BAYESIAN NONPARAMETRIC SUBSPACE ESTIMATION

Clément Elvira[(1)]*, Pierre Chainais*[(1)] *and Nicolas Dobigeon*[(2)]

[(1)] Univ. Lille, CNRS, Centrale Lille, CRIStAL, Lille, France
[(2)] Univ. Toulouse, IRIT/INP-ENSEEIHT, Toulouse, France

ABSTRACT

Principal component analysis is a widely used technique to perform dimension reduction. However, selecting a finite number of significant components is essential and remains a crucial issue. Only few attempts have proposed a probabilistic approach to derive a posterior distribution of this number of significant components. This paper introduces a Bayesian nonparametric model to jointly estimate the principal components and the corresponding intrinsic dimension. More precisely, the observations are projected onto a random orthogonal basis which is assigned a prior distribution defined on the Stiefel manifold. Then the factor scores take benefit of an Indian buffet process prior 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. The performances of the proposed approach are assessed thanks to experiments conducted on various examples.

Index Terms— Bayesian inference, dimension reduction, distribution on the Stiefel manifold, Indian buffet process.

1. INTRODUCTION

Principal component analysis (PCA) is an ubiquitous tool in signal processing and statistical data analysis. It implicitly permits a dimension reduction by projecting observations onto a subset of orthonormal vectors referred to as principal components. In its most widely admitted formulation, PCA does not derive from a probabilistic model. This can be an issue when the relevance of the selected principal components needs to be assessed. To fill this gap, Tipping and Bishop have demonstrated in [1] how PCA can be interpreted as a maximum likelihood estimator of a latent factor model, where both noise, factor and coefficients are assumed to be Gaussian distributed. The subspace to be recovered was finally inferred using an expectation-maximization (EM) algorithm, leading to the so-called probabilistic PCA (PPCA). Such an approach allows PCA to be performed while facing with missing data [1] and also can be extended to handle mixture of PCA [2]. However, selecting the optimal number of degrees of freedom has not be considered by these methods.

An inappropriate selection of the relevant subset of principal components may lead to the loss of significant information, misinterpretation. Thus, this selection is a crucial issue when resorting to PCA for dimension reduction. A few approaches have been proposed to tackle this problem in a probabilistic framework. One strategy consists in modeling principal components as random orthonormal vectors distributed over the Stiefel manifold, i.e., the set of (tall) matrices with orthonormal columns. In [3], Minka has extended PPCA by designing a prior distribution of the latent factors associated with the singular value decomposition of covariance matrices, which allows the posterior distribution over the subspace dimension to be approximated. In [4], a Stiefel manifold-based prior combined to a prior on the number of selected components leads to a variational approximation of the posterior.

In this work, we propose to investigate the use of Bayesian nonparametric inference to explore the set of subspaces and derive a posterior distribution of the intrinsic data dimension. To this aim, inspired by [5], the prior on the principal components is elected as a uniform distribution over the Stiefel manifold in dimension D. Moreover, an Indian buffet process (IBP) prior [6] is assigned to the representation coefficients to promote a parsimonious use of the corresponding components, implicitly performing a dimension reduction. The IBP can be interpreted as a distribution on the set of potentially infinite binary matrices that penalizes large matrices. Our purpose is to estimate an orthonormal basis of a subspace of dimension $K \leq D$ which is not a priori fixed but this relevant reduced dimension will be rather automatically inferred. A Markov chain Monte Carlo (MCMC) sampler is derived to approximate the posterior distribution and compute estimates of an orthonormal basis of a subspace of size $K \leq D$. The number K of selected components and their corresponding eigenvalues (energies) are inferred as well.

The sequel of this paper is organized as follows. Section 2 describes the proposed hierarchical Bayesian model. Section 3 describes the MCMC inference scheme. Section 4 illustrates the performances of the method on numerical examples. Concluding remarks are finally reported in Section 5.

2. HIERARCHICAL BAYESIAN MODEL

2.1. Representation model

Let $\mathbf{y}_n = [y_{1,n} \dots y_{D,n}]^T$ denotes a D-dimensional observation vector. The set of N observed vectors $\mathbf{y}_1, \dots, \mathbf{y}_N$ are expected to live in a K-dimensional subspace with $K \leq D$. The problem addressed in this work consists of identifying this subspace and, most importantly, the intrinsic dimension of this subspace. To do so, the observation vectors are assumed to be represented according to the following latent factor model

$$\mathbf{y}_n = \boldsymbol{P}(\mathbf{z}_n \odot \mathbf{x}_n) + \mathbf{e}_n \tag{1}$$

where $\boldsymbol{P}$ is an orthonormal basis of $\mathbb{R}^D$, i.e., $\boldsymbol{P}^T\boldsymbol{P} = \mathbf{I}$ is the identity matrix, $\mathbf{z}_n$ is a D-dimensional binary vector, $\mathbf{x}_n$ is a vector of coefficients and $\odot$ denotes the Hadamard (term-wise) product. In (1), the additive term $\mathbf{e}_n$ stands for a white Gaussian mismodeling and/or observation noise. It is worth noting that the binary coefficients $\mathbf{z}_n$ encodes the activation hence the relevance of the corresponding coefficients in $\mathbf{x}_n$ for the latent representation. Thus, the term-wise product vector $\mathbf{z}_n \odot \mathbf{x}_n$ would be referred to as the factor scores in the PCA terminology.

Part of this work has been supported by the BNPSI ANR project no. ANR-13-BS-03-0006-01.

2.2. Likelihood function

Since the noise is assumed to be a white Gaussian noise independent, i.e., $\mathbf{e}_n|\sigma^2\mathcal{N}\left(\mathbf{0},\sigma^2\mathbf{I}\right)$, the likelihood of a set of N observations assumed to be a priori independent can be written as

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

$\mathbf{Y}$ is the $D\times N$ matrix resulting from the concatenation of all observation vectors, $\mathbf{Z}$ is the binary activation matrix, $\mathbf{X}$ is the matrix of representation coefficients and $\|\boldsymbol{u}\|_2$ is the ℓ_2-norm of $\boldsymbol{u}$.

2.3. Prior distributions

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

Orthonormal basis $\boldsymbol{P}$. By definition, $\boldsymbol{P}$ is an orthonormal base and belongs to the unitary group $\mathbb{U}_D$. Since no information is available a priori about any preferred direction,a uniform distribution on $\mathbb{U}_D$ is chosen as a prior distribution on $\boldsymbol{P}$ whose pdf with respect to the Lebesgue measure is

$$f(\mathbf{P})=\frac{1}{\mathrm{Vol}(\mathbb{U}_D)}\mathbb{1}_{\mathbb{U}_D}(\boldsymbol{P}) \quad (3)$$

where the volume of the unitary group is

$$\mathrm{vol}(\mathbb{U}_D)=\frac{2^D\pi^{\frac{D^2}{2}}}{\pi^{\frac{1}{4}D(D-1)}\prod_{i=1}^{D}\Gamma\left(\frac{D}{2}-\frac{i-1}{2}\right)} \quad (4)$$

and $\mathbb{1}_{\mathbb{A}}(\cdot)$ denotes the indicator function on the set $\mathbb{A}$. Note that a subset of K columns of $\boldsymbol{P}\in\mathbb{U}_D$ belongs to the Stiefel manifold $\mathcal{S}_D^K$, the set of matrices with K orthonormal columns in dimension D (see [7] for a review of statistics on the Stiefel Manifold and corresponding sampling methods).

Indian buffet process $\mathbf{Z}$. Since the observation vectors are assumed to live in lower dimensional subspace, most of the factor scores in the vectors $\mathbf{z}_n\odot\mathbf{x}_n$ are expected to be null. To reflect this key feature, an IBP prior is assigned to the binary latent factor activation coefficient [8]

$$\mathbf{Z}|\alpha\sim\mathrm{IBP}(\alpha). \quad (5)$$

where α controls the underlying sparsity of $\mathbf{Z}$. This prior promotes a parsimonious use of the corresponding principal components $\mathbf{p}_k$. Above all, it allows the relevant subspace dimension K to be not a priori fixed and simultaneously penalizes large values of K, implicitly performing dimension reduction. Indeed, the resulting prior mean of non-zero components behaves as $\alpha\log N$. The regularization effect of the IBP is complemented by the orthogonality constraint imposed on $\boldsymbol{P}$, which also prevents any value of K to be greater than D. More specifically, the number of non-zero lines in $\mathbf{Z}$ will determine the number K of active components in $\boldsymbol{P}$ to describe observations according to the model (1).

In brief, the Indian Buffet Process (IBP) can be interpreted as a distribution over infinite binary matrices [6, 8]. It can be used to derive latent feature models where the number of features is a priori unknown. The following culinary metaphor is often employed to describe the IBP. Let consider a buffet with an infinite number of dishes. A first customer (an observation) enters into the restaurant and chooses $K_1\sim\mathcal{P}(\alpha)$ dishes (features). The next customer chooses each of these dishes with probability $m_1/2$ and then tries a Poisson random number $\mathcal{P}(\frac{\alpha}{2})$ of new dishes. Then the nth customer tries each of these dishes with probability $\frac{m_k}{n}$, where m_k is the number of times dish k has been already chosen by previous customers; then he tries $K_n\sim\mathcal{P}(\frac{\alpha}{n})$ new dishes. As a constructive consequence, some dishes are very often selected while many others are rarely chosen.

Coefficients $\mathbf{X}$. Independent Gaussian prior distributions are assigned to the individual representation coefficient 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 to make later marginalization tractable analytically (see next Section). Since the relevant quantity here is rather the ratio between the energy of each component and the noise variance, we follow the recommendation in [9] to set the variance of coefficients as a multiple of the noise variance through a Zellner's prior. Thus the prior on the coefficient x_k along component $\mathbf{p}_k$ is

$$\forall k\in\mathbb{N},\quad x_k|\delta_k^2,\sigma^2\sim\prod_{n=1}^{N}\mathcal{N}(0,\delta_k^2\sigma^2). \quad (6)$$

where the hyperparameters δ_k^2 will be the parameters of interest corresponding to the ratio between the eigenvalues of a classical PCA and the noise variance.

Noise variance σ^2. A conjugate inverse Gamma prior is assigned to σ^2

$$\sigma^2\sim\mathcal{IG}\left(a_{\sigma^2},b_{\sigma^2}\right) \quad (7)$$

where a_{σ^2} and b_{σ^2} are positive hyperparameters chosen to design a vague prior. Note that the specific choice $a_{\sigma^2}=b_{\sigma^2}=0$ would lead to a noninformative Jeffreys prior as in [9, 10]. This choice is here prohibited since it would also lead to an improper posterior distribution [11].

Hyperparameters. The set of hyperparameters is gathered in $\boldsymbol{\phi}=\left\{\delta_1^2,\ldots,\delta_k^2,\alpha\right\}$. The IBP parameter α will control the number of active latent factors while each δ_K determines 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 estimate them with the parameters of interest jointly.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. We choose a conjugate shifted Inverse Gamma distribution denoted by s$\mathcal{IG}$. More precisely, since the power of relevant components are expected to be at least of the order of magnitude of the noise variance, the prior distribution is defined over the set $(1,+\infty[$ as

$$\mathrm{p}\left(\delta_k^2|a_\delta,b_\delta\right)=\frac{b_\delta^{a_\delta}}{\gamma\left(a_\delta,0.5b_\delta\right)} \\ \times\left(\frac{1}{1+\delta_k^2}\right)^{a_\delta+1}\exp\left(-\frac{b_\delta}{1+\delta_k^2}\right)\mathbb{1}_{[1,+\infty[}(\delta_k^2) \quad (8)$$

where $\gamma(a,b)$ is the lower incomplete gamma function, a_δ and b_δ are tuned to limit the weight around $\delta_k^2=1$, e.g., $a_\delta=1$, $b_\delta=20$.

IBP parameter α. Without any prior knowledge regarding this hyperparameter, a Jeffreys prior is assigned to α

$$\mathrm{p}(\alpha) \propto \frac{1}{\alpha} \mathbb{1}_{\mathbb{R}_+}(\alpha).$$

3. METROPOLIS-WITHIN-GIBBS SAMPLER

The posterior distribution resulting from the hierarchical Bayesian model described in Section 2 is too complex to derive closed-form expressions of the Bayesian estimators associated with the parameters of interest, namely, the orthonormal matrix $\boldsymbol{P}$ and the binary matrix $\mathbf{Z}$ selecting the relevant components. To overcome this issue, a MCMC algorithm to generate samples asymptotically distributed according to the marginal posterior distributions of these parameters. These samples can be subsequently used to approximate the classical Bayesian estimators, i.e., the minimum mean square error (MMSE) and maximum a posteriori estimators. Note that other suitable Bayesian estimators have been proposed in [5,12] in the specific context of subspace estimation. The proposed MCMC algorithm is described in what follows.

3.1. 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 [13]. Here, benefiting from the conjugacy property induced by the prior in (6), we propose to marginalize over the coefficients $\mathbf{X}$

$$\mathrm{p}\left(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\phi}\right) = \int_{\mathbb{R}^{DN}} \mathrm{p}\left(\boldsymbol{\theta}, \mathbf{X}|\mathbf{Y}, \boldsymbol{\phi}\right) \mathrm{d}\mathbf{X}. \tag{9}$$

This operation goes beyond calculation convenience and leads to the following marginalized posterior distribution

$$\begin{aligned}\mathrm{p}(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{Y}) \; \propto & \left(\frac{1}{\sigma^2}\right)^{\frac{ND}{2}} \exp\left(-\frac{\mathrm{tr}(\mathbf{Y}^T\mathbf{Y})}{2\sigma^2}\right) \\ & \times \prod_{k=1}^{K} \left(\frac{1}{1+\delta_k^2}\right)^{a_\delta + \frac{1}{2}\sum_n z_{k,n}} \exp\left(-\frac{b_\delta}{1+\delta_k^2}\right) \\ & \times \prod_{k=1}^{K} \exp\left[\frac{1}{2\sigma^2}\frac{\delta_k^2}{1+\delta_k^2}\sum_n z_{k,n}\left\langle \mathbf{p}_k, \mathbf{y}_n\right\rangle^2\right] \mathbb{1}_{\mathbb{U}_D}(\mathbf{P}) \\ & \times \frac{\alpha^K}{\prod_k K_n!} e^{\alpha \sum_n \frac{1}{i}} \prod_k \frac{(N-m_k)!\,(m_k-1)!}{N!} \\ & \times (\frac{1}{\sigma^2})^{a_{\sigma^2}+1} e^{-\frac{b_{\sigma^2}}{\sigma^2}} \alpha^{-1}\end{aligned} \tag{10}$$

where $\mathrm{tr}(\cdot)$ denotes the trace operator. This, instead of sampling the joint posterior $\mathrm{p}\left(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\phi}\right)$, a MCMC algorithm is designed to sample according to (10). It consists of a Metropolis-within-Gibbs sampler whose main steps are detailed below.

3.2. Algorithm

***Sampling the binary matrix* $\mathbf{Z}$.** Let $m_{k,-n}$ denote the number of observations different from n which use the latent vector $\mathbf{p}_k$. The update goes in two steps (see [14]). First, observations for which $m_{k,-n} > 0$ are updated through a Gibbs sampling step where the scale coefficient δ_k^2 can be marginalized out. We use a Metropolis-Hastings step to sample observations for which $m_{k,-n} = 0$, also called singletons since $\mathbf{p}_k$ is then used by observation n only, and to sample new components. The move goes from state $\epsilon = \{\kappa, \mathbf{P}_{\mathrm{single}}\}$ to a new state $\epsilon^* = \left\{\kappa^*, \mathbf{P}^*_{\mathrm{single}}\right\}$ where κ is the number of singletons (potentially 0) and $\mathbf{P}_{\mathrm{single}} \triangleq [\tilde{\mathbf{p}}_1, ..., \tilde{\mathbf{p}}_\kappa]$ are the components associated to these singletons. The proposal distribution in the Metropolis-Hastings step is chosen according to the conditional model

$$\mathrm{q}\left(\kappa, \tilde{\mathbf{p}}_1, \ldots \tilde{\mathbf{p}}_\kappa\right) = \mathrm{q}\left(\kappa\right) \mathrm{J}\left(\tilde{\mathbf{p}}_1, \ldots \tilde{\mathbf{p}}_\kappa|\kappa\right) \tag{11}$$

where the following von Mises-Fisher distribution [7] is chosen as a proposal for $\mathbf{P}_{\mathrm{single}}$:

$${}_0F_1(\emptyset, D/2, \frac{1}{4}\mathbf{F}^T\mathbf{F}) \exp\left(\mathrm{tr}(\mathbf{F}^T[\tilde{\mathbf{p}}_1, \ldots \tilde{\mathbf{p}}_\kappa])\right). \tag{12}$$

In (12), the columns of $\mathbf{F}$ are the κ first eigenvectors of $\mathbf{Y}\mathbf{Y}^T$ multiplied by their corresponding eigenvalues and the function ${}_0F_1(\cdot, \cdot, \cdot)$ in the normalizing factor is the confluent hypergeometric function with a matrix argument [15].

Sampling the projection matrix $\boldsymbol{P}$. Let split the orthonormal matrix $\boldsymbol{P}$ into 2 matrices, i.e., $\boldsymbol{P} = [\boldsymbol{P}_K, \bar{\boldsymbol{P}}_K]$, namely the matrix $\boldsymbol{P}_K$ of K active components (used by at least one observed vector) and $\bar{\boldsymbol{P}}_K$, the matrix of unused components. Let $\boldsymbol{P}_{K\backslash k}$ denote the matrix obtained by removing column $\mathbf{p}_k$ from $\boldsymbol{P}_K$ and $\boldsymbol{N}_{K\backslash k}$ a matrix whose $D-K+1$ columns form an orthonormal basis in $\mathbb{U}_D$ for the space spanned by $\bar{\boldsymbol{P}}_K$, i.e., the orthogonal space of $\mathrm{span}\left(\boldsymbol{P}_K\right)$. Then, given $\boldsymbol{P}_{K\backslash k}$, the component $\mathbf{p}_k$ can be written as $\mathbf{p}_k = \boldsymbol{N}_{K\backslash k}\boldsymbol{v}_k$. Since the prior distribution of $\boldsymbol{P}$ is uniform on the unitary group $\mathbb{U}_D$, $\boldsymbol{v}_k$ is uniform on the $(D-K+1)$-dimensional unit sphere [7]. Therefore, by marginalizing $\bar{\boldsymbol{P}}_K$,

$$\begin{aligned}&\mathrm{p}(\boldsymbol{v}_k|\mathbf{Y}, \boldsymbol{P}_{K\backslash k}, \mathbf{Z}, \delta_k^2, \sigma^2) \propto \\ &\exp\left(\frac{1}{2\sigma^2}\frac{\delta_k^2}{1+\delta_k^2}\boldsymbol{v}^T\boldsymbol{N}_{K\backslash k}^T\left(\sum_{n=1}^N z_{k,n}\mathbf{y}_n^T\mathbf{y}_n\right)\boldsymbol{N}_{K\backslash k}\boldsymbol{v}_k\right)\end{aligned} \tag{13}$$

where one recognizes a Bingham distribution on the $D-K+1$ unit sphere, whose sampling can be efficiently conducted [7].

Sampling the scale coefficients δ_k^2. Thanks to the use of a conjugate s$\mathcal{IG}$ shifted inverse Gamma distribution (8) as a prior, the posterior distribution of the scale coefficients corresponding to the K active components is

$$\begin{aligned}&\forall 1 \leq k \leq K, \quad \delta_k^2|\mathbf{Y}, \mathbf{Z}, \boldsymbol{P}, \sigma^2 \sim \\ &s\mathcal{IG}\left(a_\delta + \mathbf{z}_k^T\mathbf{z}_k, b_\delta + \frac{1}{2\sigma^2}\sum_{n=1}^N z_{k,n}\left\langle \mathbf{p}_k, \mathbf{y}_n\right\rangle^2\right).\end{aligned} \tag{14}$$

Sampling the noise variance σ^2. The conditional posterior distribution of the noise variance is an inverse Gamma distribution

$$\begin{aligned}&\sigma^2|\mathbf{Y}, \mathbf{Z}, \boldsymbol{P}, \delta_k^2 \sim \mathcal{IG}\left(a_{\sigma^2} + \frac{DK}{2},\right. \\ &\left. b_{\sigma^2} + \frac{1}{2}\sum_{n=1}^N \mathbf{y}_n^T\mathbf{y}_n - \sum_k \frac{\delta_k^2}{1+\delta_k^2} z_{k,n}\left\langle \mathbf{p}_k, \mathbf{y}_n\right\rangle^2\right).\end{aligned} \tag{15}$$

4. EXPERIMENTAL RESULTS

The performances of the proposed algorithm have been evaluated on a various simulated datasets. These datasets are generated as follows: K orthonormal directions described by a matrix $\mathbf{P}_K$ are uniformly generated on the Stiefel manifold $\mathcal{S}_D^K$; N vectors of coefficients $\mathbf{x}_1 \dots \mathbf{x}_N$ of dimension K are randomly generated according to a centered Gaussian distribution with diagonal covariance matrix $(\delta_1^2\sigma^2, ..., \delta_K^2\sigma^2)$, scaled by a noise variance σ^2. The scale coefficients δ_k^2 are defined as proportional to $1/k$. Finally, N observation vectors are generated according to

$$\mathbf{y}_n = \mathbf{P}\mathbf{x}_n + \boldsymbol{e}_n \tag{16}$$

where $\boldsymbol{e}_n$ is an additive Gaussian noise of covariance matrix $\sigma^2\mathbf{I}$. As an illustration, we only report here results on 2 datasets corresponding to $(D = 16, K = 4, N = 100)$ and $(D = 36, K = 6, N = 500)$. Since all variances are scaled by the noise, only one noise level is considered, $\sigma^2 = 0.01$. For each case, we perform 20 simulations and with 500 Monte Carlo iterations after a burn-in period of 100 iterations.

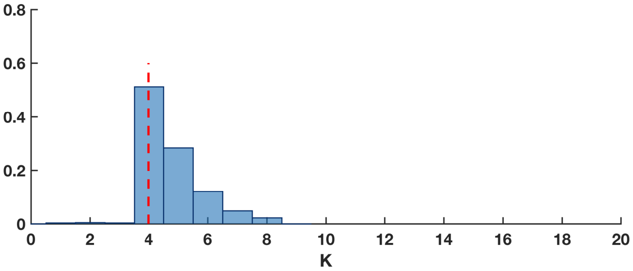

Fig. 1. Posterior distribution of K, for $D = 16$ and $N = 100$.

Fig. 1 shows the posterior distribution of K for $D = 16$ and $N = 100$. Despite a small number of iterations, the maximum of the posterior histogram corresponds to the expected dimension $K = 4$ of the latent subspace. Note that this estimator corresponds to the marginal maximum a posteriori estimator.

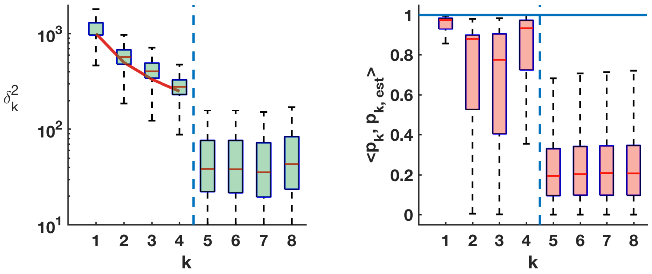

Fig. 2. Posterior distributions of the coefficients δ_K (left) and dispersion of the projection $\mathbf{P}_{\text{est}}\mathbf{P}$ (right), for $D = 16$ and $N = 100$. The red line indicates the trues value of $\delta_1^2 \dots \delta_K^2$.

Fig. 2 shows both the posterior distribution of the 8 first scale coefficients and the alignment of the true $\mathbf{P}$ with the estimated ones. The alignment is measured by the scalar product $\langle \mathbf{p}_k, \hat{\mathbf{p}}_k \rangle$ between each column of $\mathbf{P}$ and its estimate. No ordering problem is expected since the variances are different in all directions. It appears that scale coefficients are correctly inferred. Inactive components $(k \geq 5)$ are associated to coefficients with much lower alignment and are all comparable. Inferred directions correspond to actual principal components with an alignment typically higher than 0.8 in average.

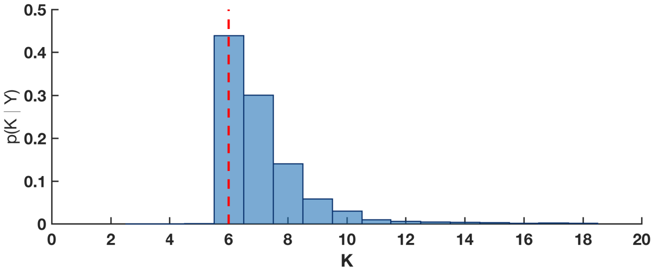

Fig. 3. Posterior distribution of K, for $D = 36$ and $N = 500$.

Fig. 3 shows results obtained on the second dataset with K for $D = 36$, $N = 500$ when $K = 6$. Again the maximum of the posterior distribution corresponds to the expected dimension of the latent subspace. Similar results are obtained from only $N = 100$ observations except that the algorithm inferred a latent subspace of dimension 7.

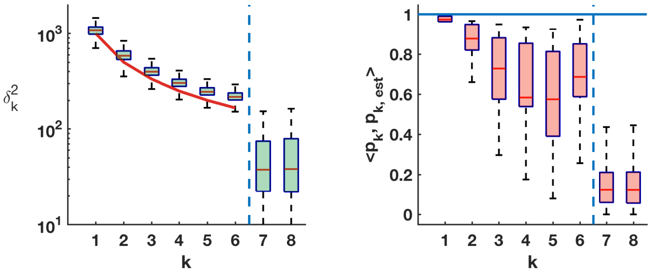

Fig. 4. Posterior distributions of the coefficients δ_K (left) and dispersion of the projection $\mathbf{P}_{\text{est}}\mathbf{P}$ (right), for $D = 36$ and $N = 500$. The red line indicates the trues value of $\delta_1^2 \dots \delta_K^2$.

Fig. 4 shows the first scale coefficients and the alignment results. As in the previous scenario, scale coefficient are correctly inferred. All inferred principal components exhibit a strong alignment with directions used for synthesis. These results own a general scope and have been reproduced in numerous other settings, not reproduced here due to lack of space.

5. CONCLUSION

This paper proposed a new Bayesian nonparametric framework to infer the intrinsic dimension of a set of observations. The model exploited the IBP, a sparse-promoting prior on potentially infinite binary matrices, coupled with a uniform distribution on the set of orthonormal bases. A Metropolis-within-Gibbs sampler was designed to successively sample all parameters according to their conditional posterior distributions. As preliminary results, performances were assessed on two synthetic datasets, demonstrating the efficiency of the proposed method in two scenarios differing by the underlying dimensions of observation space and latent space. Future works will aim at extending the proposed the model towards an application-oriented scheme such as hyperspectral unmixing, where dimension reduction plays a key role. In particular, coupling the dimension reduction step with the unmixing into a fully Bayesian framework may significantly improve the unmixing performances while avoiding painful selection of suitable spectral components.

6. REFERENCES

[1] M. E. Tipping and C. M. Bishop, “Probabilistic principal component analysis,” *J. Roy. Stat. Soc. Ser. B*, vol. 61, no. 3, pp. 611–622, Jan. 1999.

[2] ——, “Mixtures of probabilistic principal component analysers,” *Neural Comput.*, vol. 11, pp. 443–482, 1999.

[3] T. P. Minka, “Automatic choice of dimensionality for PCA,” in *Adv. in Neural Information Processing Systems (NIPS)*, vol. 13, 2000, p. 514.

[4] V. Smídl and A. Quinn, “On Bayesian principal component analysis,” *Comput. Stat. Data Anal.*, vol. 51, no. 9, pp. 4101–4123, May 2007.

[5] O. Besson, N. Dobigeon, and J.-Y. Tourneret, “Minimum mean square distance estimation of a subspace,” *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 5709–5720, Dec. 2011.

[6] T. L. Griffiths and Z. Ghahramani, “The indian buffet process: An introduction and review,” *J. Mach. Learning Research*, vol. 12, pp. 1185–1224, July 2011.

[7] P. Hoff, “Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data,” *J. Comput. and Graph. Stat.*, vol. 18, no. 2, pp. 438–456, 2009.

[8] T. Griffiths and Z. Ghahramani, “Infinite latent feature models and the Indian buffet process,” in *Adv. in Neural Information Processing Systems (NIPS)*, 2006, pp. 475–482.

[9] E. Punskaya, C. Andrieu, A. Doucet, and W. Fitzgerald, “Bayesian curve fitting using MCMC with applications to signal segmentation,” *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 747–758, Mar 2002.

[10] N. Dobigeon and J.-Y. Tourneret, “Bayesian orthogonal component analysis for sparse representation,” *IEEE Trans. Signal Process.*, vol. 58, no. 5, pp. 2675–2685, May 2010.

[11] C. P. Robert, *The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation*, 2nd ed., ser. Springer Texts in Statistics. New York: Springer-Verlag, 2007.

[12] O. Besson, N. Dobigeon, and J.-Y. Tourneret, “CS decomposition based Bayesian subspace estimation,” *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4210–4218, Aug. 2012.

[13] D. A. van Dyk and T. Park, “Partially collapsed Gibbs samplers: Theory and methods,” *J. Amer. Stat. Assoc.*, vol. 103, no. 482, pp. 790–796, June 2008.

[14] D. Knowles and Z. Ghahramani, “Infinite sparse factor analysis and infinite independent component analysis,” in *Proc. Int. Conf. Independent Component Analysis and Blind Source Separation (ICA)*, 2007.

[15] C. S. Herz, “Bessel functions of matrix argument,” *The Annals of Mathematics*, vol. 61, no. 3, p. 474, May 1955.