BAYESIAN SUBSPACE ESTIMATION USING SPARSE PROMOTING PRIOR

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

[(1)] Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, 59000 Lille, France
[(2)] University of Toulouse, IRIT/INP-ENSEEIHT, Toulouse, France

1. INTRODUCTION

Hyperspectral sensors record the light intensity beyond the visible spectra in hundreds of narrow contiguous bands. Images are characterized by a high spectral resolution but a low spatial precision due to sensors constraints. A crucial step called unmixing consists of decomposing each pixel as a combination of pure spectra, called endmembers. Endmembers act as fingerprints, improving the ability to analyse a scene.

Under reasonable assumptions, pixels are expected to live in a lower dimensional subspace whose dimension is intimately linked to the number of endmembers. The identification of this subspace yields gains in computational time, complexity and in data storage. However, determining both the relevant subspace dimension (*e.g.* the number of endmembers) and a suitable representation is a difficult problem. Existing methods are often parametric, such as thresholding the eigenvalues of Principal Component Analysis (PCA), and using eigenvectors as a subspace base (see [1] for a review). My Ph.D. aims at exploring Bayesian nonparametric inference to tackle these tasks.

We have already proposed a Bayesian formulation of anti-sparse coding [2, 3], primary motivated by the search of representations for endmembers. Anti-sparse representations aim at spreading the energy over all components uniformly.

In Data analysis, many methods exist to extract the underlying subspace. Model selection methods include criteria that quantify compromises between reconstruction and complexity (*e.g.* AIC or BIC). PCA also implicitly permits a dimension reduction by projecting observations onto a subset of orthonormal vectors. A probabilistic formulation of PCA has been proposed through factor analysis [4]. Existing extensions rely for instance on Laplace [5] or variational [6] approximations of the posterior distribution.

In this work, we investigate the use of Bayesian nonparametric inference associated to directional statistics to explore the set of subspaces. The goal is to avoid an arbitrary thresholding of eigenvalues as often done for PCA. We derive an empirical posterior distribution of bases of the latent subspace, where coefficients, *e.g.* projections, have been marginalized out. We proposed to use MCMC methods to sample according to this posterior and approximate estimators. In addition to subspace estimation, different indicators permit to assess the relevance of inferred projectors.

2. METHOD

2.1. Model

Let $\boldsymbol{y} = [y_1 \dots y_D]^t$ denote a D-dimensional observation vector. We propose the following latent factor model

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

where $\boldsymbol{P}$ is an orthonormal basis of $\mathbb{R}^D$ (*e.g.* the eigenvector of a classical PCA), $\boldsymbol{z}$ is a binary vector, $\boldsymbol{x}$ is the set of coefficients and $\odot$ denotes the term-wise product. The vector $\boldsymbol{n}$ is a white Gaussian noise of variance σ^2 leading to a quadratic discrepancy measure. Let $\boldsymbol{Y}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ denote the matrices resulting respectively from the concatenation of all vectors $\boldsymbol{y}$, $\boldsymbol{x}$ and $\boldsymbol{z}$.

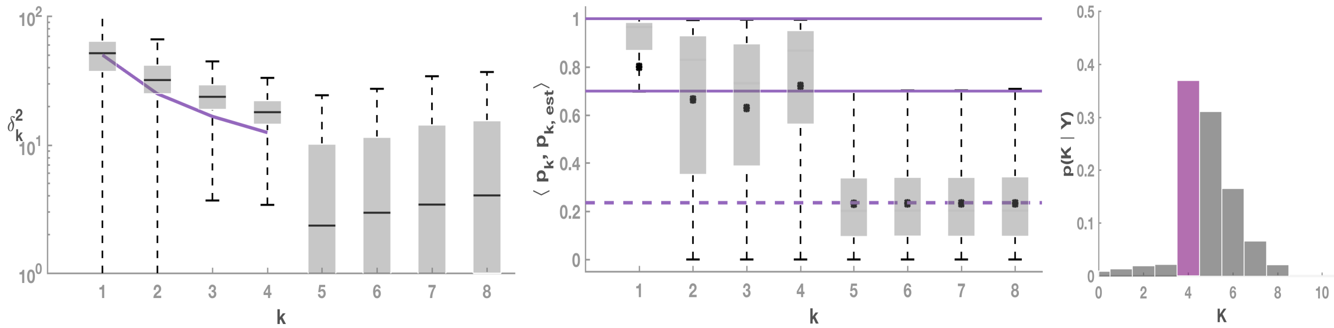
The aim is to estimate an orthonormal basis of a subspace of dimension $K \leq D$ that is not fixed in advance, to avoid the choice of a threshold. To that purpose, a uniform distribution over the set of orthonormal matrices in dimension D is elected as prior distribution over the projectors $\boldsymbol{P}$. We propose to use the binary property of $\boldsymbol{Z}$ to favor a parsimonious use of the components, in the spirit of a dimension reduction. The Indian buffet process (IBP) [7] is chosen as a prior distribution over the binary coefficients. The Indian buffet process (IBP) [7] can be seen as a distribution over the set of binary matrices with a potentially infinite number of rows. It penalizes large matrices. Moreover, the number of active rows of $\boldsymbol{Z}$ is limited to D because of the orthogonal constraints. We expect that the regularizing effect of the IBP still holds in low dimensions ($K \ll D$).

Independent Gaussian prior distributions are assigned to projection components gathered in $\boldsymbol{X}$. We choose to scale the variance in each direction to the noise. Thus, the variance σ_k^2 in a direction is expressed through $\sigma_k^2 = \delta_k^2 \sigma^2$, where δ_k^2 correspond to the ratio between the eigenvalues of PCA and the noise variance. All hyperpriors are chosen as non informative as possible, such that no parameter tuning is required.

2.2. Algorithm

Since the posterior distribution is too complex to analytically derive Bayesian estimators, a Monte Carlo Markov Chain (MCMC) sampler is derived to approximate the

*Part of this work has been funded thanks to the BNPSI ANR project no. ANR-13-BS-03-0006-01.
Corresponding author. clement.elvira@ec-lille.fr



(a) (b) (c)

Fig. 1. Posterior distributions of K (a), scale factors δ^2 (b) and dispersion of the projection measured through $\boldsymbol{P}_{\text{est}}^t \boldsymbol{P}_{\text{true}}$ (c).

marginal posterior distribution of the set of parameter $\boldsymbol{\theta} = \boldsymbol{P}, \boldsymbol{Z}, \sigma^2, (\delta_k^2)$

$$\mathrm{p}\left(\boldsymbol{\theta} \mid \boldsymbol{Y}\right) = \int \mathrm{p}\left(\boldsymbol{\theta}, \boldsymbol{X} \mid \boldsymbol{Y}\right) \mathrm{d}\boldsymbol{X} \tag{2}$$

where the projections $\boldsymbol{X}$ are integrated out. We thus obtain estimates of the orthonormal basis of a subspace of size $K \leq D$ with their corresponding eigenvalues (energies), and K, the number of selected components. The MCMC algorithm consists of a Metropolis-within-Gibbs sampler where parameters are iteratively sampled according to their respective conditional posterior distribution.

3. EXPERIMENTAL RESULTS

We propose to illustrate the relevance of the method on a simple simulated dataset. The following experiment is repeated 50 times. $N = 100$ observations of dimension $D = 16$ through the generative model $\boldsymbol{y} = \boldsymbol{P}\boldsymbol{x} + \boldsymbol{n}$ where $\boldsymbol{P}$ is a random orthonormal base where only $K = 4$ components are active, the coefficient $\boldsymbol{x}$ are Gaussian whose variance are proportional to x^{-1}, *e.g.* $[50\sigma^2, 25\sigma^2, 16\sigma^2, 12\sigma^2]$ and scale to the noise variance $\sigma^2 = 0.1$. We run for each simulation 1000 iterations including 50 burn-in iterations.

Figs. 1a to 1c plot three posterior distributions resulting from the concatenation of all iterations. One can see that the first four scale factors δ_k^2 are correctly recovered (Fig. 1a) with low relative variance compared to factors that correspond to inactive ones. This trends is confirmed by the posterior distribution of the alignment of the true $\boldsymbol{P}$ with the estimated ones (Fig. 1b). The first four orthogonal directions have an alignment in average higher than 0.7 while all the others are close to 0.23. Note that 0.23 corresponds to the expected value (represented by the horizontal dashed line) if the component is uniformly distributed on the subspace orthogonal of the previous ones. These results agree with the Maximum A Posteriori estimators extracted from the posterior distribution of K (Fig. 1c).

4. CONCLUSION

This Ph.D. work proposes a new Bayesian nonparametric framework to infer the intrinsic dimension of a set of observations. The MCMC framework permits to build a comprehensive statistical description of the solution through posterior distributions. These indicators could lead for instance to statistical test to help decision making.

Future works include both numerical and methodological investigations. We plan to explore the ability of the algorithm to recover components whose corresponding energy is below the average noise level (*i.e.*, where $\delta^2 \leq 1$). The approach could be validated through machine learning tasks on simple datasets. Conversely, we plan to incorporate the dimension reduction into a Bayesian formulation of these tasks, such as linear classification. Such methods may improve performances in hypespectral unmixing, where the dimension reduction plays a key role as explained Section 1.

5. REFERENCES

[1] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.*, 2012.

[2] C. Elvira, P. Chainais, and N. Dobigeon, "Democratic prior for anti-sparse coding," in *SSP*, 2016.

[3] C. Elvira, P. Chainais, and N. Dobigeon, "Bayesian anti-sparse coding," *submitted in TSP*, 2016, preprint available at https://arxiv.org/abs/1512.06086.

[4] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *"J. R. Stat. Soc. Series B*, 1999.

[5] T. P. Minka, "Automatic choice of dimensionality for pca," *NIPS*, vol. 13, pp. 514, 2000.

[6] V. Šmídl and A. Quinn, "On bayesian principal component analysis," *Comp. Stat. and Data Analysis*, 2007.

[7] T. L. Griffiths and Z. Ghahramani, "The indian buffet process: An introduction and review," *J. Mach. Learn. Res.*, 2011.