Bayesian nonparametric subspace estimation

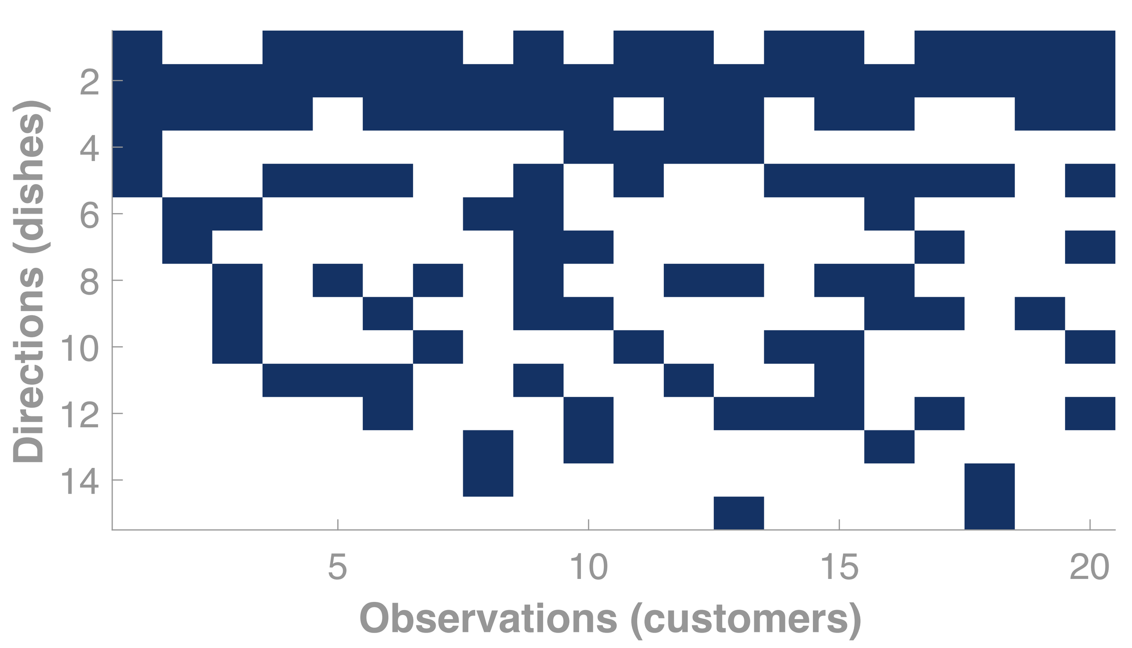
Clément Elvira[♪,♩], Pierre Chainais[♪] and Nicolas Dobigeon[♫]

[♪] University of Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL and [♫] University of Toulouse, IRIT/INP-ENSEEIHT, CNRS

Context and contributions

Reduction of dimension is an ubiquitous pre-processing in numerous signal processing and machine learning tasks. The choice of the number of **principal components** K retained has a significant impact on performances. Existing methods able to infer K generally rely on RJMCMC methods, variational Bayes approximations or information criteria. We propose a **nonparametric Bayesian** modeling of subspace estimation whose **size is not known in advance**. We propose a Gibbs sampler as inference scheme. We build estimators w.r.t. to all possible subspaces and **discuss their consistency**. The method is validated on synthetic dataset and two real tasks.

Indian Buffet Process = prior over potentially infinite sparse binary matrices with $\mathbb{E}[\boldsymbol{K}] = \boldsymbol{\alpha} \log(\boldsymbol{N})$ (**regularizing effect**)



Distributions on the Stiefel manifold $\mathcal{S}_L$

$\mathrm{vMF}(\mathbf{P}|\boldsymbol{F}) \sim \exp \mathrm{tr}\left[\boldsymbol{F}^t\mathbf{P}\right]$

$\mathrm{Bingham} \sim \exp \mathrm{tr}\left[\Lambda \mathbf{P}^t \boldsymbol{A}\mathbf{P}\right]$

Trick: write $\mathbf{p} = \mathbf{N}\boldsymbol{v}$ such that $\|\boldsymbol{v}\|_2 = 1$,

$\mathbf{N}$ an orthonormal basis of $\mathbf{P}^{\perp}_{\setminus k}$, $\boldsymbol{p} = \mathbf{N}\boldsymbol{v}$

Bayesian model

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

$\boldsymbol{y}_n \in \mathbb{R}^D$, the observation vector

$\mathbf{P} = [\mathbf{p}_1 \ldots \mathbf{p}_D, \mathbf{0} \ldots]$, $\mathbf{P}^t\mathbf{P} = \mathbf{I}_D$, $[\mathbf{p}_1 \ldots \mathbf{p}_D] \sim \mathcal{U}_{\mathcal{S}_D}$

$\boldsymbol{x}_n = [x_{1,n} \ldots]$ where $x_{k,n} \sim \mathcal{N}(0, \delta_k^2\sigma^2)$

$\boldsymbol{Z} = [\boldsymbol{z}_1 \ldots \boldsymbol{z}_n] \sim \mathrm{IBP}(\alpha)$ a binary matrix

$\boldsymbol{e}_n \sim \mathcal{N}(0, \sigma^2\mathbf{I}_D)$

$\boldsymbol{\theta} = \{\boldsymbol{\delta}^2, \sigma^2, \alpha\}$ vague conjugate priors

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

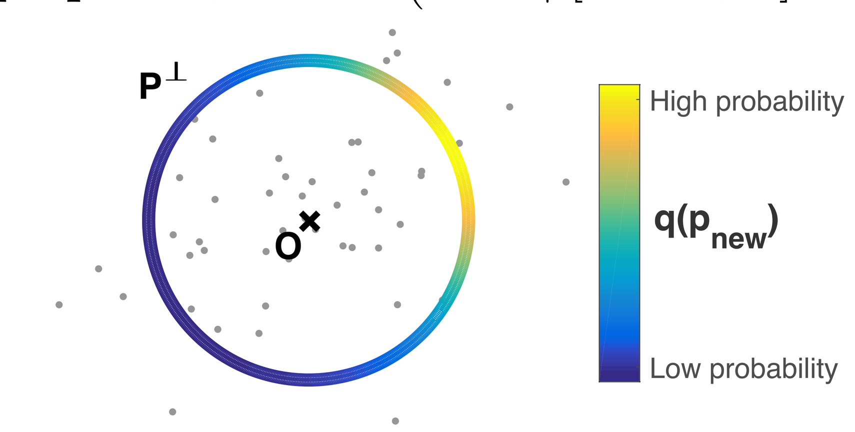
Sampling active directions

$$\mathbf{p}_k|\boldsymbol{Y}, \boldsymbol{P}_{\setminus k} \overset{d}{\sim} \mathrm{Bingham}\left(\sum z_{n,k} f(\boldsymbol{\delta}, \sigma^2)\mathbf{N}^t\boldsymbol{y}, \boldsymbol{y}^t\mathbf{N}\right)$$

Gibbs Sampler

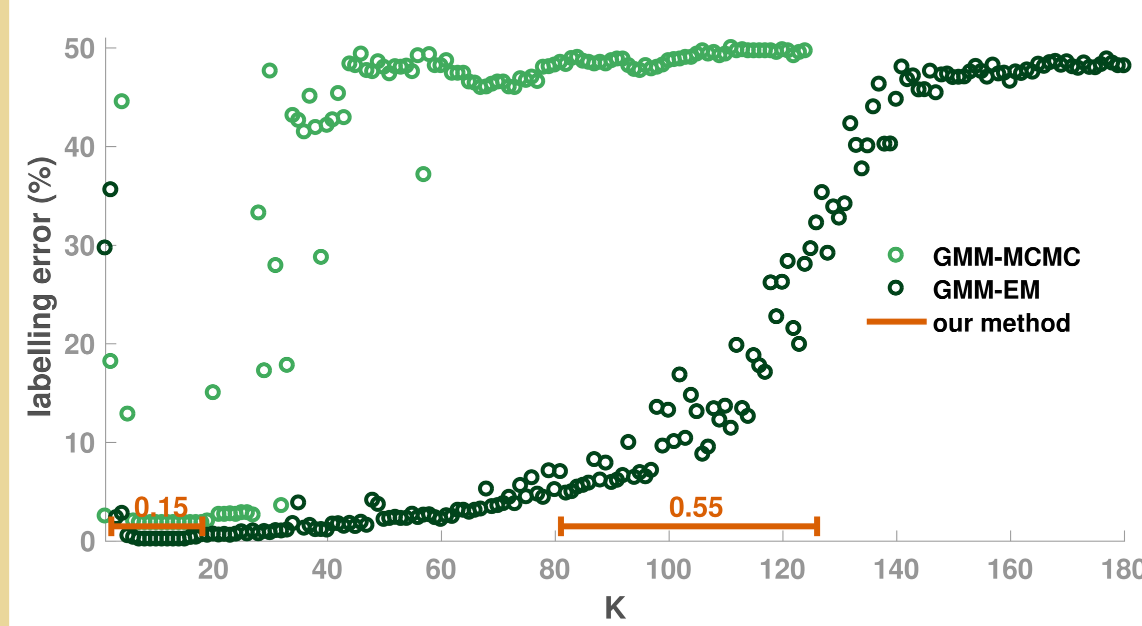
foreach *Iteration t* **do**
  Sample the non-singleton feature of $\boldsymbol{Z}$;
  Add/suppress directions $\sim$ von Mises Fisher;
  $\mathbf{P} \sim$ Bingham;
  $\boldsymbol{\delta}, \sigma^2, \alpha \sim$ conjugate distributions;
end

Explore new direction: Metropolis Hastings step with proposal $\mathrm{q} \overset{d}{=} \mathrm{vMF}\left(\mathbf{P}_{\text{new}} \mid [\boldsymbol{p}_1 \ldots \boldsymbol{p}_K]^{\perp}, \boldsymbol{Y}, \sigma^2\right)$



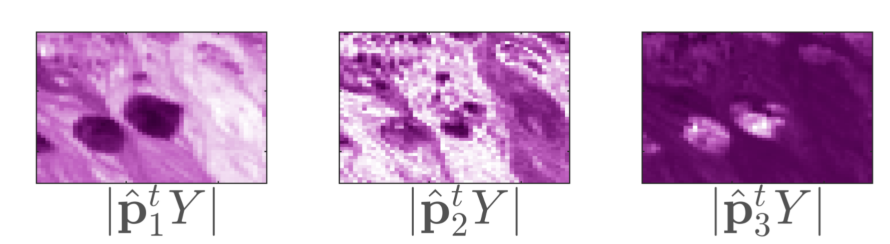
Applications

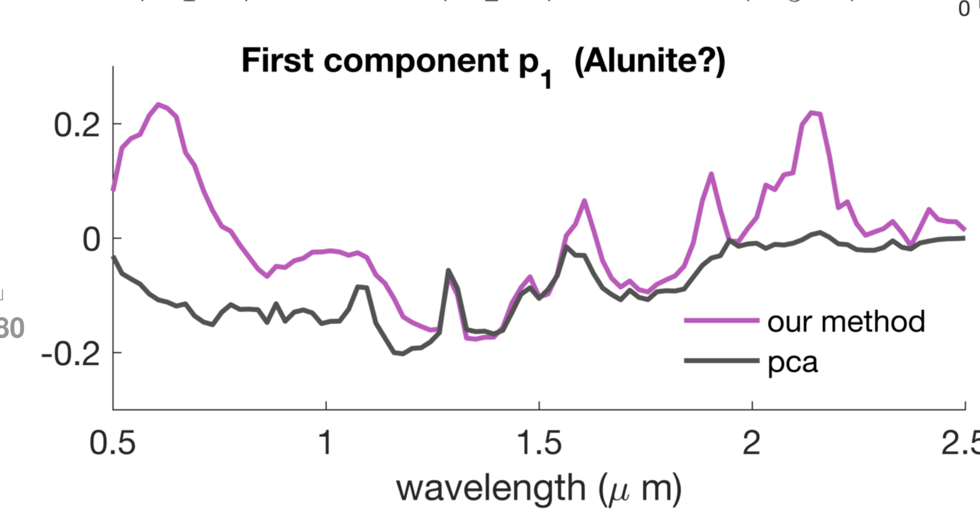
Mnist



- 600 images, digits 6 and 7, $\simeq 550$ pixels *(after removing those with null variance)*
- $\boldsymbol{x} \sim \pi_1\mathcal{N}_1 + (1-\pi_1)\mathcal{N}_2$ here
- compared with a MCMC- and EM-based inferences of a Gaussian mixture model for a varying number K of principal components
- our method reaches 1.5% of labeling error
- 2 areas are explored by the posterior $K|\boldsymbol{Y}$

Hyperspectral





- Cuprite-Hill dataset : $\simeq$ 100 dimensions and 500 observations
- linear model hypothesis : voxel $= \mathbf{E}\boldsymbol{a} +$ noise, with $\|\boldsymbol{a}\|_1 = 1$ and $\boldsymbol{a} > 0$.
- $\hat{K} = 15$, ground truth ~ 10
- interpretability of the first directions
- next step: include unmixing!

Estimating the number of components

The natural estimator is inconsistent

$$\forall k, \ \limsup_{N\to+\infty} \ \mathrm{P}\left(K_N = k \mid \boldsymbol{y}_1 \ldots \boldsymbol{y}_N, \alpha\right) < 1 \ \ \textit{with probability 1}$$

If $\boldsymbol{y} \sim \mathcal{N}(0, \sigma^2\mathrm{I})$, a stronger result holds

$$\mathrm{P}\left[K_N = 0|\boldsymbol{y}_1 \ldots \boldsymbol{y}_N, \alpha, \sigma^2\right] \xrightarrow[N\to+\infty]{a.s.} 0$$

Proposed methodology

All posteriors should be used to infer $\hat{K}$

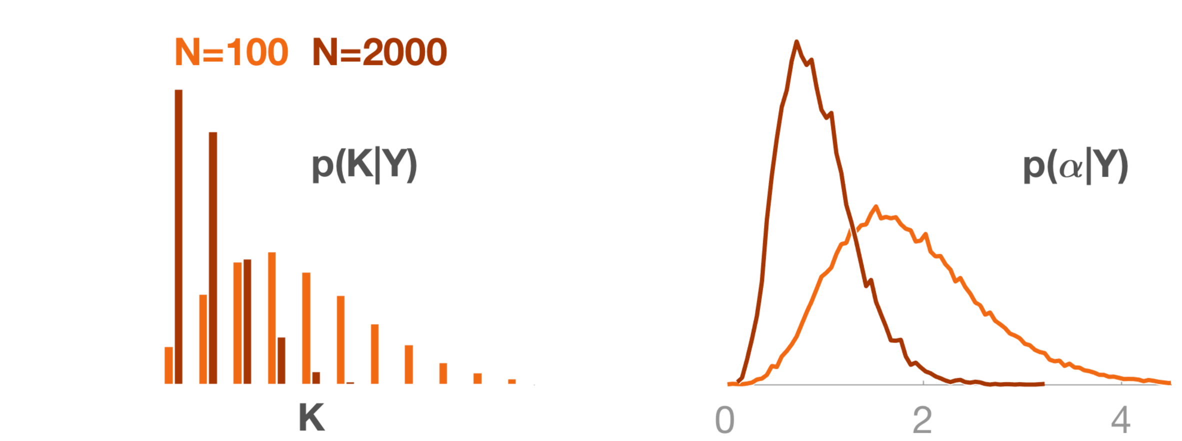
Let $\boldsymbol{u} \in \mathbb{R}^L$, $\|\boldsymbol{u}\|_2 = 1$, $\boldsymbol{V} \sim \mathcal{U}_{\mathcal{S}_L}$, $W = |\langle \boldsymbol{u}, \boldsymbol{V}\rangle|$

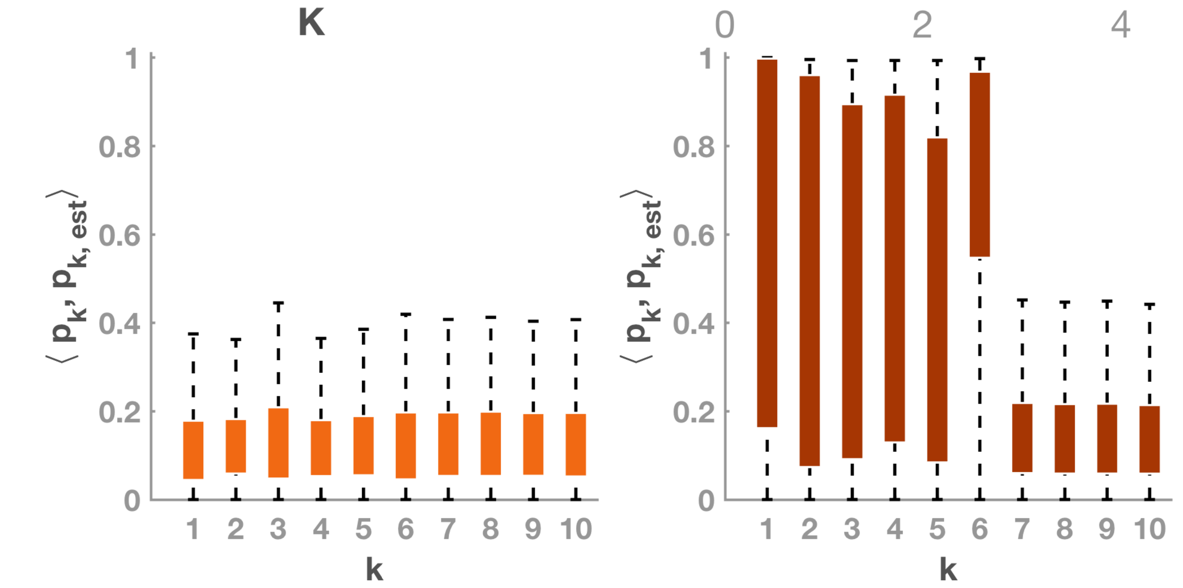
Compare the marginal distributions $\langle \boldsymbol{u}, \boldsymbol{p}_k^{(t)}\rangle$ to $\langle \boldsymbol{u}, \boldsymbol{V}\rangle$ for various L using CDF

$$\mathrm{p}_U\left(W \le \lambda\right) = \frac{\mathrm{vol}\left(\mathcal{S}_{L-2}\right)}{\mathrm{vol}\left(\mathcal{S}_{L-1}\right)} 2\int_0^{\lambda} \left(1-w^2\right)^{(L-3)/2} \mathrm{d}w$$

e.g., statistical tests.

Empirical observations





The estimator $\hat{K}|\boldsymbol{Y}$ empirically behaves correctly for large N, where $\mathbf{P}, \boldsymbol{Z}, \boldsymbol{\delta}, \alpha$ and σ^2 have been marginalized out.

Several settings are tested :
noise as input signal(✓), anisotropic noise(✓), scale factors δ^2 below the noise level(✗).

References

[1] C. Elvira, P. Chainais, and N. Dobigeon, "Bayesian antisparse coding," *IEEE Trans. Signal Process.*, vol. 65, pp. 1660–1672, April 2017.

[2] C. Elvira, P. Chainais, and N. Dobigeon, "Bayesian nonparametric modeling of latent subspace," in prep.

[♩] This work was supported in part by the BNPSI ANR Project ANR-13-BS-03-0006-01 and the Fondation Centrale Initiative.