broderick et al (2013) and is given by

$$\mathrm{P}\left[\mathbf{Z}|\alpha\right] = \frac{\alpha^K e^{-\alpha \sum_n \frac{1}{n}}}{\prod_{i=1}^{2^N-1} K_i!} \prod_{k=1}^{K} \frac{(N-m_k)!(m_k-1)!}{N!} \tag{7}$$

where K_i denotes the number of times a *history* has appeared: the term *history* refers to a realization of the binary vector of size N formed by the rows $(z_{k,\cdot})$ of $\mathbf{Z}$. Thus, there are $2^N - 1$ possibilities. The IBP can also be interpreted as the asymptotic distribution of a beta Bernoulli process where the beta process has been marginalized out (Thibaux et al 2007). A stick-breaking construction has been also proposed by Teh et al (2007). We emphasize that the IBP of parameter α is a α-sparsity promoting prior since the expected number of non-zero coefficient in $\mathbf{Z}$ is of order $\alpha N \log N$.

3 Bayesian nonparametric principal component analysis (BNP-PCA)

This section introduces a Bayesian method called BNP-PCA for dimension reduction that includes the a priori unknown number of underlying components into the model. The latent factor model and the associated likelihood function are first introduced in Section 3.1. The prior model is described in Section 3.2. A Monte Carlo-based inference scheme will be proposed in Section 4.

3.1 Proposed latent factor model

Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ denote the $D \times N$-matrix of observation vectors $\mathbf{y}_n = [y_{1,n}, \ldots, y_{D,n}]^T$. For sake of simplicity but without loss of generality, the sample mean vector $\bar{\mathbf{y}} \triangleq \frac{1}{N}\sum_{n=1}^{N} \mathbf{y}_n$ is assumed to be zero. Data are supposed to live in an unknown subspace of dimension $K \leq D$. The problem addressed here is thus to identify both the latent subspace and its dimension. To this aim, the observation vectors are assumed to be represented according to the following latent factor model

$$\forall n \in \{1, \ldots, N\}, \quad \mathbf{y}_n = \mathbf{P}(\mathbf{z}_n \odot \mathbf{x}_n) + \mathbf{e}_n \tag{8}$$

where $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_D]$ is an orthonormal base of $\mathbb{R}^D$, i.e., $\mathbf{P}^T\mathbf{P} = \mathbb{I}_D$, $\mathbf{z}_n = [z_{1,n}, \ldots, z_{D,n}]^T$ is a binary vector, $\mathbf{x}_n = [x_{1,n}, \ldots, x_{D,n}]^T$ is a vector of coefficients and $\odot$ denotes the Hadamard (term-wise) product. In Eq. (8), the additive term $\mathbf{e}_n$ can stand for a measurement noise or a modeling error and is assumed to be white and Gaussian with variance σ^2. It is worth noting that the binary variable $z_{k,n}$ ($k \in \{1, \ldots, D\}$) explicitly encodes the activation hence the relevance of the coefficient $x_{k,n}$ and of the corresponding direction $\mathbf{p}_k$ for the latent representation. Thus, the term-wise product vectors $\mathbf{s}_n \triangleq \mathbf{z}_n \odot \mathbf{x}_n$ would be referred to as *factor scores* in the PCA terminology. This is the reason why we call this approach Bayesian nonparametric principal component analysis or BNP-PCA.

The likelihood function is obtained by exploiting the Gaussian property of the additive white noise term. The likelihood of the set of N observed vectors assumed to be a priori independent can be written as

$$f(\mathbf{Y}|\mathbf{P}, \mathbf{Z}, \mathbf{X}, \sigma^2) = (2\pi\sigma^2)^{-DN/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^{N} \|\mathbf{y}_n - \mathbf{P}(\mathbf{z}_n \odot \mathbf{x}_n)\|_2^2 \right), \tag{9}$$

where $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]$ is the binary activation matrix, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ is the matrix of representation coefficients and $\| \cdot \|_2$ stands for the ℓ_2-norm.

3.2 Prior distributions

The unknown parameters associated with the likelihood function are the orthonormal basis $\mathbf{P}$, the binary matrix $\mathbf{Z}$, the coefficients $\mathbf{X}$ and the noise variance σ^2. Let define the corresponding set of parameters as $\boldsymbol{\theta} = (\mathbf{P}, \mathbf{Z}, \sigma^2)$, leaving $\mathbf{X}$ apart for future marginalization.

Orthonormal basis P. By definition, $\mathbf{P}$ is an orthonormal basis and belongs to the orthogonal group $\mathcal{O}_D$. Since no information is available a priori about any preferred direction, a uniform distribution on $\mathcal{O}_D$ is chosen as a prior distribution on $\mathbf{P}$ whose probability density function (pdf) with respect to the Lebesgue measure is given by Eq. (4).

Indian buffet process Z. Since the observation vectors are assumed to live in a lower dimensional subspace, most of the factor scores in the vectors $\mathbf{z}_n \odot \mathbf{x}_n$ are expected to be zero. To reflect this key feature, an IBP prior IBP(α) is assigned to the binary latent factor activation coefficients, as discussed in Section 2.2. The parameter α controls the underlying sparsity of $\mathbf{Z}$. Note that the IBP is a prior over binary matrices with a potentially infinite number of rows K. However any factor model underlied by a matrix $\mathbf{Z}$ with $K > D$ will occur with null probability due to to the orthogonality of $\mathbf{P}$. Our purpose is to combine the flexibility of the IBP prior with the search for an orthogonal projector.

Coefficients X. Independent Gaussian prior distributions are assigned to the individual representation coefficients gathered in the matrix $\mathbf{X}$. This choice can be easily motivated for large N by the central limit theorem since these coefficients are expected to result from orthogonal projections of the observed vectors onto the identified basis. Moreover, it has the great advantage of being conjugate to make later marginalization tractable analytically (see next section). To reflect the fact that the relevance of a given direction $\mathbf{p}_k$ is assessed by the ratio between the energy of the corresponding representation coefficients in x_k and the noise variance σ^2, we follow the recommendation of Punskaya et al (2002) to define the prior variances of these coefficients as multiples of the noise variance through a Zellner's prior

$$\forall k \in \mathbb{N}, \quad \mathbf{x}_k | \delta_k^2, \sigma^2 \sim \prod_{n=1}^{N} \mathcal{N}(0, \delta_k^2 \sigma^2). \tag{10}$$

Along this interpretation, the hyperparameters δ_k^2 would correspond to the ratios between the eigenvalues of a classical PCA and the noise variance.

Noise variance σ^2. A non informative Jeffreys' prior is assigned to σ^2

$$f(\sigma^2) \propto \frac{1}{\sigma^2} \mathbb{1}_{\mathbb{R}_+}\left(\sigma^2\right). \tag{11}$$

Hyperparameters. The set of hyperparameters is gathered in $\boldsymbol{\phi} = \{\boldsymbol{\delta}, \alpha\}$ with $\boldsymbol{\delta} = \left\{\delta_1^2, \ldots, \delta_K^2\right\}$. The IBP parameter α will control the mean number of active latent factors while each hyperparameter δ_k^2 scales the power of each component $\mathbf{p}_k$ with respect to the noise variance σ^2. In this work, we propose to include them into the Bayesian model and to jointly estimate them with the parameters of interest. This hierarchical Bayesian approach requires to define priors for these hyperparameters (usually referred to as hyperpriors), which are summarized below.

Scale parameters δ_k^2. The powers of relevant components are expected to be at least of the order of magnitude of the noise variance. Thus, the scale parameters δ_k^2 are assumed to be a priori independent and identically distributed according to a conjugate shifted inverse gamma (sIG, see Appendix B for more details) distribution defined over $\mathbb{R}_+$ as in (Godsill 2010)

$$\begin{aligned} \mathrm{p}_{\mathrm{sIG}}\left(\delta_k^2 | a_\delta, b_\delta\right) = & \frac{b_\delta^{a_\delta}}{\gamma\left(a_\delta, b_\delta\right)} \\ & \left(\frac{1}{1+\delta_k^2}\right)^{a_\delta+1} \exp\left(-\frac{b_\delta}{1+\delta_k^2}\right) \mathbb{1}_{\mathbb{R}_+}\left(\delta_k^2\right) \end{aligned} \tag{12}$$

where $\gamma(a, b)$ is the lower incomplete gamma function and a_δ and b_δ are positive hyperparameters chosen to design a vague prior, typically $a = 1$ and $b = 0.1$. Note that the specific choice $a_\delta = b_\delta = 0$ would lead to a noninformative Jeffreys prior (Punskaya et al 2002). However, this choice is prohibited here since it would also lead to an improper posterior distribution (Robert 2007).

IBP parameter α. Without any prior knowledge regarding this hyperparameter, a Jeffreys prior is assigned to α. As shown in Appendix C, the corresponding pdf is given by

$$f(\alpha) \propto \frac{1}{\alpha} \mathbb{1}_{\mathbb{R}_+}(\alpha). \tag{13}$$

4 Inference: MCMC algorithms

The posterior distribution resulting from the hierarchical Bayesian model for BNP-PCA described in Section 3 is too complex to derive closed-form expressions of the Bayesian estimators associated with the parameters of interest, namely, the orthonormal matrix $\mathbf{P}$ and the binary matrix $\mathbf{Z}$ selecting the relevant components.

To overcome this issue, this section introduces a MCMC algorithm to generate samples asymptotically distributed according to the posterior distribution of interest. It also describes a practical way of using these samples to approximate Bayesian estimators.

4.1 Marginalized posterior distribution

A common tool to reduce the dimension of the space to be explored while resorting to MCMC consists in marginalizing the full posterior distribution with respect to some parameters. In general, the resulting collapsed sampler exhibits faster convergence and better mixing properties (D. A. van Dyk and Park 2008). Here, taking benefit from the conjugacy property induced by the prior in Eq. (10), we propose to marginalize over the coefficients $\mathbf{X}$ according to the following hierarchical model

$$f\left(\boldsymbol{\theta}, \boldsymbol{\phi} | \mathbf{Y}\right) = \int_{\mathbb{R}^{DN}} f\left(\mathbf{Y} | \boldsymbol{\theta}, \mathbf{X}\right) f\left(\boldsymbol{\theta}, \mathbf{X} | \boldsymbol{\phi}\right) f\left(\boldsymbol{\phi}\right) \mathrm{d}\mathbf{X}. \tag{14}$$

Calculations detailed in Appendix A lead to the marginalized posterior distribution

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{\phi} | \mathbf{Y}) \;=\; & \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{ND}{2}} \exp\left(-\frac{\mathrm{tr}(\mathbf{Y}^T\mathbf{Y})}{2\sigma^2}\right) \\ & \times \prod_{k=1}^{K} \exp\left[\frac{1}{2\sigma^2}\frac{\delta_k^2}{1+\delta_k^2}\sum_n z_{k,n}\left\langle \mathbf{p}_k, \mathbf{y}_n \right\rangle^2\right] \\ & \times \prod_{k=1}^{K}\left(\frac{1}{1+\delta_k^2}\right)^{a_\delta + \frac{1}{2}\sum_n z_{k,n}} \exp\left(-\frac{b_\delta}{1+\delta_k^2}\right) \\ & \times \frac{\alpha^K}{\prod_k K_n!} e^{-\alpha \sum_n \frac{1}{i}} \prod_k \frac{(N-m_k)!\,(m_k-1)!}{N!} \\ & \times \left(\frac{b_\delta^{a_\delta}}{\gamma\left(a_\delta, b_\delta\right)}\right)^K \left(\sigma^2\right)^{-1} \alpha^{-1} \mathbb{1}_{\mathbb{U}_D}(\mathbf{P}). \end{aligned} \tag{15}$$

Note that, since the main objective of this work is to recover a lower dimensional subspace (and not necessarily the representation coefficients of observations on this subspace), this marginalization goes beyond a crude sake of algorithmic convenience. It is also worth noting that it is still possible to marginalize with respect to the scale parameters δ_k^2. This finding will be exploited in Section 4.2.

4.2 MCMC algorithm

The proposed MCMC algorithm includes the sampling of $\mathbf{Z}$ described in Algo. 1 and is summarized in Algo. 2. It implements a Gibbs sampling to generate samples asymptotically distributed according to Eq. (15). This section derives the conditional distributions associated with the parameters and hyperparameters.

Sampling the binary matrix Z. The matrix $\mathbf{Z}$ is updated as suggested by Knowles and Ghahramani (2011), see Algo. 1. Let $m_k(n) = \sum_{i \neq n} z_{k,i}$ the number of observations different from n which actually use the direction $\mathbf{p}_k$, *i.e.* verifying

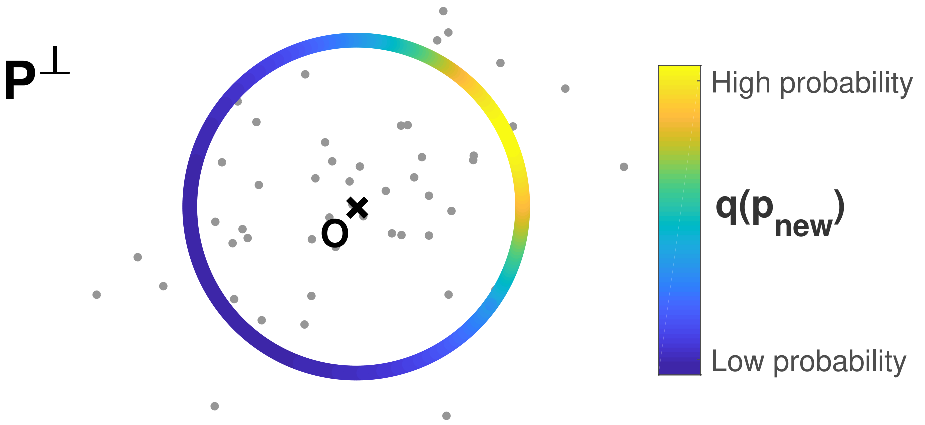

Fig. 1 An example of the proposition of new directions when $\mathbf{P}^{\perp}$ is 2 dimensional and $\kappa^{\star} = 1$. Gray dots are observations projected on $\mathbf{P}^{\perp}$. The colored circle is the pdf of the proposal distribution when $\kappa^{\star} = 1$.

$z_{k,i} = 1$ for $i \neq n$. Directions for which $m_k(n) = 0$ are called *singletons* and the corresponding indices are gathered in a set denoted by $\mathcal{J}_n$. Conversely, directions for which $m_k(n) > 0$ are referred to as *non-singletons* and the set of corresponding indices is denoted by $\mathcal{I}_n$. Note that $\forall n$, $\mathcal{I}_n \cup \mathcal{J}_n = \{1, \dots, K\}$. First, non-singletons are updated through a Gibbs sampling step where δ_k^2 can be marginalized out. One has

$$\begin{aligned}&\frac{\mathrm{P}\left(z_{k,n} = 1 | \mathbf{Y}, \mathbf{P}, \sigma^2\right)}{\mathrm{P}\left(z_{k,n} = 0 | \mathbf{Y}, \mathbf{P}, \sigma^2\right)} = \\ &\frac{m_k(n)}{N - 1 - m_k(n)} \exp\left(\frac{1}{2\sigma^2}\left(\mathbf{p}_k^T \mathbf{y}_n\right)^2\right) \times \\ &\frac{\gamma\left(a+1, b + \frac{1}{2\sigma^2}\left(\mathbf{p}_k^T \mathbf{y}_n\right)^2\right)}{\gamma(a,b)} \frac{b^a}{\left(b + \frac{1}{2\sigma^2}\left(\mathbf{p}_k^T \mathbf{y}_n\right)^2\right)^{a+1}}\end{aligned} \tag{16}$$

where

$$a = a_\delta + \sum_{i=1, i\neq n}^{N} z_{k,i} \tag{17}$$

$$b = b_\delta + \frac{1}{2\sigma^2} \sum_{i=1, i\neq n}^{N} z_{k,i} \left(\mathbf{p}_k^T \mathbf{y}_i\right)^2. \tag{18}$$

A Metropolis Hastings step is used to update singletons. Let $\kappa = \mathrm{card}\left(\mathcal{J}_n\right)$ be the number of singletons, $\mathbf{P}_{\mathcal{I}_n}$ and $\mathbf{P}_{\mathcal{J}_n} \triangleq [\tilde{\mathbf{p}}_1, \dots \tilde{\mathbf{p}}_\kappa]$ be the sub-matrices of $\mathbf{P}$ with indices in $\mathcal{I}_n$ and $\mathcal{J}_n$, respectively. The move goes from a current state $\mathbf{s} = \{\kappa, \mathbf{P}_{\mathcal{J}_n}\}$ to a new state $\mathbf{s}^\star = \left\{\kappa^\star, \mathbf{P}^\star_{\mathcal{J}_n^\star}\right\}$. The proposal distribution in the Metropolis-Hastings step is chosen according to the conditional model

$$\mathrm{q}\left(\kappa^\star, \mathbf{P}^\star_{\mathcal{J}_n^\star} | \kappa, \mathbf{P}_{\mathcal{J}_n}, \mathbf{P}_{\mathcal{I}_n}\right) = \mathrm{q}\left(\kappa^\star | \mathbf{P}_{\mathcal{I}_n}\right) \mathrm{q}\left(\mathbf{P}^\star_{\mathcal{J}_n^\star} | \kappa^\star, \mathbf{P}_{\mathcal{I}_n}\right). \tag{19}$$

Algorithm 1: Detailed procedure to sample $\mathbf{Z}$

```
Input: $\mathbf{Y}, \mathbf{Z}^{(t-1)}, \mathbf{P}^{(t-1)}, \sigma^{2\,(t-1)}, \delta_k^{2\,(t-1)}$
Let $\mathbf{P}^{(t-\frac{1}{2})} = \mathbf{P}^{(t-1)}$ ;
for $n \leftarrow 1$ to $N$ do
    // Identify shared directions and singletons
    for $k \leftarrow 1$ to $K$ do
        Compute $m_k(n) = \sum_{l \neq n} z_{k,l}^{(t-1)}$;
    end
    Let $\mathcal{I}_n \triangleq \{k,\ m_k(n) > 0\}$ ;
    Let $\mathcal{J}_n \triangleq \{k,\ m_k(n) = 0\}$ ;
    // Sample shared directions
    foreach $k$ in $\mathcal{I}$ do
        Sample $z_{k,n}^{(t)}$ according to Eq. (16) ;
    end
    // Define set of singletons
    Let $\kappa \triangleq \mathrm{card}(\mathcal{J}_n)$ ;
    Let $\mathbf{P}_{\mathcal{I}_n} \triangleq [\mathbf{p}_k, k \in \mathcal{I}_n]$. ;
    // Sample new number of singletons
    Sample $\kappa^\star$ according to Eq. (20) ;
    // Sample iteratively new directions
    Let $\mathbf{P}^\star_{\mathcal{J}_n^\star} = [\,]$ ;
    for $k \leftarrow 1$ to $\kappa^\star$ do
        Let $\mathbf{N}$ an orthonormal basis of $[\mathbf{P}_{\mathcal{I}_n}, \mathbf{P}^\star_{\mathcal{J}_n^\star}]^\perp$;
        Let $\mathbf{v}$ the first eigenvector of $\mathbf{N}^T\mathbf{Y}\mathbf{Y}^T\mathbf{N}$ and $\lambda$ its associated eigenvalue ;
        Sample $\mathbf{p}_k^\star \sim \mathrm{vMF}(\mathbf{v}, \lambda)$ ;
        Update $\mathbf{P}^\star_{\mathcal{J}_n^\star} = [\mathbf{P}^\star_{\mathcal{J}_n^\star}, \mathbf{p}_k^\star]$ ;
    end
    // Metropolis Hasting step
    Compute $u_{\mathbf{s}\to\mathbf{s}^\star}$ according to Eq. (21) ;
    Sample $u \sim \mathcal{U}([0,1])$ ;
    if $u \leq u_{\mathbf{s}\to\mathbf{s}^\star}$ then
        Set $\mathbf{s} = \mathbf{s}^\star$ and update $\mathbf{P}^{(t-\frac{1}{2})}$ ;
        Update $K = K - \kappa + \kappa^\star$ ;
    end
end
Output: $\mathbf{Z}^{(t)}, \mathbf{P}^{(t-\frac{1}{2})}$.
```

Note that the proposal distribution Eq.(19) is conditioned to $\mathbf{P}_{\mathcal{I}_n}$. This choice is legit since the goal is to sample the conditional distribution $f(\mathbf{Z}, \mathbf{P}_{\mathcal{J}_n} | \mathbf{Y}, \mathbf{P}_{\mathcal{I}_n}, \sigma^2)$. Close to the structure of the IBP, we propose to use for $\mathrm{q}\left(\kappa^\star | \mathbf{P}_{\mathcal{I}_n}\right)$ a Poisson distribution $\mathcal{P}(\alpha)$ combined with a mass $\mathrm{card}(\mathcal{I}_n)/D$ on 0:

$$\mathrm{q}\left(\kappa^\star | \mathbf{P}_{\mathcal{I}_n}\right) = \frac{\mathrm{card}(\mathcal{I}_n)}{D}\delta_0(\kappa) + \left(1 - \frac{\mathrm{card}(\mathcal{I}_n)}{D}\right)\mathcal{P}(\alpha) \tag{20}$$

Recall that $\mathrm{card}(\mathcal{I}_n)$ is the number of coefficients $z_{k,n} = 1$ of the nth column of $\mathbf{Z}$ that are not singletons (singletons$\Leftrightarrow z_{k,n} = 1$ & $\forall i \neq n,\ z_{k,i} = 0$). Once $\kappa^\star$ has been chosen, a new matrix $\mathbf{Z}^\star$ is formed by concatenating columns with indices in $\mathcal{I}_n$ and $\kappa^\star$ rows with zeros everywhere except ones at the n^{th} position (or column).

For $\mathbf{P}_{\mathcal{J}_n}$, a von Mises-Fisher distribution vMF($\mathbf{C}$), see Section 2.1, is chosen as a proposal. The columns of $\mathbf{C}$ are built from the κ first eigenvectors of the projection of $\mathbf{Y}\mathbf{Y}^T$ on the orthogonal of $\mathbf{P}_{\mathcal{I}_n}$, i.e. the span of singletons and unused directions. The columns of $\mathbf{C}$ are then multiplied by their corresponding eigenvalues. Figure 1 illustrates the procedure to add one new direction, $\kappa^\star = 1$, on a simple example in dimension 2.
The move $\mathbf{s} \to \mathbf{s}^\star$ is then accepted with probability

$$u_{\mathbf{s}\to\mathbf{s}^\star} = \frac{f\left(\mathbf{Y}|\mathbf{P}_{\mathcal{I}_n}, \mathbf{P}^\star_{\mathcal{J}_n}, \mathbf{Z}^\star, \sigma^2\right)}{f\left(\mathbf{Y}|\mathbf{P}_{\mathcal{I}_n}, \mathbf{P}_{\mathcal{J}_n}, \mathbf{Z}, \sigma^2\right)} \frac{\mathrm{p}\left(\mathbf{s}^\star\right) \mathrm{q}(\mathbf{s}|\mathbf{s}^\star, \mathbf{P}_{\mathcal{I}_n})}{\mathrm{p}\left(\mathbf{s}\right) \mathrm{q}(\mathbf{s}^\star|\mathbf{s}, \mathbf{P}_{\mathcal{I}_n})} \tag{21}$$

The full procedure is summarized in Algo. 2.

Algorithm 2: Gibbs sampler

```
Input: $\mathbf{Y}$, $n_{\mathrm{mc}}$
for $t \leftarrow 1$ to $n_{\mathrm{mc}}$ do
    // Update directions and handle singletons
    Sample $\mathbf{Z}^{(t)}$ and $\mathbf{P}^{(t-\frac{1}{2})}$ as described in Alg. 1 ;
    // Update activated directions and weights.
    for $k \leftarrow 1$ to $K$ do
        Compute $\mathbf{N}_{K\backslash k}$, a basis of $\mathbf{P}_{\backslash k}^{\perp\,(t-\frac{1}{2})}$ ;
        Sample $\mathbf{v}_k$ according to Eq. (22) ;
        Set $\mathbf{p}_k^{(t)} = \mathbf{N}_{K\backslash k}\mathbf{v}_k$ ;
        Sample $\delta_k^{2\,(t)}$ according to Eq. (24) ;
    end
    // Update hyperparameters.
    Sample $\sigma^{2\,(t)}$ according to Eq. (25) ;
    Sample $\alpha^{(t)}$ according to Eq. (26) ;
end
Output: A collection of samples $\left\{\mathbf{P}^{(t)}, \mathbf{Z}^{(t)}, \delta_k^{2\,(t)}, \sigma^{2\,(t)}, \alpha^{(t)}\right\}_{t=n_{\mathrm{burn}}+1}^{n_{mc}}$
        asymptotically distributed according to Eq. (15).
```

Sampling the orthonormal basis P.
Let $\mathcal{A} \subset \{1, \ldots, D\}$ denote the set of K indices corresponding to the active directions in $\mathbf{P}$, i.e., the K columns of $\mathbf{P}$ actually used by at least one observed vector: $\forall k \leq K$, $\exists n$ s.t. $z_{k,n} = 1$. Matrix $\mathbf{P}$ can be split into 2 parts $\mathbf{P} = [\mathbf{P}_{\mathcal{A}}, \mathbf{P}_{\bar{\mathcal{A}}}]$. The matrix $\mathbf{P}_{\mathcal{A}}$ features the K active directions and $\mathbf{P}_{\bar{\mathcal{A}}}$ the $(D-K)$ unused components. Let $\mathbf{P}_{\mathcal{A}\backslash k}$ denote the matrix obtained by removing the column $\mathbf{p}_k$ from $\mathbf{P}_{\mathcal{A}}$ and $\mathbf{N}_{\mathcal{A}\backslash k}$ a matrix whose $(D-K+1)$ columns form an orthonormal basis for the orthogonal of $\mathbf{P}_{\mathcal{A}\backslash k}$. Since $\mathbf{p}_k \in \mathbf{P}^{\perp}_{\mathcal{A}\backslash k}$ it can be written as $\mathbf{p}_k = \mathbf{N}_{\mathcal{A}\backslash k}\mathbf{v}_k$. Since the prior distribution of $\mathbf{P}$ is uniform on the orthogonal group $\mathcal{O}_D$, $\mathbf{v}_k$ is uniform on the $(D-K+1)$-dimensional unit sphere (Hoff 2009). By marginalizing $\mathbf{P}_{\bar{\mathcal{A}}}$, one obtains

$$\begin{aligned} &f(\mathbf{v}_k|\mathbf{Y}, \mathbf{P}_{\mathcal{A}\backslash k}, \mathbf{Z}, \delta_k^2, \sigma^2) \propto \\ &\exp\left(\frac{1}{2\sigma^2}\frac{\delta_k^2}{1+\delta_k^2}\mathbf{v}_k^T\mathbf{N}_{\mathcal{A}\backslash k}^T\left(\sum_{n=1}^{N} z_{k,n}\mathbf{y}_n\mathbf{y}_n^T\right)\mathbf{N}_{\mathcal{A}\backslash k}\mathbf{v}_k\right) \end{aligned} \tag{22}$$

which is a Bingham distribution on the $(D-K+1)$-unit sphere, see Section 2.1. As a consequence,

$$\begin{aligned} f(\mathbf{p}_k|\mathbf{Y},\mathbf{P}_{\mathcal{A}\setminus k},\mathbf{Z},\delta_k^2,\sigma^2) \propto \\ \exp\left(\frac{1}{2\sigma^2}\frac{\delta_k^2}{1+\delta_k^2}\mathbf{p}_k^T\left(\sum_{n=1}^N z_{k,n}\mathbf{y}_n\mathbf{y}_n^T\right)\mathbf{p}_k\right) \end{aligned} \tag{23}$$

Sampling the scale parameters δ_k^2. The posterior distribution of δ_k^2 for all k, can be rewritten as

$$\begin{aligned} f\left(\delta_k^2|\mathbf{P},\mathbf{Z},\sigma^2\right) &\propto \left(\frac{1}{1+\delta_k^2}\right)^{a_\delta+\frac{1}{2}\sum_n z_{k,n}+1} \\ \exp&\left[-\frac{1}{1+\delta_k^2}\left(b_\delta+\frac{1}{2\sigma^2}\sum_{k,n} z_{k,n}\left(\mathbf{p}_k^T\mathbf{y}_n\right)^2\right)\right]. \end{aligned} \tag{24}$$

which is a shifted Inverse Gamma distribution.
Sampling the noise variance σ^2. By looking carefully at (15), one obtains

$$\begin{aligned} \sigma^2|\mathbf{Y},\mathbf{Z},\mathbf{P},\boldsymbol{\delta} \sim \mathcal{IG}\Bigg(\frac{ND}{2}, \\ \frac{1}{2}\operatorname{tr}\left(\mathbf{Y}\mathbf{Y}^T\right)-\sum_{k,n}\frac{1}{2}\frac{\delta_k^2}{1+\delta_k^2}z_{k,n}\left(\mathbf{y}_n^T\mathbf{p}_k\right)^2\Bigg). \end{aligned} \tag{25}$$

Sampling the IBP parameter α. The conditional posterior distribution of α is gamma distributed

$$\alpha|\mathbf{Y},\mathbf{Z} \sim \mathcal{G}\left(K,\sum_{n=1}^N\frac{1}{n}\right). \tag{26}$$

Algo. 2 describes the full sampling procedure.

5 Estimators: theoretical properties

Since one motivation of the proposed BNP-PCA approach is its expected ability to identify a relevant number of degrees of freedom of the proposed model, this section focuses on this aspect. Section 5.1 derives theoretical results concerning the marginal maximum a posteriori (MAP) estimator of K associated with the proposed IBP-based model. In particular, Theorem 1 apparently brings some bad news by showing that this estimator is not consistent when the parameter α of the IBP is fixed. Similar results have been reported by Chen et al (2016) on an empirical basis only. Note that our approach considers α as an unknown parameter as well, which may explain the good behaviour observed experimentally in Section 6. Section 5.2 proposes an efficient way to select the right number of components based on simple statistical tests. Section 5.3 deals with estimators of other parameters.

5.1 Posterior distribution of the subspace dimension

The consistency of Dirichlet process mixture models (DPMMs) for Bayesian density estimation has been widely studied, see Ghosal (2009) and references therein. For instance, posterior consistency of such DPMMs with a normal kernel has been obtained by Ghosal et al (1999). While such results tend to motivate the use of nonparametric priors, a certain care should be paid regarding the behaviour of any posterior distribution. For instance, McCullagh and Yang (2008) have provided both experimental and analytical results about the ability of DPMMs to identify and separate two clusters. More recently, Miller and Harrison (2013, 2014) have shown that the posterior distribution of the number of clusters of DPMMs and Pitman-Yor process mixture models are not consistent. When the number of observations tends to infinity, the marginal posterior does not concentrate around any particular value, despite the existence of concentration rates. Fewer results are available when an IBP is used, see for instance Chen et al (2016) where posterior contraction rates are established for phylogenetic models.

The following theorem shows that the marginal MAP estimator of the number of components K is not consistent when conditioned upon (fixed) α.

Theorem 1 *Let $\mathbf{Y}_N = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ denote a matrix of N D-dimensional observations. Let K_N denotes the random variable associated with the latent subspace dimension of the model described in Section 3. Then, the two following assertions*

$$\forall k < D \quad \limsup_{N\to\infty} \; \mathrm{P}\big[K_N = k \mid \mathbf{Y}_N, \alpha\big] \qquad < 1 \tag{27}$$

$$\limsup_{N\to\infty} \; \mathrm{P}\big[K_N = D \mid \mathbf{Y}_N, \alpha\big] \qquad > 0 \tag{28}$$

are true.

Proof See Appendix F.

As discussed in the proof, Eq. (27) can be extended to a wider range of models, while Eq. (28) results from the orthogonality constraint. Up to our knowledge, no similar results have been derived for the IBP. We emphasize that Theorem 1 does not claim that the marginal MAP estimator of the subspace dimension defined as

$$\widehat{K}_{\mathrm{mMAP},\alpha} = \underset{k\in\{0,\ldots,D\}}{\operatorname{argmax}} \; \mathrm{P}\big[K = k \mid \mathbf{Y}_N, \alpha\big] \tag{29}$$

is biased or irrelevant. However, a corollary of Eq. (27) is that this estimator is not consistent. This can be explained by a certain leakage of the whole mass towards the probability of having $K = D$, as shown by Eq. (28). To overcome this issue, instead of resorting to the conventional marginal MAP estimator of the dimension, an alternative strategy will be proposed in Section 5.2 to identify the dimension of the relevant subspace.

By considering an additional hypothesis on the distribution of the measurements $\mathbf{Y}_N$, the following theorem states an interesting result.

Theorem 2 *Let $\mathbf{y}_1, \ldots, \mathbf{y}_N$ be N D-dimensional observations independently and identically distributed according to a centered Gaussian distribution of common variance $\sigma_{\mathbf{y}}^2$. Then*

$$\mathrm{P}\big(K_N = 0 | \mathbf{Y}_N, \alpha, \sigma_{\mathbf{y}}^2\big) \xrightarrow[N\to+\infty]{\text{a.s.}} 0. \tag{30}$$

Proof See Appendix G.

Two distinct interpretations of this theorem can be proposed. Indeed the Gaussian assumption is used twice in this case: both the data and the noise are Gaussian. On one hand, if Gaussian measurements are interpreted as noise, i.e., $\mathbf{y}_n = \boldsymbol{\varepsilon}_n$ and $\sigma_{\mathbf{y}}^2 = \sigma^2$ in the proposed latent factor model (2), the expected dimension of the latent subspace should be 0. Theorem 2 states that this will almost surely not be the case, so that $\widehat{K}_N$ is inconsistent. On the other hand, the same Theorem 2 can be positively interpreted since one would rather expect to find $\widehat{K}_N = D$ since white Gaussian noise spreads its energy equally in every direction. With respect to this second interpretation, $\widehat{K}_N$ may be considered as consistent.

In the present approach, we consider that a latent subspace is meaningful as soon as it permits to distinguish a signal from white Gaussian noise: we stick to the first interpretation of Theorem 2 and consider that $\widehat{K}_N$ is inconsistent. Finally, we emphasize that the two theorems above are related to posterior estimators of K conditioned upon α and possibly σ^2. A posterior estimator $\widehat{K}_{\text{mMAP}}$ will be defined later by Eq.(31) where parameters α and σ^2 are marginalized. Experiments conducted in Section 6 will show that this $\widehat{K}_{\text{mMAP}}$ seems to be asymptotically consistent.

5.2 Selecting the number of components

As emphasized in Section 5.1, the posterior probabilities $\mathrm{P}\big[K|\mathbf{Y},\alpha\big]$ may not to be sufficient to properly derive reliable estimates of the subspace dimension and select the number of relevant directions. However, the proposed BNP-PCA considers the IBP parameter α as unknown. Then one can define the marginalized MAP estimate

$$\widehat{K}_{\text{mMAP}} = \underset{k\in\{0,\ldots,D\}}{\operatorname{argmax}} \ \mathrm{P}\big[K = k \mid \mathbf{Y}\big]. \tag{31}$$

The numerical study of Section 6 will show that it seems to be consistent contrary to $\widehat{K}_{\text{mMAP},\alpha}$. As a consequence, as soon as sufficient amount of data is available, one may use $\widehat{K}_{\text{mMAP}}$ for model selection.

Another possibility, with theoretical guarantees, is to take advantage of the posterior distribution of the principal components $\mathbf{P}$ and to use statistical tests. In accordance with the notations introduced in Section 4.2, let $\mathcal{A} \subset \{1,\ldots,D\}$ denote the set of K indices corresponding to the estimated active directions in $\mathbf{P}$. Elaborating on (23), the posterior distribution of $\mathbf{p}_k$, $\forall k \in \bar{\mathcal{A}}$, can be expressed thanks to a $(D-K)$-dimensional unit-norm random vector $\boldsymbol{w}_k = \mathbf{N}_{\mathcal{A}}^T\mathbf{p}_k$ whose distribution is given by

$$f\left(\boldsymbol{w}_k|\mathbf{Y}_N,\mathbf{P}_{\mathcal{A}},\mathbf{Z},\delta_k^2,\sigma^2\right) \propto \exp\left(\boldsymbol{w}_k^T\boldsymbol{\Lambda}_{k,N}\boldsymbol{w}_k\right) \tag{32}$$

where

$$\boldsymbol{\Lambda}_{k,N} = \gamma_k \sum_{n=1}^{N} \mathbf{N}_{\mathcal{A}}^T\mathbf{y}_n\mathbf{y}_n^T\mathbf{N}_{\mathcal{A}} \tag{33}$$

where $\mathbf{N}_{\mathcal{A}}$ is a $D\times(D-K)$ orthogonal matrix which spans the null space of $\mathbf{P}_{\mathcal{A}}$; γ_k depends on σ^2 and δ_k^2. Interestingly, if $\mathbf{P}_{\mathcal{A}}$ correctly identifies the unknown signal subspace of dimension K, any component $\mathbf{p}_\ell$, $\ell \in \bar{\mathcal{A}}$ is actually a non-relevant direction. According to the latent factor model (2), the projected vectors $\mathbf{N}_{\bar{\mathcal{A}}}^T \mathbf{y}_n$ $(n = 1, \ldots, N)$ in (33) should reduce to white Gaussian noises so that

$$\lim_{N \to +\infty} \frac{1}{N} \boldsymbol{\Lambda}_{k,N} = \gamma_k \sigma^2 \mathbb{I}_{D-K} \tag{34}$$

where $\mathbb{I}_{D-K}$ is the $(D-K)$ identity matrix. This means that the posterior distribution (32) of the $\boldsymbol{w}_\ell$ tends to be uniform over the $(D-K)$-dimensional sphere. Let $L = D - K$ and $\boldsymbol{\mathcal{W}}_{\bar{\mathcal{A}}}$ the $L \times L$ orthogonal matrix whose columns are the vectors $\{\boldsymbol{w}_\ell\}_{\ell \in \bar{\mathcal{A}}}$. One could think of building tests of goodness-of-fit able to identify the maximum dimension $L = D - K \in \{0, \ldots, D\}$ for which $\boldsymbol{\mathcal{W}}_{\bar{\mathcal{A}}}$ remains uniformly distributed over the orthogonal group $\mathcal{O}_L$. However, since $\boldsymbol{\mathcal{W}}_{\bar{\mathcal{A}}}$ lives in a possibly high dimensional space, this testing procedure would be inefficient to provide a reliable decision rule. As an alternative, we propose to conduct a statistical tests on the set of the following $L = D - K$ absolute scalar products

$$\omega_\ell \triangleq |\boldsymbol{w}_\ell^T \mathbf{u}_\ell|, \quad \ell \in \bar{\mathcal{A}}, \tag{35}$$

where the $\{\mathbf{u}_\ell\}_{\ell \in \bar{\mathcal{A}}}$ is a set of L arbitrary L-dimensional unit-norm vectors, for instance uniformly distributed on the sphere. Indeed, if $\boldsymbol{\mathcal{W}}_{\bar{\mathcal{A}}}$ is uniformly distributed over the orthogonal group $\mathcal{O}_L$, the distribution of the L-dimensional random vector $\boldsymbol{\omega}_{\bar{\mathcal{A}}}$ whose components are given by (35) can be easily derived as stated by the following theorem.

Theorem 3 *Let $K \in \{0, \ldots, D\}$, $\boldsymbol{\mathcal{W}} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{D-K}]^T$ be a random matrix uniformly distributed on the orthogonal group $\mathcal{O}_{D-K}$, and $\mathbf{u}_1, \ldots, \mathbf{u}_{D-K}$ be $L = (D-K)$ arbitray unit-norm L-dimensional vectors. Let $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_L]^T$ such that $\omega_\ell \triangleq |\boldsymbol{w}_\ell^T \mathbf{u}_\ell|$. Then, the components of $\boldsymbol{\omega}$ are identically distributed and the cumulative distribution (cdf) of any component ω_ℓ is given by*

$$\begin{aligned} \mathrm{P}(\omega_\ell \le \lambda) &= \frac{\mathrm{vol}(\mathcal{O}_{L-2})}{\mathrm{vol}(\mathcal{O}_{L-1})} 2 \int_0^\lambda \left(1 - z^2\right)^{(L-3)/2} \mathrm{d}z \\ &= 2\lambda_l \frac{\mathrm{vol}(\mathcal{O}_{L-2})}{\mathrm{vol}(\mathcal{O}_{L-1})} {}_2F_1\left(\frac{1}{2}, -\frac{L-3}{2}; \frac{3}{2}; \lambda^2\right). \end{aligned} \tag{36}$$

Proof See Appendix E.

Note that the ω_ℓ can be interpreted as generalized cosines in dimension $L = D - K$. The distribution Eq. (36) depends on the difference $D - K$ only. Fig. 2 shows the empirical and theoretical pdf's associated with the cdf (36) for various values of $D - K$.

We propose to use Theorem 3 to design the following Kolmogorov-Smirnov test of goodness-of-fit applied to the matrices $\left\{\mathbf{P}^{(t)}\right\}_{t=n_{\mathrm{bi}}}^{n_{\mathrm{mc}}}$ generated by the Gibbs sampler detailed in Algo. 2. For a given candidate $\mathcal{A}$ of K indices associated with the subspace spanned by $\mathbf{P}_{\mathcal{A}}$, one can test whether the remaining set $\bar{\mathcal{A}}$ of indices corresponds to directions $\mathbf{P}_{\bar{\mathcal{A}}}$ uniformly distributed over the orthogonal group $\mathcal{O}_{D-K}$. Thanks to Theorem 3 this is equivalent to test whether the absolute scalar products (35) are distributed according to (36). Note also that the random variables $\{\omega_\ell\}_{\ell \in \bar{\mathcal{A}}}$ form a set of identically distributed components of a L-dimensional random vector $\boldsymbol{\omega}_{\bar{\mathcal{A}}}$. This permits to use a single statistical test to be performed for

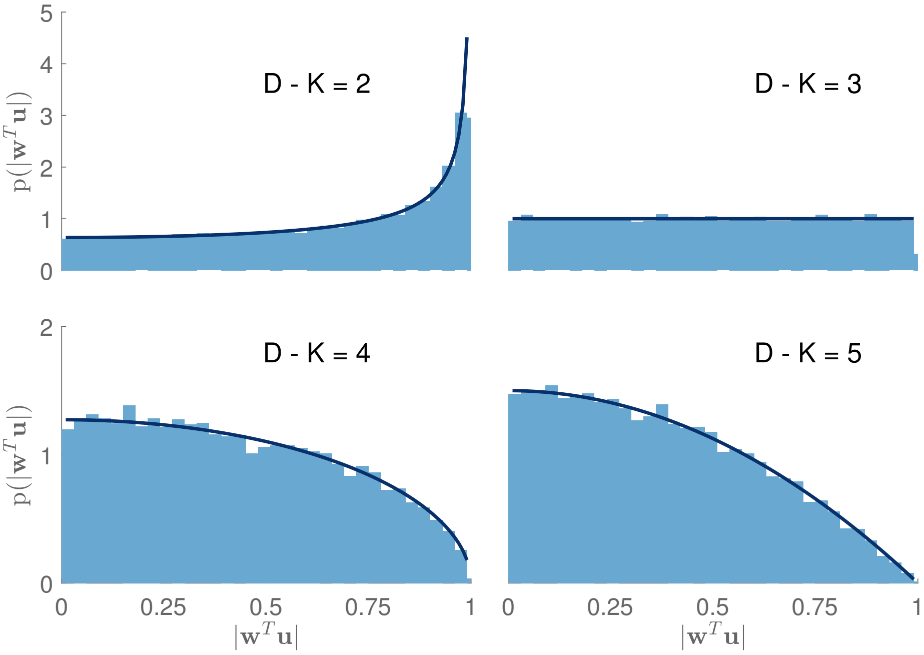

Fig. 2 Empirical (light blue bars, computed from 20000 samples) and theoretical (dark blue lines) pdf's associated with the cdf (36) for 4 different values of the dimension.

Algorithm 3: Selecting the number of relevant directions

```
Input: level of KS test; a collection of samples $\left\{\mathbf{p}_1^{(t)}, \ldots \mathbf{p}_D^{(t)}, \mathbf{Z}^{(t)}\right\}_{t=n_{\text{burn}}+1}^{T_{MC}}$
       generated by Alg. 2.
For each iteration, relabel the directions $\mathbf{p}_k^{(t)}$ w.r.t. their frequency of activation, given
  by $\mathbf{Z}^{(t)}$;
Sample $\mathbf{u}_1 \ldots \mathbf{u}_D \overset{\text{i.i.d.}}{\sim} \mathcal{S}_D^1$ ;
for $K \leftarrow 1$ to $D-1$ do
    for $t \leftarrow n_{\text{burn}}+1$ to $n_{\text{burn}}+n_{\text{iter}}$ do
        Let $\mathbf{N}_K$ be a basis of the orthogonal of $\mathbf{p}_1^{(t)} \ldots \mathbf{p}_K^{(t)}$ ;
        Compute $\omega_{K+1}^{K\,(t)} \triangleq \|\mathbf{N}_K^T \mathbf{u}_{K+1}\|^{-1} |\mathbf{p}_{K+1}^{(t)\,T} \mathbf{N}_K^T \mathbf{u}_{K+1}|, \ldots$
          $\omega_D^{K\,(t)} \triangleq \|\mathbf{N}_K^T \mathbf{u}_D\|^{-1} |\mathbf{p}_D^{(t)\,T} \mathbf{N}_K^T \mathbf{u}_D|$ ;
    end
    Stack the $\omega_{K+1}^{K\,(t)}, \ldots \omega_D^{K\,(t)}$ into a single collection of samples in view of
      Kolmogorov-Smirnov's test ;
    if $\mathcal{H}_K$ is not rejected then
        $\hat{K}_{\text{KS}} = K$ ;
        break;
    end
end
Output: $\hat{K}_{\text{KS}}$, an estimator of the number of relevant components.
```

each dimension candidate K iteratively in increasing or decreasing order, rather than $D - K$ multiple tests. The null hypothesis is defined as

$$\mathcal{H}_0^{(K)} : \omega_\ell \overset{\text{cdf}}{\sim} (36), \quad \forall \ell \in \bar{\mathcal{A}} = \{D - K + 1, ..., D\} \tag{37}$$

Obviously, if this null hypothesis is accepted for a given set $\bar{\mathcal{A}}$ of $D - K$ indices, it will be accepted for any subset of lower dimension. Conversely, if this null hypothesis is rejected for some K and a given set $\bar{\mathcal{A}}$ of $D - K$ indices, it will be definitely rejected for any superset of $\bar{\mathcal{A}}$, that is for subspace dimensions smaller than K. Since the objective of the proposed procedure is to identify an a priori small number K of relevant components (and not a lower or upper bound), this hypothesis should be tested for an increasing number K of active components. As a result, the following estimator $\widehat{K}_{\text{KS}}$ of the number of active components is finally proposed:

$$\widehat{K}_{\text{KS}} = \min \left\{ K \in \{0, \ldots, D\} \mid \mathcal{H}_0^{(K)} \text{ is accepted} \right\}. \tag{38}$$

By convention, $\mathcal{H}_0^{(D)}$ is accepted when $\mathcal{H}_0^{(K)}$ has been rejected for all $K \in \{0, \ldots, D - 1\}$: thus the model would identify data to white Gaussian noise with no special direction. Algo. 3 describes the full procedure.

5.3 Estimating other parameters

This section discusses the derivation of estimates associated with the remaining parameters, other than the dimension K of the subspace. Regarding the orthonormal matrix $\mathbf{P}$ of which the K first columns span the signal subspace, it is not recommended to use a simple average of the samples $\mathbf{P}^{(t)}$ generated by the MCMC algorithm to approximate the minimum mean square error (MMSE) estimator. Indeed, the Markov chain targets a highly multimodal distribution with modes that depend on the current state of the dimension $K^{(t)}$. In particular, at a given iteration t, the last $D - K^{(t)}$ columns of $\mathbf{P}^{(t)}$ are directly drawn from a uniform prior. One alternative is to compute the MMSE estimator conditioned upon an estimate $\widehat{K}$ of the relevant dimension. This can be easily done by averaging the samples $\mathbf{P}^{(t)}$ corresponding to the iterations t for which $K^{(t)} = \widehat{K}$. A similar procedure applies for the binary matrix $\mathbf{Z}$. Note that in the specific context of parametric subspace estimation, other Bayesian estimators have been proposed by Besson et al (2011, 2012).

Remark: the posterior distribution of the scale parameters $\boldsymbol{\delta} = \left\{\delta_1^2, \ldots, \delta_K^2\right\}$, where the matrix $\mathbf{P}$ has been marginalized, cannot be derived analytically. This posterior distribution can be derived explicitly in some very particular cases only, assuming that the binary matrix $\mathbf{Z}$ is the $K \times N$ matrix $\mathbf{1}_{K,N}$ with only 1's everywhere, see App. H for details. The resulting posterior involves a generalized hypergeometric function of two matrices that could be used as a measure of mismatch between the magnitudes of the eigenvalues of covariance matrices. We leave this open question for future work.

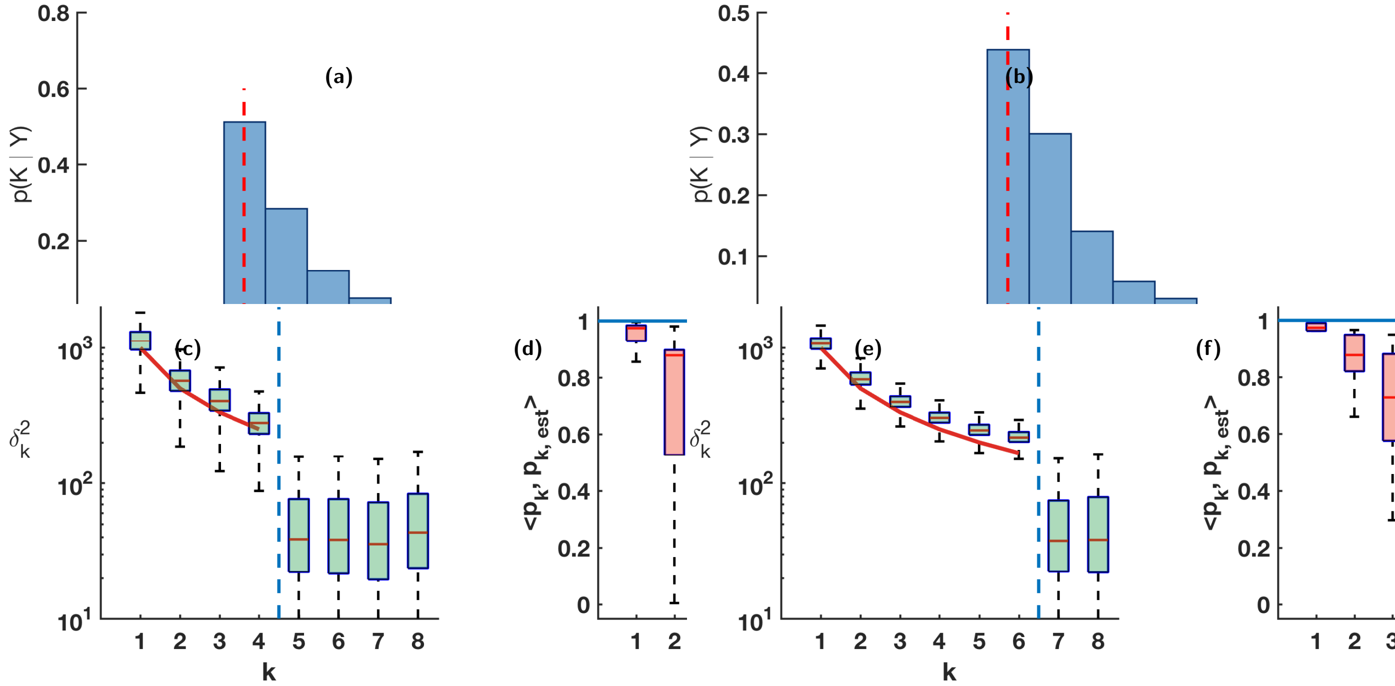

Fig. 3 Top : posterior distribution of K for (a) $D = 16, N = 100$, and (b) $D = 36, N = 500$. Bottom : posterior distributions of (c) & (e) scale factors δ_K^2 and (d) & (f) dispersion of the projection $\widehat{\mathbf{P}}^T\mathbf{P}$ for $D = 16, N = 100$ and $D = 36, N = 500$, respectively. The red lines indicates the true values of $\delta_1^2 \ldots \delta_K^2$.

6 Performance assessment of BNP-PCA

The performance of the proposed BNP-PCA is assessed on datasets simulated according to the linear model

$$\mathbf{y}_n = \mathbf{H}\mathbf{u}_n + \mathbf{e}_n \tag{39}$$

where $\mathbf{e}_n$ is an additive Gaussian noise of covariance matrix $\sigma^2\mathbb{I}_D$ and the quantities $\mathbf{H}$ and $\mathbf{u}_n$ are specified as follows. First, for a given dimension D of the observations, K orthonormal directions are gathered in a $D \times K$ matrix $\mathbf{H}$ which is uniformly generated on the Stiefel manifold $\mathcal{S}_D^K$. Then, N representation vectors $\mathbf{u}_1, \ldots, \mathbf{u}_N$ of dimension K are identically and independently generated according to a centered Gaussian distribution with a diagonal covariance matrix $\boldsymbol{\Sigma} = \operatorname{diag}\left\{\delta_1^2\sigma^2, \ldots, \delta_K^2\sigma^2\right\}$ where the scale factors $\delta_1^2, \ldots, \delta_K^2$ control the relevance of a particular direction. Equivalently, by choosing different values for the scale factors, this model also conveniently permits to consider the case of an anisotropic noise corrupting an isotropic latent subspace. In the following, the choice of these scale factors will be specified in four typical scenarios.

Since each scale factors δ_k^2 controls the signal-to-noise ratio in each direction, a unique value $\sigma^2 = 0.01$ of the noise variance is considered without loss of generality. Several dimensions D and K are considered for various numbers of observations N.

The proposed Gibbs sampler has been run during 1000 iterations after a burn-in period of 100 iterations.

6.1 Scale factors and alignment of components

The performances of the proposed algorithm have been first evaluated on various simulated datasets. As an illustration, we report here the results on 2 datasets corresponding to $(D = 16, K = 4, N = 100)$ and $(D = 36, K = 6, N = 500)$ and where the scale coefficients δ_k^2 are defined as proportional to $1/k$.
Fig. 3(a) & (b) show the posterior distributions of K for $(D = 16, N = 100)$ and $(D = 36, N = 500)$, respectively. We observe that the maximum of the two posterior histograms correspond to the expected dimension, *i.e.*, $K = 4$ for $D = 16$ and $K = 6$ for $D = 36$. Note that this estimator corresponds to the marginal maximum a posteriori estimator defined by Eq. (31). These two examples suggest that the marginal MAP estimator $K|\mathbf{Y}$ seems to be consistent since it is able to recover the expected dimension. This is in contrast with the behaviour of the conditional MAP estimator $K|\mathbf{Y}, \alpha$ that is known to be inconsistent from Theorems 1 and 2. Section 6.2 will come back to this question in more details. We do not comment on the behaviour of $\widehat{K}_{\text{KS}}$ based on KS tests here: in such simple scenarios, $\widehat{K}_{\text{KS}}$ and $\widehat{K}_{\text{KS}}$ always give the same results.
Fig. 3 (c) & (e) show the posterior distributions of the 8 first scale factors. Fig. 3(d) & (f) show the alignment of the true $\mathbf{p}_k$ with the estimated $\widehat{\mathbf{p}}_k$; see Fig. 3(c)&(d) for $D = 16, N = 100$ and Fig. 3(e)&(f) for $D = 36, N = 500$. The alignment is measured by the scalar product $\langle \mathbf{p}_k, \widehat{\mathbf{p}}_k \rangle$ between each column of $\mathbf{P}$ and its estimate. No ordering problem is expected here since the variances are sufficiently different in every direction. In both cases, it appears that scale factors are correctly identified. We observe that inferred directions correspond to actual principal components with an alignment typically higher than 0.8 on average. All other components, for $k \geq 5$ on Fig. 3(d) and $k \geq 7$ on Fig. 3(f)), are considered as inactive since associated to components with comparable factors and much lower alignment. This observation motivated the procedure proposed in Section 5.2 elaborated on KS tests to build the estimator $\widehat{K}_{\text{KS}}$, see Eq. (38). Recall that $\widehat{K}_{\text{KS}}$ will be especially useful when the signal to noise ratio is close to 1 for some components, that is $\delta_k^2 \simeq 1$.

These first experiments show that the proposed BNP-PCA is able to identify the relevant latent subspace through its dimension K as well as principal components $\mathbf{p}_k$ and their corresponding scale factors δ_k^2. They also indicate that $\widehat{K}_{\text{mMAP}}$ seems to be consistent in contrast with $\widehat{K}_{\text{mMAP},\alpha}$, see Theorems 1 & 2 of Section 5.1.

6.2 Marginal MAP estimator of the latent dimension

This section experimentally investigates the behaviour of the marginal MAP estimator $\widehat{K}_{\text{mMAP}}$ of the dimension of the latent subspace defined by (31). Note that this estimator is different from the marginal MAP estimator $\widehat{K}_{\text{mMAP},\alpha}$ defined in (29) which was still conditioned upon α. Indeed, in the Bayesian model proposed in Section 3, a prior distribution is assigned to the hyperparameter α which is thus

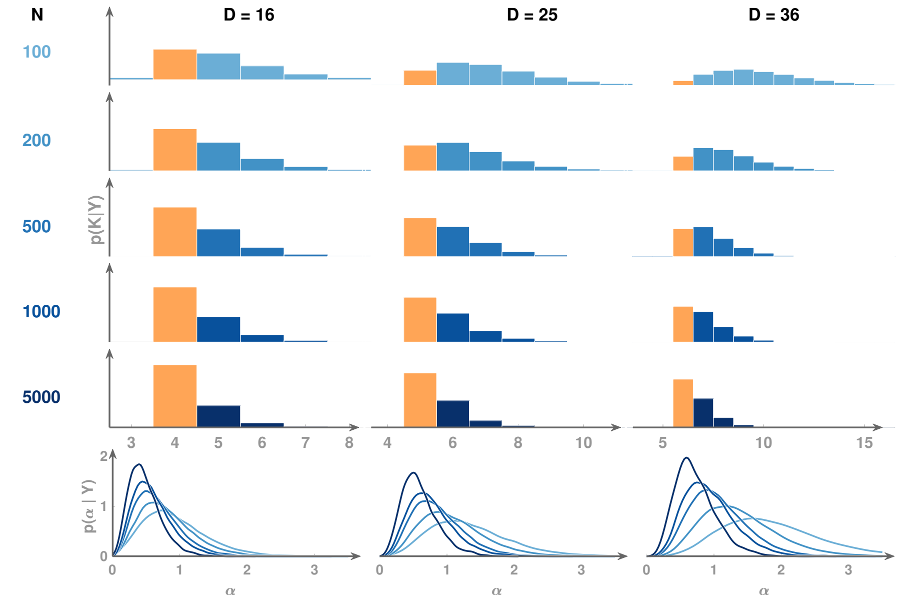

Fig. 4 Empirical posterior probabilities $\mathrm{P}[K = k|\mathbf{Y}]$ of the latent dimension for (left) $D = 16$, (center) $D = 25$, (right)$D = 36$ and $N \in \{100, 200, 500, 1000, 5000\}$. The orange bars indicate the true dimension K of the latent subspace. Bottom plots are the empirical marginal posterior distributions $f(\alpha|\mathbf{Y})$ where the number of observations N increases when the line color goes from light to dark blue lines.

jointly inferred with the parameters of interest. While Theorem 1 of section 5.1 says that $\widehat{K}_{\mathrm{mMAP},\alpha}$ with fixed α is inconsistent, we will empirically show that $\widehat{K}_{\mathrm{mMAP}}$ seems to be consistent.

Fig. 4 shows the empirical posterior probabilities $\mathrm{P}[K = k|\mathbf{Y}]$ when all the scaling factors have been fixed to values significantly higher than 1, such that $\delta_k^2 = 50/k$, $1 \leq k \leq K$. Actual subspace dimensions are $K = \sqrt{D}$ for $D \in \{16, 25, 36\}$. This figure shows that, for $D = 16$, the marginal MAP estimator $\widehat{K}_{\mathrm{mMAP}}$ correctly recovers the latent dimension for all values of N. The proposed model needs around $N = 500$ observations for $D = 25$, and $N = 1000$ for $D = 36$. All posteriors seem to concentrate around the true value $K = \sqrt{D}$ as the number of observations increases: these numerical results suggest a consistent behaviour of the estimator.

These findings do not contradict Theorem 1 which states that the marginal MAP estimator of K is inconsistent *for fixed* α. In contrast, sampling α jointly with the other parameters leads to a marginal MAP estimator $\widehat{K}_{\mathrm{mMAP}}$ which seems to be consistent, at least based on our numerical experiments. By examining the empirical marginal posterior distributions $f(\alpha|\mathbf{Y})$ reported in Fig. 4 (last row), one can note that this distribution seems to get closer to 0 as the number of observations N increases. Exploiting the fact that $\mathbb{E}[K]$ a priori scales as $\alpha \log(N)$, the posterior behaviour of the latent subspace dimension seems to result from a decreasing estimated value of α, this is expected. Moreover, recall that Theorem 1 states that the marginal posterior probabilities $\mathrm{P}[K_N = k|\mathbf{Y}_N, \alpha]$ does not admit

Table 2 Results of Kolmogorov-Smirnov goodness-of-fit tests at level 0.05 averaged over 20 Monte Carlo simulations when the signal is made of $N = 500$ D-dimensional realizations of an isotropic Gaussian noise. Scores reported in each column correspond to the probability of rejecting the null hypothesis for a subspace of candidate dimension K.

K	**0**	**1**	**2**	**3**	**4**	**5**
$D = 9$	0.05	0.05	0.05	0.05	0.05	0
$D = 16$	0.05	0	0.05	0.05	0	0
$D = 25$	0.05	0.1	0.05	0.1	0.05	0
$D = 36$	0.05	0.05	0.05	0.05	0.05	0.05

1 as a limit for any value k. However, it does not state that the mode cannot converge to the true value.

Finally, let us recommend that a certain care be taken anyway when resorting to these posterior probabilities. We have shown that the proposed estimator $\widehat{K}_{\text{mMAP}}$ can exhibit a good asymptotic behaviour, but how this asymptote behaves still seems to depend both on the generative model and the experiment settings and is out of the scope of the present paper.

6.3 The BNP-PCA of white Gaussian noise

In this experiment, the scaling parameters are all chosen as $\delta_k^2 = 0$, leading to observed measurements $\mathbf{y}_n$ $(n = 1, \ldots, N)$ only composed of white Gaussian noise. In this particular case, data do not live in a particular subspace. The purpose of this first basic experiment is to check whether the algorithm is able to detect that no component is relevant, i.e., $K = 0$ since data behaves like white Gaussian noise. More precisely, since the signal is only composed of isotropic noise, the empirical covariance matrix of the observed vectors verifies

$$\lim_{N \to +\infty} N^{-1}\mathbf{Y}\mathbf{Y}^T = \sigma^2 \mathbb{I}_D. \tag{40}$$

According to Section 5.2 and Theorem 3, the posterior distribution of a potential active direction in (23) should asymptotically tend to be $\propto \exp\left(\mathbf{p}^T\mathbf{p}/4\right)$ that is constant since $\mathbf{p}^T\mathbf{p} = 1$ by definition: one expects that the $\mathbf{p}_k$ be uniformly distributed on the unit sphere. BNP-PCA estimates scale factors that are all comparable given the prior. Therefore BNP-PCA does not identify any special latent subspace in this case.

Table 2 shows the results provided by the Kolmogorov-Smirnov (KS) goodness-of-fit test described in Section 5.2. More precisely, for $N = 500$ and $D \in \{9, 16, 25, 36\}$, Table 2 reports the probability of rejecting the null hypothesis $\mathcal{H}_0^{(K)}$ in (37) for candidate dimensions $K \in \{0, \ldots, 5\}$ of the latent subspace, i.e., $L = D - K \in \ldots \{D, \ldots, D - 5\}$. These results computed from 20 Monte Carlo simulations show that the null hypothesis is very often rejected with a probability of the order of 0.05, which corresponds to the chosen rejection level of the KS test here: it is considered as accepted (not rejected). Similar results are obtained for $K \in \{6, \ldots, D\}$. As expected, the estimator $\widehat{K}_{\text{KS}}$ defined by (38) well recovers the actual dimension of the latent subspace, i.e., $K = 0$ here since the data is simply white Gaussian noise only.

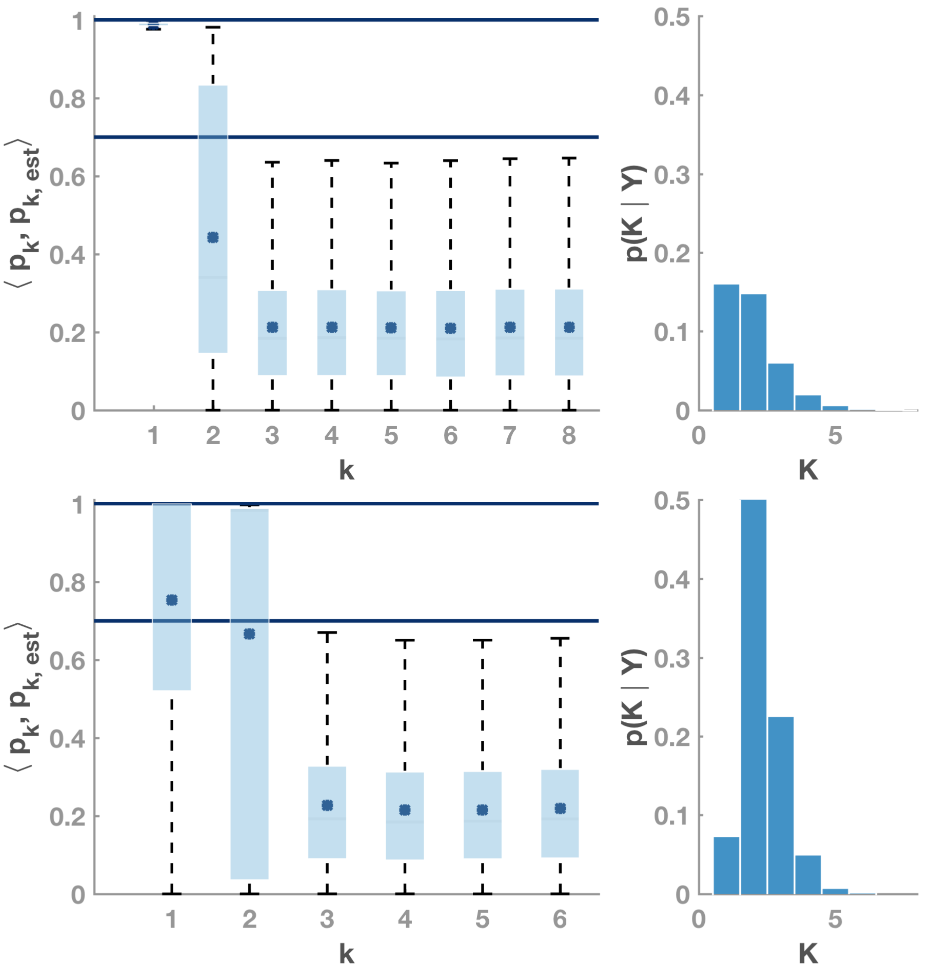

Fig. 5 Marginal posterior distributions in case of signal with anisotropic noise, for $D = 16$, $N = 200$(top) and $N = 2000$(bottom).

6.4 Influence of the distribution of scaling factors

The third experiment aims at investigating two aspects of BNP-PCA. The first question is how far principal components are well recovered. The second aspects concerns the limitations of the proposed method when some scaling factors δ_k^2 are below 1, leading to poorly relevant directions of the latent subspace with respect to the noise level. More precisely, N measurement vectors have been generated according to the model (39) with $N \in \{200, 2000\}$, $D = 16$ and $K = 16$ with scaling factors $\delta_k^2 = 10/k^{2.2}$ $(k = 1, \ldots, K)$, such that the first 5 scaling factors are $[10, 2.2, 0.9, 0.5, 0.3]$; only 2 are larger than 1. This setting permits to play with individual signal-to-noise ratios specified in each direction. Since the scaling factors δ_k^2 are lower than 1 for $k \geq 3$, not all directions are expected to be recovered.

Fig. 5 (right) shows the empirical marginal posterior probability of the latent dimension. These probabilities lead to marginal MAP estimators (31) of the latent dimension equal to $\widehat{K}_{\text{mMAP}} = 2$ for both cases ($N = 200$ and $N = 2000$). The alternative estimator $\widehat{K}_{\text{KS}}$ of the latent subspace derived from the Kolmogorov-Smirnov test (see Section 5.2) leads to estimates between 2 (65% provides $\widehat{K}_{\text{KS}} = 2$ for $N = 200$) and 3 (95% provides $\widehat{K}_{\text{KS}} = 3$ for $N = 2000$). These experiments indicate that BNP-PCA fails to detect principal components weaker than the noise level.

Fig. 5 (left) depicts the estimated inner products $\langle \mathbf{p}_k, \hat{\mathbf{p}}_k \rangle$ and corresponding confidence intervals computed from 50 Monte Carlo simulations where $\hat{\mathbf{p}}_k$ denote the estimated direction vectors. A high score (like a cosine) indicates a good alignment of the vectors, thus a correct recovery of the corresponding latent direction. This figure shows that, for $N = 200$ (top), the proposed model accurately identifies the first component only among the two expected from $\widehat{K}_{\text{mMAP}} = 2$. For larger $N = 2000$ (bottom) the alignment is better and the 2 predicted components are well recovered as attested by the good alignment between the $\hat{\mathbf{p}}_k$ and $\mathbf{p}_k$. However, in both cases, the proposed strategy is not able to extract components with scaling factors δ_k^2 smaller than 1: they are identified to noise, as expected from signal-to-noise ratios.

7 Applications

7.1 BNP-PCA and clustering

To illustrate the flexibility of the proposed model, a simple experiment where the dimension reduction is combined with a linear binary classifier is presented. The representation coefficients in Eq. (8) are now modeled by a mixture of two Gaussian distributions corresponding to 2 distinct clusters

$$\forall n, \quad \mathbf{x}_n \sim \pi \, \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Delta}_0) + (1 - \pi) \, \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Delta}_1), \tag{41}$$

where $\boldsymbol{\mu}_i = [\mu_{i,1}, \ldots, \mu_{i,K}]^T$ and $\boldsymbol{\Delta}_i = \text{diag}\left\{\delta_{i,1}^2, \ldots, \delta_{i,K}^2\right\}$ for $i \in \{0, 1\}$ are respectively the mean and the covariance matrix associated with each class. A common centered Gaussian distribution is used as the prior distribution for the mean vectors $\boldsymbol{\mu}_i$ ($i \in \{0, 1\}$) assumed to be a priori independent, i.e., $\boldsymbol{\mu}_i \sim \mathcal{N}\left(\mathbf{0}, s^2 \mathbb{I}\right)$. Note that the use of non-informative priors are prohibited here due to posterior consistency. Additionally, a binary label vector $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_N]^T$ which indicates whether the nth observation belongs to the class $\mathcal{C}_0$ or $\mathcal{C}_1$ is assigned equiprobable prior probabilities and will be jointly estimated with the parameters of interest. Analytical marginalization w.r.t. to the scale factors remains tractable. All prior distributions are conjugate, yielding conditional posterior distributions that can be easily derived and sampled as described in Section 4.2.

Results on a subset of the MNIST database. The performance of the proposed algorithm is illustrated on a subset of the MNIST database[1], obtained by extracting the first 200 images associated with the digits 6 and 7. Each image is

[1] Available online at http://ufldl.stanford.edu/wiki/index.php/Using_the_MNIST_Dataset

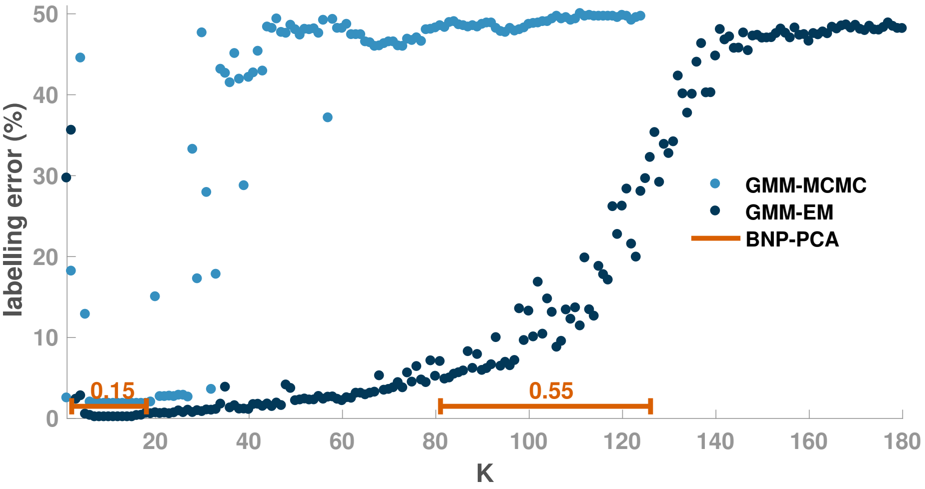

Fig. 6 Clustering results for the 200 first images of the MNIST database for digits 6 and 7.

encoded as a vector in lexicographic order where pixels with null variance (i.e., pixels mainly located in the image corners) have been removed, leading to observation vectors of dimension $D = 572$. The objective of this experiment is to evaluate the need and impact for dimension reduction for this binary classification task. The results provided by the proposed method are compared with those obtained by using an expectation-maximization (EM) algorithm[2] as well as an MCMC algorithm, both inferring the parameters associated with the conventional Gaussian mixture model (41) described above. Both algorithms, denoted respectively by GMM-EM and GMM-MCMC, are preceded by a supervised dimension reduction preprocessing which consists in computing the first K principal components, for a wide set of dimensions K. We emphasize that the proposed BNP-PCA approach combined to an MCMC algorithm for inference addresses jointly the dimension reduction and classification tasks as well as it identifies the dimension of the relevant latent subspace and estimates the noise level.

To overcome the problem of label switching inherent to MCMC sampling of mixture models, the samples generated from the proposed Bayesian nonparametric approach and the Bayesian parametric GMM-MCMC algorithms are postprocessed appropriately (Marin and Robert 2007, Chapter 6-4). More precisely, first, the two farthest observation vectors (in term of Euclidean distance) are assumed to belong to distinct classes. Gibbs sampler iterations leading to equal labels for these two observations are discarded. For remaining iterations, all the generated labels are reassigned in agreement with consistent labels for these two particular observations.

Classification performance is evaluated by the resulting labeling errors. All results have been averaged over 20 Monte Carlo simulations.

Fig. 6 shows the clustering results for the 2 parametric methods compared to BNP-PCA. Both parametric methods, GMM-EM and GMM-MCMC, show labeling errors close to 1% when using few principal components as input features, but

[2] Available through the *gmdistribution* class of MATLAB.

exhibit a phase transition leading to error up to 50% when retaining too much principal components. Note that the phase transition occurs later for the EM-based algorithm that seems to be more robust, but a more elaborated MCMC method may have exhibited a similar performance. The proposed Bayesian nonparametric method shows an average labeling error of about 1.5%. Fig. 6 indicates the typical ranges of values visited by the sampled latent dimension (brown lines). The intervals $K \in [3, 18]$ and $K \in [83, 130]$ correspond to 70% of the samples. It is noticeable that the two parametric methods reach their best performance when considering a number K of principal components belonging to the first interval.

7.2 Hyperspectral subspace identification

As a second pratical illustration, the BN-PCA is employed to solve a key preprocessing task for the analysis of hyperspectral images. An hyperspectral image consists of a collection of several hundreds or thousands of 2D images acquired in narrow and contiguous spectral bands. Such images can be interpreted as a collection of spectra measured at each pixel location. A classical objective is the recovering of spectral signatures of the materials that are present in the scene as well as their spatial distributions over the scene. A common assumption in spectral unmixing is to consider that each measured spectrim is a noisy convex combination of the unknown elementary spectral signatures called *endmembers*. The combination coefficients correspond to the unknown proportions to be estimated. Thus this so-called spectral unmixing can be formulated as a classical blind source separation or nonnegative matrix factorization problem. One crucial issue lies in the fact that the number R of endmembers (i.e., the order of decomposition/factorization) present in the image is generally unknown in most applicative scenarios. However, under the hypothesis of a linear mixing model, measurements should lie in a K-dimensional linear subspace with $K = R - 1$. As a consequence, most of the spectral unmixing techniques first estimate the relevant latent subspace by a dimension reduction step such as PCA. Then one usually considers (Bioucas-Dias et al 2012) that the number of materials present in the scene is $R = K + 1$. Precisely, the proposed BNP-PCA can identify the number R of components that are significant in an hyperspectral image.

A real hyperspectral image, referred to as “Cuprite hill” and acquired by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over Cuprite, Nevada, is considered. The image of interest consists of 1250 pixels observed in 190 spectral bands after spatial subsampling in horizontal and vertical directions of a factor 2 and after removing the spectral bands of low SNR typically corresponding to the water absorption bands. Then the hyperspectral image has been whitened according to the noise covariance matrix estimated by the strategy described by Bioucas-Dias and Nascimento (2008).

The proposed BNP-PCA based method is compared to the generic methods referred to as L-S and OVPCA introduced by Minka (2000) and Smídl and Quinn (2007), respectively, as well as to the hyperspectral-specific subspace identification algorithm HySime (Bioucas-Dias and Nascimento 2008). The proposed Gibbs sampler has been run during 1100 iterations including a burn-in period of 100 iterations.

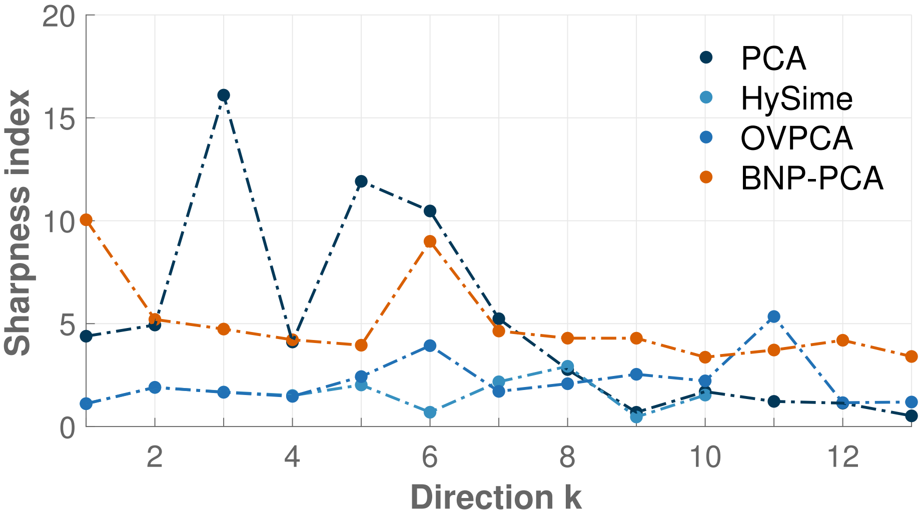

Fig. 7 Sharpness index of the images resulting from the projection onto the directions inferred by PCA (dark blue) and the proposed method (light blue).

The HySime algorithm estimates a hyperspectral subspace of dimension $\widehat{K} = 10$ while L-S and OVPCA lead to $\widehat{K} = 25$ and $\widehat{K} = 23$, respectively. There is no oracle correct number of materials or dimension of the latent subspace. Examining the crude mapping of the materials conducted by Clark et al (1993) and Clark et al (2003) permits to state that it is highly unlikely that more than 15 materials are present in the considered region of interest. Specialists generally agree about a number of components between 10 and 15. It appears that both HySime and OVPCA overestimate the number of endmembers Using BNP-PCA on the same dataset, the marginal MAP estimator defined by 31 yields $\widehat{K}_{\text{mMAP}} = 25$ while the implementation of the Kolmogorov-Smirnov goodness-of-fit test detailed in Section 5.2 leads to a latent subspace dimension estimate $\widehat{K}_{\text{KS}} = 13$ which is quite coherent with the expected value.

To evaluate the relevance of the K directions recovered by BNP-PCA, the measured hyperspectral spectra are orthogonally projected on each direction $\mathbf{p}_1, \ldots, \mathbf{p}_K$. The resulting K images are supposed to explain most of the information contained in the original hyperspectral image with respect to each endmember. They are expected to individually provide relevant interpretation of the scene. The sharpness index introduced by Blanchet and Moisan (2012) as a ground truth-free image quality measure is computed on each image. Figure 7 features the corresponding scores for each direction. These values are compared with those similarly obtained by a standard PCA. Figure 7 shows that our method consistently provides better scores, except for components 3, 5 and 6. This can be empirically explained by the fact that more spatial information (structure and texture) has been recovered by BNP-PCA due to its sparsity promoting property. It ensures a better separation between relevant components and purely random white process than the images projected on the principal components identified by a standard PCA.

8 Conclusion

This paper indroduces a Bayesian nonparametric principal component analysis (BNP-PCA). This approach permits to infer the orthonormal basis of a latent subspace in which the signal lives as an information distinct from white Gaussian noise. It relies on the use of an Indian buffet process (IBP) prior which permits to deal with a family of models with a potentially infinite number of degrees of freedom. The IBP features two regularizing properties: it promotes sparsity and penalizes the number of degrees of freedom.

Algorithms implementing a Markov chain Monte Carlo (MCMC) sampling are described for all parameters according to their conditional posterior distributions. BNP-PCA appears to be close to completely nonparametric since no parameter tuning or initialization is needed and the most general priors are used. Compared to a parametric approach based on RJ-MCMC, the Markov chain is much easier to implement and mixes much more rapidly. One limitation of the proposed approach is the use of MCMC for inference: faster estimates may be obtained by resorting to variational inference for instance.

Since one may be interested in a BNP approach to estimate the dimension K of the latent subspace (or equivalently the number of degrees of freedom), we have studied the theoretical properties of some estimators based on BNP-PCA in the case where the parameter α of the IBP is fixed. Theorems 1 & 2 show that the marginal MAP (mMAP) estimate of K is not consistent in this case: its posterior does not asymptotically concentrate on any particular value as the number of observations increases.

Numerical experiments show that the proposed BNP-PCA that considers the parameter α of the IBP as an unknown parameter yields very good results. In particular, experimental results indicate that the mMAP estimate of K seems to be consistent (as soon as α is not fixed anymore). To make our approach even more robust, we have elaborated on a Kolmogorov-Smirnov test to propose a method to accurately identify the dimension of the relevant latent subspace. An expected limitation is that a principal component may not be recovered when its energy/eigenvalue is below the noise level. Finally, we have applied BNP-PCA to two classical problems: clustering based on Gaussian models mixture applied to the MNIST dataset and linear unmixing of hyperspectral images (or more generally matrix factorization). The clustering performance of the proposed approach is very good. The inspection of the significance of the elementary images (also called endmembers) estimated from a hyperspectral image is in favour of BNP-PCA compared to standard PCA: each component seems to extract more detailed information as attested by image-guided diagnosis. Performed on real datasets, these experiments show that BNP-PCA can be used in a general Bayesian model and yield good performance on real applications. Again we emphasize that the resulting approach will call for very few parameter tuning only.

Based on these encouraging results, future work will aim at studying the consistency of both the new KS-based estimator and the marginal MAP estimator when the IBP parameter has been marginalized. We plan to use BNP-PCA as a subspace identification strategy in a refined linear hyperspectral unmixing method.

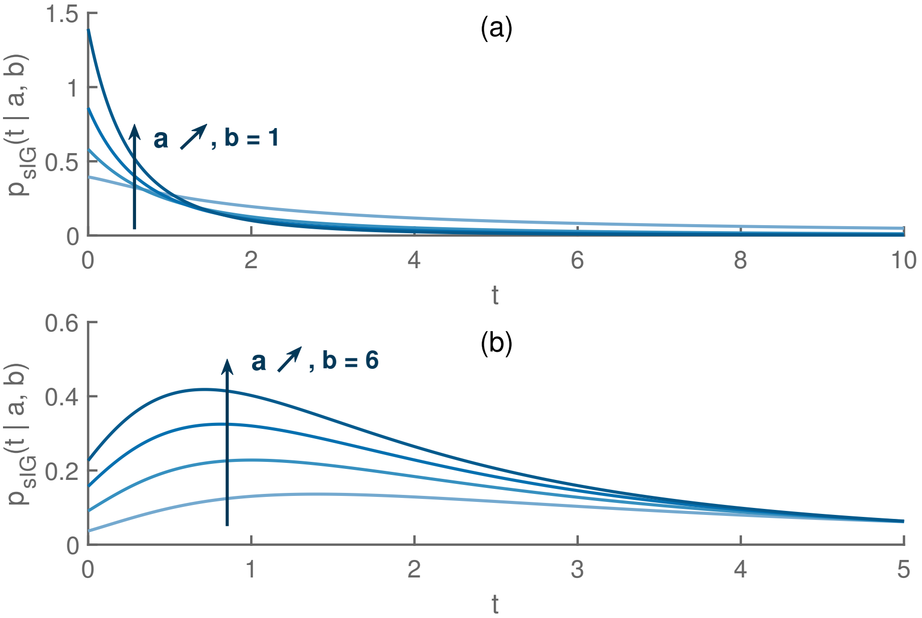

Fig. 8 pdf of the sIG distribution for (a) $a = 0.25, 1, 1.5, 2$ and $b = 1$, and (b) $a = 1.5, 2, 2.3, 2.5$ and $b = 6$.

A Marginalized posteriori distribution

The marginal posterior distribution is obtained by computing

$$f(\boldsymbol{\theta}, \phi | \mathbf{Y}) = \int_{\mathbb{R}^{DN}} f(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{X}) \, f(\boldsymbol{\theta}, \mathbf{X}|\phi) \, f(\phi) \, \mathrm{d}\mathbf{X}.$$

The rationale of the proof is to split the exponential in two. The coefficients $x_{k,n}$ corresponding to non activated block in $\mathbf{Z}$, *i.e.*, for which $z_{k,n} = 0$, vanish. The remaining constant is $\prod_{k=1}^{K} (2\pi\delta_k^2)^{-\mathbf{z}_k^T \mathbf{z}_k/2}$ where $\mathbf{z}_k$ denotes the k^{th} row.

The remaining exponential term becomes

$$-\frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(\|\mathbf{y}_n - \sum_{\substack{k \\ z_{k,n}=1}} \mathbf{p}_k \mathbf{x}_n\|_2^2 + \sum_{\substack{k \\ z_{k,n}=1}} \frac{1}{\delta_k^2} \mathbf{x}_n^T \mathbf{x}_n \right). \tag{42}$$

The ℓ_2 norm in Eq. (42) can be easily simplified since $\mathbf{p}_l^T \mathbf{p}_m = \delta_{l,m}$ where $\delta_{l,m}$ is the Kronecker symbol. In addition, the posterior in Eq. (42) is conjugated to a Gaussian distribution. The remaining terms after integration are a constant $\left(2\pi\delta_k^2\sigma^2/(1+\delta_k^2)\right)^{\mathbf{z}_k^T \mathbf{z}_k/2}$ as well as terms proportional to $\mathbf{y}_n^T \mathbf{p}_k \mathbf{p}_k^T \mathbf{y}_n$ which can be rewritten as $\left(\mathbf{p}_k^T \mathbf{y}_n\right)^2$. The marginal posterior Eq. (15) is obtained by combining all these terms.

B Shifted inverse gamma distribution

The sIG pdf is defined for all real $x > 0$ by

$$\mathrm{p}_{\mathrm{sIG}}(x|a,b) = \frac{b^a}{\gamma(a,b)} (1+x)^{-(a+1)} \exp\left(-\frac{b}{1+x}\right) \tag{43}$$

with shape parameter a and rate parameter b, and $\gamma(a, b) = \int_0^a t^{b-1} e^{-t} \mathrm{d}\, t$ is the lower incomplete gamma function. If $b > a + 1$, it is easy to see that the pdf has a unique maximum in $\frac{b}{a+1} - 1$, but no maximum otherwise. Fig. 8 displays the pdf of the sIG distributions for several values of a and b.

if $X \sim \mathrm{sIG}(a, b)$, the two first moments of X are given by

$$\begin{aligned} \mathbb{E}[X] &= b \frac{\gamma(a-1, b)}{\gamma(a, b)} - 1 && (44) \\ \mathrm{var}(X) &= b^2 \left(\frac{\gamma(a-2, b)}{\gamma(a, b)} - \left(\frac{\gamma(a-1, b)}{\gamma(a, b)} \right)^2 \right). && (45) \end{aligned}$$

Note finally that the sIG distribution can be easily sampled by resorting to the change of variable $u = 1 + \delta_k^2$ where u^{-1} follows a Gamma distribution of parameters a_δ and b_δ truncated on the segment $(0, 1)$.

C Jeffreys' prior for the IBP hyperparameter

By definition, the Jeffreys' prior is given by (Marin and Robert 2007, Ch. 2)

$$f(\alpha) \propto \sqrt{\mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\, \alpha} \log \mathrm{P}\left[\mathbf{Z} | \alpha\right] \right)^2 \right]}. \tag{46}$$

Since $\frac{\mathrm{d}}{\mathrm{d}\, \alpha} \log \mathrm{P}\left[\mathbf{Z} | \alpha\right] = \frac{K}{\alpha} - \sum_{n=1}^{N} \frac{1}{n}$, and does not depend on $\mathbf{Z}$,

$$\mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\, \alpha} \log \mathrm{P}\left[\mathbf{Z} | \alpha\right] \right)^2 \right] = \left(\frac{K}{\alpha} - \sum_{n=1}^{N} \frac{1}{n} \right)^2. \tag{47}$$

Thus $f(\alpha) \propto \alpha^{-1}$.

D Marginalized posterior distribution

The marginal posterior distribution is obtained by integrating the marginal posterior given by Eq. (15) with respect to the parameters δ^2 and α. By mean of conjugacy, straightforward computations lead to

$$\begin{aligned} f\left(\mathbf{P}, \mathbf{Z}, \sigma^2 | \mathbf{Y}\right) = & \left(\frac{1}{2\pi\sigma^2} \right)^{ND/2} \exp\left[\mathrm{trace}\left[-\frac{1}{2\sigma^2} \mathbf{Y}\mathbf{Y}^T \right] \right] \\ & \times \left(\frac{b_\delta^{a_\delta}}{\gamma(a_\delta, b_\delta)} \right)^K \prod_{k=1}^{K} \frac{\gamma(a_k, b_k)}{b_k^{a_k}} \exp\left(\frac{1}{2\sigma^2} \sum_n \left(\mathbf{p}_k^T \mathbf{y}_n \right)^2 \right) \\ & \times \left(\sum_n \frac{1}{n} \right)^{-K} \frac{\Gamma(K)}{\prod_k K_n!} \prod_k \frac{(N - m_k)!\, (m_k - 1)!}{N!} \mathbb{1}_{\mathbb{U}_D}(\mathbf{P}), \end{aligned} \tag{48}$$

where for all k

$$\begin{aligned} a_k &= a_\delta + \mathbf{z}_k^T \mathbf{z}_k \\ b_k &= b_\delta + \frac{1}{2\sigma^2} \sum_n \left(\mathbf{p}_k^T \mathbf{y}_n \right)^2. \end{aligned}$$

E Law and expectation of scalar product

This section derives the marginal distribution of the projections evoked in Theorem 3 under the uniform distribution over $\mathcal{S}_{D-K}^D$.

Area element of the sphere. The rationale of the proof is to adapt the vector to the area element in the D-dimensional Euclidean space expressed in spherical coordinate. The D-dimensional element parametrized by $D-1$ angles is given by

$$\mathrm{d}S^{D}=\sin^{D-2}(\phi_1)\sin^{D-3}(\phi_2)\ldots\sin(\phi_{D-2})\mathrm{d}\phi_1\ldots\mathrm{d}\phi_{D-1},$$

and the Cartesian coordinates $\boldsymbol{v}_1\ldots\boldsymbol{v}_D$ of a vector $\boldsymbol{v}$ are given by

$$\begin{aligned}\boldsymbol{v}_1 &=\cos(\phi_1)\\ \boldsymbol{v}_2 &=\sin(\phi_1)\cos(\phi_2)\\ &\vdots\\ \boldsymbol{v}_{D-1} &=\sin(\phi_1)\ldots\sin(\phi_{D-2})\cos(\phi_{D-1})\\ \boldsymbol{v}_D &=\sin(\phi_1)\ldots\sin(\phi_{D-1}).\end{aligned}$$

The proof considers a non explicit rotation applied to $\boldsymbol{u}$ such that only the last component $\boldsymbol{v}_D$ is involved in the scalar product.

Proof. Let $\mathbf{u}$ be a unit vector of $\mathbb{R}^L$. See L here as the size of the orthogonal of the relevant component, $L=D-K$. Let $\boldsymbol{\nu}$ be a random variable uniformly distributed on the L-dimensional unit sphere. Let also w be the random variable associated to the scalar product $w=|\langle\mathbf{u},\boldsymbol{\nu}\rangle|=|\boldsymbol{\nu}^T\mathbf{u}|$. The density of w will be obtain from the cdf

$$\mathrm{p}_w\left(w\leq\lambda\right)=\mathrm{p}_{\boldsymbol{\nu}}\left(|\boldsymbol{\nu}^T\mathbf{u}|\leq\lambda\right)=\int\mathbf{1}_{|\boldsymbol{\nu}^T\mathbf{u}|}(\boldsymbol{\nu})\mathrm{d}\boldsymbol{\nu},\tag{49}$$

where the sum appearing in the last equation is expressed w.r.t. the Haar measure on the sphere.

Let $\mathbf{R}$ the rotation matrix such that $\mathbf{e}=\mathbf{Ru}$ where $\boldsymbol{e}=[1,0,0,\ldots]$. Since the Haar measure is invariant under rotation, Eq. (49) becomes, once rewritten w.r.t. the area element $\mathrm{d}S^{L-1}$

$$\mathrm{p}\left(w\leq\lambda\right)=\frac{1}{\mathcal{S}_{L-1}}\int\mathbf{1}_{|\cos(\phi_1)|\leq\lambda}(\boldsymbol{v})\mathrm{d}S^{L-1}.$$

Since $|\cos(\phi_1)|\leq\lambda$ if ϕ_1 belongs to the set $[\arccos(\lambda),\pi-\arccos(\lambda)]$, one have, by means of symmetry around $\pi/2$

$$\begin{aligned}\mathrm{p}\left(w\leq\lambda\right)=&\frac{2}{\mathcal{S}_{L-1}}\int_{\phi_1=\arccos(\lambda)}^{\pi/2}\int_{\phi_2\ldots\phi_{L-2}=0}^{\pi}\int_{\phi_{L-1}=0}^{2\pi}\\ &\sin^{L-2}(\phi_1)\ldots\sin(\phi_{L-2})\mathrm{d}\phi_1\ldots\mathrm{d}\phi_{L-1}\\ &=2\frac{\mathcal{S}_{L-2}}{\mathcal{S}_{L-1}}\int_{\phi_1=\arccos(\lambda)}^{\pi/2}\sin^{L-2}(\phi_1)\mathrm{d}\phi_1,\end{aligned}$$

which is only composed of independent sum. By recognizing the area of the $L-2$-sphere and by defining the change of variable $y=\cos(\phi_1)$, one have

$$\mathrm{p}\left(w\leq\lambda\right)=\frac{\mathcal{S}_{L-2}}{\mathcal{S}_{L-1}}\,2\int_0^{\lambda}\sin^{L-3}(\arccos(y))\mathrm{d}y.$$

Knowing that $\sin(\arccos(y))$ can be rewritten as $\sqrt{1-y^2}$, one obtains, after two changes of variable

$$\begin{aligned}\int_0^{\lambda}\sin^{L-3}(\arccos(y))\mathrm{d}y&=\int_0^{\lambda}\left(1-y^2\right)^{L-3}\mathrm{d}y\\ &=\lambda\int_0^{1}\left(1-\lambda^2y^2\right)^{L-3}\mathrm{d}y\\ &=\frac{\lambda}{2}\int_0^{1}\left(1-\lambda^2z\right)^{L-3}z^{-1/2}\mathrm{d}z.\end{aligned}$$

The sum can be resolved using Corollary 1.6.3.2 page 36 in Gupta and Nagar (1999) with parameters $\alpha = \frac{1}{2}$, $\beta = -\frac{L-3}{2}$, $\gamma = \frac{3}{2}$ and $R = \lambda^2$, leading to

$$\int_0^\lambda \sin^{L-3}(\arccos(y))\mathrm{d}y = 2\lambda \; {}_2F_1\left(\frac{1}{2}, -\frac{L-3}{2}; \frac{3}{2}; \lambda^2\right),$$

which is the expected result.

F Inconsistency of the marginal MAP estimator of the latent dimension

We emphasize that the proof is conducted with arguments similar to the one in Miller and Harrison (2013).

Let first introduce a few notations. We call $\mathcal{A}(K, N)$ the set all binary matrices $\mathbf{Z}$ with K rows and N columns. For every binary matrix $\mathbf{Z}$, we call $\mathcal{B}(\mathbf{Z})$ the set of matrices $\mathbf{Z}'$ which are identical to $\mathbf{Z}$ except that a new line have been added with only one active element. The notation $\mathbf{Z}'(j)$ will seldom be employed, where j indicates the index of the new active element. Finally, let $c_N(K, \alpha)$ be the quantity

$$c_N(K,\alpha) \triangleq \max_{\mathbf{Z}\in\mathcal{A}(K,N)} \; \max_{\mathbf{Z}'\in\mathcal{B}(\mathbf{Z})} \; \frac{\mathrm{P}[\mathbf{Z}|\alpha]}{\mathrm{P}[\mathbf{Z}'|\alpha]}. \tag{50}$$

F.1 Two lemmas

Let first consider the two following lemmas

Lemma 1 *For all α, K*

$$\limsup_{N\to+\infty} \; \frac{1}{N} c_N(K,\alpha) \le +\infty. \tag{51}$$

Proof Let N, K be two positive integers, $\mathbf{Z}, \mathbf{Z}'$ two binary matrices belonging respectively to $\mathcal{A}(K, N)$ and $\mathcal{B}(\mathbf{Z})$.

According to Eq. (7), one have, by noting K^h_{new} the number of column in $\mathbf{Z}'$ identical to the added one,

$$\frac{\mathrm{P}[\mathbf{Z}|\alpha]}{\mathrm{P}[\mathbf{Z}'|\alpha]} \le \frac{N}{\alpha} K^h_{\text{new}} \le \frac{K}{\alpha} N,$$

which lead to the expected result. □

Lemma 2 *Let $\mathbf{Z}, \mathbf{Z}'$ be respectively two elements of $\mathcal{A}(K, N)$ and $\mathcal{B}(\mathbf{Z})$. Thus,*

$$\mathrm{p}\left(Y_{1:N} \mid \mathbf{Z}\right) \le \kappa \, \mathrm{p}\left(Y_{1:N} \mid \mathbf{Z}'\right), \tag{52}$$

where

$$\kappa = b_\delta \frac{\gamma(a_\delta, b_\delta)}{\gamma(a_\delta + 1, b_\delta)}. \tag{53}$$

Proof Let Θ be the set of all parameters and hyperperameters, such that

$$\mathrm{p}(\mathbf{Y}|\mathbf{Z}) = \int_\Theta \mathrm{p}(\mathbf{Y}|\theta, \mathbf{Z})\mathrm{p}(\theta|\mathbf{Z})\mathrm{d}\theta.$$

Let $\mathbf{Z}'$ be an element of $\mathcal{B}(\mathbf{Z})$, and j be the index of the active element in the new line. Note the activation of the j^{th} element adds a term of the form

$$\frac{1}{1+\delta^2_{K+1}} \exp\left(\frac{\delta^2_{K+1}}{1+\delta^2_{K+1}} \frac{\left(\mathbf{y}_j^T \mathbf{p}_{K+1}\right)^2}{\sigma^2}\right). \tag{54}$$

The term in the exponential is always positive, so the exponential can be minored by 1. By integrating w.r.t. δ_{K+1}^2, one has

$$\mathrm{p}\left(Y_{1:N} \mid \mathbf{Z}'\right) \geq \frac{b_\delta^{a_\delta}}{\gamma(a_\delta, b_\delta)} \frac{\gamma(a_\delta+1, b_\delta)}{b_\delta^{a_\delta+1}} \mathrm{p}\left(Y_{1:N} \mid \mathbf{Z}\right),$$

which completes the proof. □

F.2 proof

For all integer j in $[\![1, N]\!]$

$$\begin{aligned} &\mathrm{p}\big(\mathbf{Y}, K_N = K|\alpha\big) \\ &\quad= \sum_{\mathbf{Z}_K \in \mathcal{A}(K,N)} \mathrm{P}\left[\mathbf{Z}_K\right] \mathrm{p}(\mathbf{Y}|\mathbf{Z}_K, \alpha) \\ &\quad\leq \sum_{\mathbf{Z}_K \in \mathcal{A}(K,N)} N c_N(K,\alpha)\, \mathrm{P}\left[\mathbf{Z}'(j) \mid \alpha\right] \kappa\, \mathrm{p}\left(\mathbf{Y} \mid \mathbf{Z}'(j), \alpha\right). \end{aligned} \tag{55}$$

where the last inequality has been obtained using both Lemmas 1 and 2. Since this inequality is true for all j, one can average over all values of j, leading to

$$\begin{aligned} &\mathrm{p}\big(\mathbf{Y}, K_N = K|\alpha\big) \\ &\leq \sum_{\mathbf{Z}_K \in \mathcal{A}(K,N)} \sum_{j=1}^{N} \kappa c_N(K,\alpha)\, \mathrm{P}\left[\mathbf{Z}'(j) \mid \alpha\right] \;\; \mathrm{p}\left(\mathbf{Y} \mid \mathbf{Z}'(j), \alpha\right) \\ &\leq \kappa c_N(K,\alpha) \sum_{\mathbf{Z}_K \in \mathcal{A}(K,N)} \sum_{\mathbf{Z}' \in \mathcal{A}(K+1,\alpha)} \mathrm{p}\left(\mathbf{Y}|\mathbf{Z}'(j)|\alpha\right) \mathbf{1}_{\mathbf{Z}' \in \mathcal{B}(\mathbf{Z})} \\ &\leq \kappa c_N(K,\alpha) \sum_{\mathbf{Z}' \in \mathcal{A}(K+1,\alpha)} \mathrm{card}\Big\{\mathbf{Z}, \mathbf{Z}' \in \mathcal{B}(\mathbf{Z})\Big\} \mathrm{p}\left(\mathbf{Y}|\mathbf{Z}'(j)|\alpha\right) \mathbf{1}_{\mathbf{Z}' \in \mathcal{B}(\mathbf{Z})}. \end{aligned}$$

However, for each matrix $\mathbf{Z}'$ in $\mathcal{A}(K+1,\alpha)$, there are at most one matrix $\mathbf{Z}$ verifying the condition, leading to

$$\begin{aligned} &\mathrm{p}\big(\mathbf{Y}, K_N = K|\alpha\big) \\ &\leq \kappa c_N(K,\alpha) \sum_{\mathbf{Z}' \in \mathcal{A}(K+1,\alpha)} \mathrm{p}\left(\mathbf{Y}|\mathbf{Z}'(j)|\alpha\right) \mathbf{1}_{\mathbf{Z}' \in \mathcal{B}(\mathbf{Z})}. \end{aligned} \tag{56}$$

From now, the proof is almost finished. By the Bayes rule, one has for $K < D$

$$\begin{aligned} &\mathrm{p}\Big(K_N = K|\mathbf{Y}, \alpha\Big) \\ &\quad= \frac{\mathrm{p}\left(K_N = K, \mathbf{Y} \mid \alpha\right)}{\sum_{k=0}^{\infty} \mathrm{p}\left(K_N = K, \mathbf{Y}, \alpha\right)} \\ &\quad< \frac{\mathrm{p}\left(K_N = K, \mathbf{Y}|\alpha\right)}{\mathrm{p}\left(K_N = K, \mathbf{Y}, \alpha\right) + \mathrm{p}\left(K_N = K+1|\mathbf{Y}, \alpha\right)} \\ &\quad< \frac{c_N(K,\alpha)\kappa}{c_N(K,\alpha)\kappa + 1} \\ &\quad< 1. \end{aligned}$$

finally, for $K = D$

$$\begin{aligned}
\mathrm{p}\Big(K_N =& D|\mathbf{Y},\alpha\Big)\\
&= \frac{\mathrm{p}\left(K_N = D, \mathbf{Y}|\alpha\right)}{\sum_{k=0}^{\infty} \mathrm{p}\left(K_N = k \mid \mathbf{Y},\alpha\right)}\\
&\geq \frac{\mathrm{p}\left(K_N = D, \mathbf{Y}|\alpha\right)}{\sum_{k=0}^{D} \left(c_N(k,\alpha)\kappa\right)^{K-k} \mathrm{p}\left(K_N = k|\mathbf{Y},\alpha\right)}\\
&\geq \frac{1}{\sum_{k=0}^{D} \left(c_N(k,\alpha)\kappa\right)^{D-k}}\\
&\geq \frac{1}{1+\sum_{k=1}^{D} \left(c_N(K,\alpha)\kappa\right)^{K}}\\
&> 0.
\end{aligned}$$

One can see from the last couple of equations that the result stated in Eq. (27) can be generalized to all models based on an IBP and verifying Lemma 2. However, the result in Eq.(28) results from the orthogonality constraints.

G Severe inconsistency in case of a simple generative model

Assumes that for all n, $\mathbf{y_n} \sim \mathcal{N}(0,\sigma^2\mathbb{I}_D)$ and g be the quantity

$$\begin{aligned}
g(\mathbf{Y},\mathbf{Z},\mathbf{P},\delta^2) = \frac{\mathcal{K}(a_{\delta^2},b_{\delta^2})^K}{\mathrm{vol}(\mathcal{S}_D)} \prod_{k=1}^{K} \left(\frac{1}{1+\delta_k^2}\right)^{a_{\delta^2}+\boldsymbol{z}_k^T\boldsymbol{z}_k}\\
\exp\left[-\frac{1}{1+\delta_k^2}\left(b_{\delta^2}+\frac{1}{2\sigma^2}\sum_{n=1}^{N} z_{k,n}\left(\mathbf{p}_k^T\mathbf{y}_n\right)^2\right)\right],
\end{aligned}$$

i.e., $g \propto \mathrm{p}\left(\mathbf{Z},\mathbf{P},\delta^2|\mathbf{Y},\sigma^2,\alpha\right)$. Let emphasize that g is intimately linked to a probability distribution.

Let K_N be again the random variable associated to the latent subspace dimension. One has, by definition

$$\begin{aligned}
\mathrm{P}\left[K_N = 0|\mathbf{Y},\sigma^2,\alpha\right] &= \frac{\mathrm{p}\left(K_N = 0,\mathbf{Y}|\sigma^2,\alpha\right)}{\sum_{K=1}^{+\infty} \mathrm{p}\left(K_N = K,\mathbf{Y}|\sigma^2,\alpha\right)}\\
&\leq \frac{1}{1+\frac{\mathrm{p}\left(K_N=1,\mathbf{Y}|\sigma^2,\alpha\right)}{\mathrm{p}\left(K_N=0,\mathbf{Y}|\sigma^2,\alpha\right)}}.
\end{aligned} \tag{57}$$

The quantity appearing in the denominator of Eq. (57) can be rewritten

$$\begin{aligned}
&\frac{\mathrm{p}\left(K_N = 1,\mathbf{Y}|\sigma^2,\alpha\right)}{\mathrm{p}\left(K_N = 0,\mathbf{Y}|\sigma^2,\alpha\right)}\\
&= \sum_{\mathbf{Z},K_N=1} \int_{\mathcal{S}_D}\int_{\mathbb{R}_+} g(\mathbf{Y},\mathbf{Z},\delta^2,\mathbf{P})\mathrm{d}\sigma^2\mathrm{d}\mathbf{P}\mathrm{d}\delta^2\ \frac{\mathrm{P}[\mathbf{Z}|\alpha]}{\mathrm{P}[\mathbf{0}|\alpha]}\\
&= \sum_{\mathbf{Z},K_N=1} \int_{\mathcal{S}_D}\int_{\mathbb{R}_+} g(\mathbf{Y},\mathbf{Z},\delta^2,\mathbf{P})\mathrm{d}\sigma^2\mathrm{d}\mathbf{P}\mathrm{d}\delta^2\\
&\times\ \alpha\frac{(N-\mathbf{z}_1^t\mathbf{z}_1)!(\mathbf{z}_1^t\mathbf{z}_1-1)!}{N!}.
\end{aligned}$$

Since the matrix $\mathbf{Z}$ appearing in the former equation has only one row, one can decompose the sum over the number of active component and the number of instance,

$$\frac{\mathrm{p}\left(K_N = 1, \mathbf{Y}|\sigma^2, \alpha\right)}{\mathrm{p}\left(K_N = 0, \mathbf{Y}|\sigma^2, \alpha\right)} =$$
$$\sum_{l=1}^{N} \sum_{\mathbf{Z}, K_N=1, \mathbf{z}_1\mathbf{z}_1^T=l} \frac{\alpha}{l} \frac{1}{\binom{N}{l}} \int_{\mathcal{S}_D} \int_{\mathbb{R}_+} g(\mathbf{Y}, \mathbf{Z}, \delta^2, \mathbf{P}) \mathrm{d}\sigma^2 \mathrm{d}\mathbf{P} \mathrm{d}\delta^2.$$

Let define for each l the U-statistic

$$U_l(\boldsymbol{Y}) \triangleq \frac{1}{\binom{N}{l}} \sum_{\substack{\mathbf{Z}, K_N=1, \\ \mathbf{z}_1\mathbf{z}_1^T=l}} \int_{\mathcal{S}_D \cup \mathbb{R}_+} g(\mathbf{Y}, \mathbf{Z}, \delta^2, \mathbf{P}) \mathrm{d}\sigma^2 \mathrm{d}\mathbf{P} \mathrm{d}\delta^2, \tag{58}$$

where the support of each permutation is given by the l active components of $\boldsymbol{Z}$. By the strong law of large number (Hoeffding 1961), for all l,

$$U_l(\boldsymbol{Y}) \underset{N \to +\infty}{\overset{a.s.}{\longrightarrow}} \mathbb{E}_{\boldsymbol{Y}} \left[\int_{\mathcal{S}_D \cup \mathbb{R}_+} g(\mathbf{Y}, \mathbf{Z}, \delta^2, \mathbf{P}) \mathrm{d}\sigma^2 \mathrm{d}\mathbf{P} \mathrm{d}\delta^2 \right] = 1. \tag{59}$$

The former equality holds since the quantity under the expectation is a density. Consequently, for all $L \leq N$

$$\frac{\mathrm{p}\left(K_N = 1, \mathbf{Y}|\sigma^2, \alpha\right)}{\mathrm{p}\left(K_N = 0, \mathbf{Y}|\sigma^2, \alpha\right)} \geq \sum_{l=1}^{L} \frac{\alpha}{l} U_l(\boldsymbol{Y}) \underset{N \to +\infty}{\overset{a.s.}{\longrightarrow}} \sum_{l=1}^{L} \frac{\alpha}{l}.$$

Since the former equality is true for all L, and that the harmonic series $\sum_l \frac{1}{l}$ diverges, the quantity $\frac{\mathrm{p}\left(K_N=1, \mathbf{Y}|\sigma^2, \alpha\right)}{\mathrm{p}\left(K_N=0, \mathbf{Y}|\sigma^2, \alpha\right)}$ goes to infinity almost surely as N increases. This complete the proof.

H Marginal posterior distribution of the scale parameters

In the general case, the posterior distribution of the scale parameters $\boldsymbol{\delta} = \left\{\delta_1^2, \ldots, \delta_K^2\right\}$, where the orthogonal matrix $\mathbf{P}$ has been marginalized, cannot be derived analytically. However, assuming that the binary matrix $\mathbf{Z}$ is the $K \times N$ matrix $\mathbf{1}_{K,N}$ with only 1's everywhere, this posterior distribution can be derived explicitly. In particular, when $K = D$

$$\begin{aligned} & f\left(\boldsymbol{\delta}|\mathbf{Y}, \sigma^2, \alpha, \mathbf{Z} = \mathbf{1}_{D,D}\right) \propto \\ & \prod_{k=1}^{D} \left(\frac{1}{1+\delta_k^2}\right)^{a_\delta + 1} \exp\left(-\frac{b_\delta}{1+\delta_k^2}\right) \\ & \times {}_0\mathrm{F}_0\left(\emptyset, \emptyset, \frac{1}{\sigma^2}\mathbf{Y}\mathbf{Y}^T - \lambda \mathbb{I}_D, \boldsymbol{\Delta}_{\boldsymbol{\delta}}\right) \mathrm{etr}\left(\lambda \boldsymbol{\Delta}_{\boldsymbol{\delta}}\right) \end{aligned} \tag{60}$$

with $\lambda \in (0, \frac{1}{\sigma^2}\rho_{\min})$ where[3] $\rho_{\min}$ is the minimum eigenvalue of $\mathbf{Y}\mathbf{Y}^T$, $\boldsymbol{\Delta}_{\boldsymbol{\delta}}$ is a $D \times D$ diagonal matrix formed by the ratios $\delta_k^2/(1+\delta_k^2)$ and ${}_0\mathrm{F}_0$ is a generalized hypergeometric function of two matrices. In particular, this function is defined by

$${}_0\mathrm{F}_0(\emptyset, \emptyset, \mathbf{A}, \mathbf{B}) = \sum_{k=1}^{\infty} \sum_{\kappa \vdash k} \frac{C_\kappa(\mathbf{A}) C_\kappa(\mathbf{B})}{C_\kappa(\mathbb{I}_D) k!} \tag{61}$$

[3] Note that the positive real number λ has no particular interpretation and is only introduced here for convenience.

where $\kappa \vdash k$ denotes the integer partitions of k, $C_\kappa(\mathbf{A})$ is a zonal polynomial defined by the eigenvalues of $\mathbf{A}$ (Muirhead 1982, Ch. 7). Despite recent advances in numerical evaluation of zonal polynomials due to, e.g., Koev and Edelman (2006), this quantity remains difficult to be computed. However, it can be interpreted as a measure of mismatch between the magnitudes of the principal components recovered by PCA (through the eigenvalues of $\frac{1}{\sigma^2}\mathbf{Y}\mathbf{Y}^T - \lambda\mathbb{I}_D$) and the magnitudes of the relevant components identified by the proposed procedure (in $\boldsymbol{\Delta_\delta}$).

More generally, this hypergeometric function can be advocated for as an elegant way to compare two positive definite matrices using their respective eigenvalues. This finding would suggest the design of an appropriate metric which allows two covariance matrices to be compared regardless of their respective induced orientations.

References

Archambeau C, Delannay N, Verleysen M (2008) Mixtures of robust probabilistic principal component analyzers. Neurocomputing 71(7-9):1274 – 1282, progress in Modeling, Theory, and Application of Computational Intelligenc15th European Symposium on Artificial Neural Networks 200715th European Symposium on Artificial Neural Networks 2007 2

Besson O, Dobigeon N, Tourneret JY (2011) Minimum mean square distance estimation of a subspace. IEEE Trans Signal Process 59(12):5709–5720 3, 17

Besson O, Dobigeon N, Tourneret JY (2012) CS decomposition based Bayesian subspace estimation. IEEE Trans Signal Process 60(8):4210–4218 17

Bioucas-Dias J, Nascimento J (2008) Hyperspectral subspace identification. Geoscience and Remote Sensing, IEEE Transactions on 46(8):2435–2445 25

Bioucas-Dias JM, Nascimento JMP (2008) Hyperspectral subspace identification. IEEE Trans Geosci and Remote Sens 46(8):2435–2445 25

Bioucas-Dias JM, Plaza A, Dobigeon N, Parente M, Du Q, Gader P, Chanussot J (2012) Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. IEEE J Sel Topics Appl Earth Observations and Remote Sens 5(2):354–379 25

Bishop CM (1999a) Bayesian pca. In: Kearns MJ, Solla SA, Cohn DA (eds) Advances in Neural Information Processing Systems 11, MIT Press, pp 382–388 2

Bishop CM (1999b) Variational principal components. In: Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99, IEE, vol 1, pp 509–514 2

Blanchet G, Moisan L (2012) An explicit sharpness index related to global phase coherence. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1065–1068 26

Bolton RJ, Hand DJ, Webb AR (2003) Projection techniques for nonlinear principal component analysis. Statistics and Computing 13(3):267–276 2

Broderick T, Jordan MI, Pitman J (2013) Cluster and feature modeling from combinatorial stochastic processes. Statist Sci 28(3):289–312 5

Chen M, Gao C, Zhao H (2016) Posterior contraction rates of the phylogenetic indian buffet processes. Bayesian Anal 11(2):477–497 12, 13

Clark RN, Swayze GA, Gallagher A (1993) Mapping minerals with imaging spectroscopy. US Geological Survey, Office of Mineral Resources Bulletin 2039:141–150 26

Clark RN, Swayze GA, Livo KE, Kokaly RF, Sutley SJ, Dalton JB, McDougal RR, Gent CA (2003) Imaging spectroscopy: Earth and planetary remote sensing with the USGS Tetracorder and expert systems. J Geophys Res 108(E12):5–1–5–44 26

D A van Dyk, Park T (2008) Partially collapsed Gibbs samplers: Theory and methods. J Amer Stat Assoc 103(482):790–796 8

Elvira C, Chainais P, Dobigeon N (2017) Bayesian nonparametric subspace estimation. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 2247–2251, DOI 10.1109/ICASSP.2017.7952556 3

Ghosal S (2009) The dirichlet process, related priors and posterior asymptotics. In: Hjort NL, Holmes C, Muller P, Walker SG (eds) Bayesian Nonparametrics:, Cambridge University Press, Cambridge, pp 35–79 13

Ghosal S, Ghosh JK, Ramamoorthi RV (1999) Posterior consistency of dirichlet mixtures in density estimation. Ann Statist 27(1):143–158 13

Godsill S (2010) The shifted inverse-gamma model for noise-floor estimation in archived audio recordings. Signal Processing 90(4):991–999 7

Green PJ (1995) Reversible jump Markov Chain Monte Carlo methods computation and Bayesian model determination. Biometrika 82(4):711–732 3

Griffiths TL, Ghahramani Z (2011) The indian buffet process: An introduction and review. J Mach Learning Research 12:1185–1224 5

Gupta A, Nagar (1999) Matrix Variate Distributions, 1st edn. Monographs and Surveys in Pure and Applied Mathematics, Chapman and Hall/CRC 31

Herz CS (1955) Bessel functions of matrix argument. The Annals of Mathematics 61(3):474 4

Hoeffding W (1961) The strong law of large numbers for u-statistics. Institute of Statistics mimeo series 302 34

Hoff P (2009) Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data. J Comput and Graph Stat 18(2):438–456 11

Jolliffe IT (1986) Principal Component Analysis. Springer-Verlag, New York 2

Knowles D, Ghahramani Z (2011) Nonparametric Bayesian sparse factor models with application to gene expression modeling. Ann Appl Stat 5(2B):1534–1552 8

Koev P, Edelman A (2006) The efficient evaluation of the hypergeometric function of a matrix argument. Mathematics of Computation 75(254):833–847 35

Lawrence N (2005) Probabilistic non-linear principal component analysis with gaussian process latent variable models. J Mach Learn Res 6:1783–1816 2

Lian H (2009) Bayesian nonlinear principal component analysis using random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(4):749–754 2

MacKay D (1995) Ensemble learning and evidence maximization. Tech. rep., Adv. in Neural Information Processing Systems (NIPS) 2

Marin JM, Robert CP (2007) Bayesian Core: A Practical Approach to Computational Bayesian Statistics. Springer, New York, NY, USA 24, 29

McCullagh P, Yang J (2008) How many clusters? Bayesian Anal 3(1):101–120 13

Miller JW, Harrison MT (2013) A simple example of dirichlet process mixture inconsistency for the number of components. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in Neural Information Processing Systems 26, Curran Associates, Inc., pp 199–206 13, 31

Miller JW, Harrison MT (2014) Inconsistency of pitman-yor process mixtures for the number of components. J Mach Learn Res 15(1):3333–3370 13

Minka TP (2000) Automatic choice of dimensionality for PCA. In: Adv. in Neural Information Processing Systems (NIPS), vol 13, p 514 3, 25

Muirhead RJ (1982) Aspects of multivariate statistical theory. Wiley series in probability and mathematical statistics. Probability and mathematical statistics, Wiley 35

Müller P, Mitra R (2013) Bayesian nonparametric inference – why and how. Bayesian Anal 8(2):269–302 3

Punskaya E, Andrieu C, Doucet A, Fitzgerald W (2002) Bayesian curve fitting using MCMC with applications to signal segmentation. IEEE Trans Signal Process 50(3):747–758 6, 7

Robert CP (2007) The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation, 2nd edn. Springer Texts in Statistics, Springer-Verlag, New York 7

Schmitt E, Vakili K (2016) The fasthcs algorithm for robust pca. Statistics and Computing 26(6):1229–1242 2

Smídl V, Quinn A (2007) On Bayesian principal component analysis. Comput Stat Data Anal 51(9):4101–4123 3, 25

Teh YW, Görür D, Ghahramani Z (2007) Stick-breaking construction for the Indian buffet process. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, vol 11 5

Thibaux R, Thibaux R, Jordan MI (2007) Hierarchical beta processes and the Indian buffet process. In practical nonparametric and semiparametric Bayesian statistics 2007:227–242 5

Tipping ME, Bishop CM (1999a) Mixtures of probabilistic principal component analyzers. Neural Comput 11(2):443–482 2

Tipping ME, Bishop CM (1999b) Probabilistic principal component analysis. J Roy Stat Soc Ser B 61(3):611–622 2

Zhang Z, Zhang Z, Chan KL, Kwok JT, Yeung Dy (2004) Bayesian inference on principal component analysis using reversible jump markov chain monte carlo. Proceedings of the Nineteenth National Conference on Artificial Intelligence 3

Acknowledgements Part of this work has been funded thanks to the BNPSI ANR project no. ANR-13-BS-03-0006-01.