Bayesian nonparametric subspace estimation

Clément Elvira, Pierre Chainais, Nicolas Dobigeon

IEEE ICASSP'17 New Orleans
March 7, 2017

Lower dimensional representation

subspace estimation $\quad Y \in \mathbb{R}^{N\times D} \to \mathbb{R}^{N\times K}$, with $K \ll D$

$$\boldsymbol{y}_n = \mathbf{P}\boldsymbol{x}_n + \boldsymbol{e}_n = \sum_{k=1}^{K} x_{n,k}\, \mathbf{p}_k + \text{noise}$$

choice of K $\to$ relevance of the dimension reduction
$\to$ critical impact on performances

e.g., Hyperspectral unmixing

A review of subspace estimation

- the most ubiquitous tool : PCA
- probabilistic PCA $\rightarrow$ latent factor analysis *Tipping and Bishop (1999)*

- difficulty : estimate the covariance matrix
- K as a latent variable

  $\rightarrow$ approximate the posterior *Minka (2000)*, *Šmídl and Quinn (2007)*

  $\rightarrow$ RJMCMC *Zhang et al. (2004)*

- related methodology : sparse PCA *Zou et al. (2006)*

Contributions

- Bayesian nonparametric sparse PCA
- IBP + Stiefel manifold
- asymptotic consistency of $K|\boldsymbol{Y}$
- numerical study
- applications : Hyperspectral + MNIST

Proposed model

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

$\boldsymbol{y}_1 \dots \boldsymbol{y}_N \in \mathbb{R}^D$, N observations

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

$\mathbf{P} = [\mathbf{p}_1 \dots \mathbf{p}_K, \mathbf{p}_{K+1} \dots \mathbf{p}_D]$, $\mathbf{P}^t\mathbf{P} = \mathbf{I}_D$, and $\mathbf{P} \sim \mathcal{U}_{\mathcal{O}_D}$

$\boldsymbol{Z} = [\boldsymbol{z}_1 \dots \boldsymbol{z}_N] \sim \mathrm{IBP}(\alpha)$ binary matrix $\rightarrow K$

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

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

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

The Indian Buffet Process prior *Griffiths and Ghahramani (2006)*

prior on binary matrices $\rightarrow$ sparsity

potentially infinite number of rows

column $\rightarrow$ observations, rows $\rightarrow$ feature

regularizing effect : $\mathbb{E}[K] = \alpha \log(N)$

$\rightarrow$ dimension reduction effect

Culinary metaphor

Algorithm

foreach *Iteration* t **do**

```
// K is sampled here
```

for $n \leftarrow 1$ **to** N **do**

sample active $(z_{k,n})_k$;

add / suppress directions $\sim$ von Mises Fischer;

end

```
// K is fixed
```

sample direction energy $\delta \sim$ conjugate shifted Gamma;

foreach *active direction* k **do**

$\mathbf{p}_k | \mathbf{P}_{\setminus k} \sim$ Bingham;

end

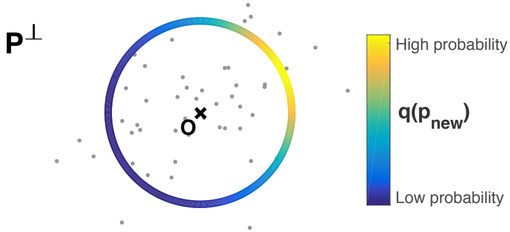
$\sigma^2, \alpha \sim$ conjugate distribution;

end

Posterior of interest

Explore new direction(s)

1. number of directions $\kappa^* \sim \mathcal{P}(\alpha)$
2. new directions $\mathbf{P}_{\text{new}} | \kappa^* \overset{d}{\sim} \text{q} = \text{von Mises Fischer}$
3. accept new state $\rightarrow$ Metropolis Hastings

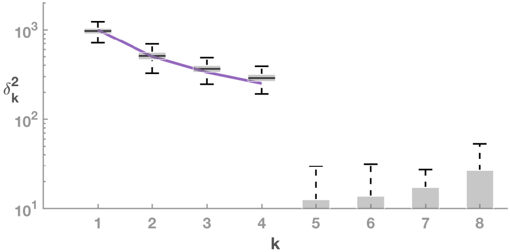


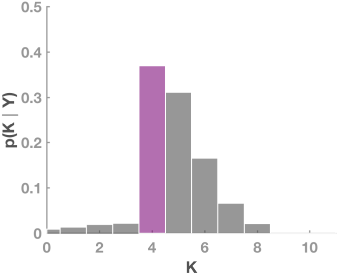
Examples

Generate observations

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

with arbitrary $\mathbf{P}$, $D = 16$, $K_{\text{true}} = 4$, $N = 500$ $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$





Marginal MAP estimation of K

Theoretical result

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

+ special case for white noise

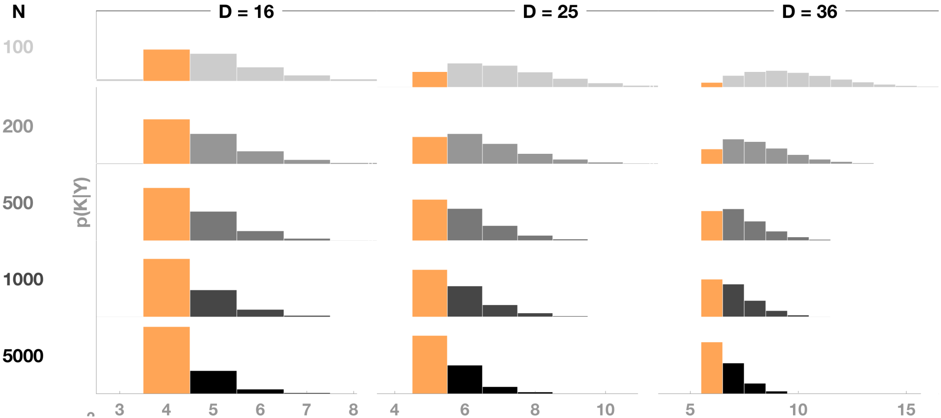
similar to *Miller and Harrison (2014)* for clustering with CRP

$\Rightarrow$ The MAP estimator of $K|\alpha$ is not consistent

This is bad news

Once K is known : subspace estimation *Besson et al. (2011)*

If the MAP is marginalized w.r.t. α



In summary

Inconsistent $\arg\ \max\ \mathrm{p}\left(K|\boldsymbol{Y},\alpha\right)$ mass never tends to 1

Empirically consistent $\arg\ \max\ \mathrm{p}\left(K|\boldsymbol{Y}\right)$ but no theoretical guarantees yet

$K|\boldsymbol{Y}$ is consistent for $N \gg D$

need for more reliable estimator

Proposed method to infer K

Key idea

K relevant directions $\rightarrow$ $D - K$ irrelevant directions

irrelevant $\rightarrow$ close to uniformly distributed on $\mathbf{P}^{\perp}_{\text{relevant}}$

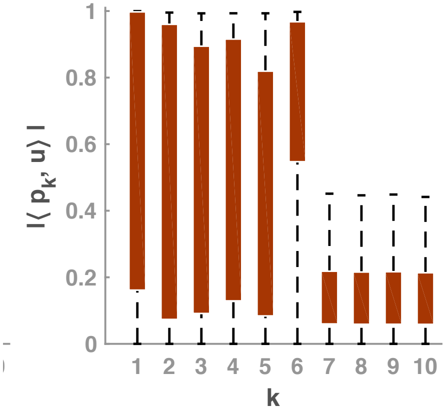
if $\|\boldsymbol{u}\|_2 = 1$, $\mathbf{p}$ uniform on $\mathcal{O}_{D-K}$, $W_{D-K} = |\langle \boldsymbol{u}, \mathbf{p} \rangle|$

$$\mathrm{p}\left(W_{D-K} \leq \lambda\right) = \frac{\mathrm{vol}\left(\mathcal{S}_{D-K-2}\right)}{\mathrm{vol}\left(\mathcal{S}_{D-K-1}\right)} \, 2 \int_0^{\lambda} \left(1 - w^2\right)^{(D-K-3)/2} \mathrm{d}w$$

Statistical test

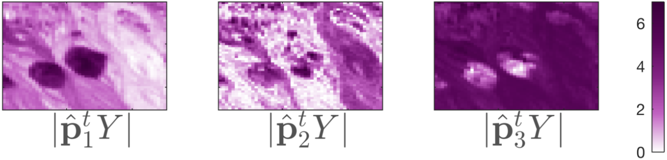
$H_0^{k,K} : |\langle \boldsymbol{u}, \mathbf{p}_k \rangle| \sim W_{D-K}$

$$\mathrm{p}\left(W_{D-K} \leq \lambda\right) = \frac{\mathrm{vol}\left(\mathcal{S}_{D-K-2}\right)}{\mathrm{vol}\left(\mathcal{S}_{D-K-1}\right)}\, 2 \int_0^{\lambda} \left(1 - w^2\right)^{(D-K-3)/2} \mathrm{d}w$$

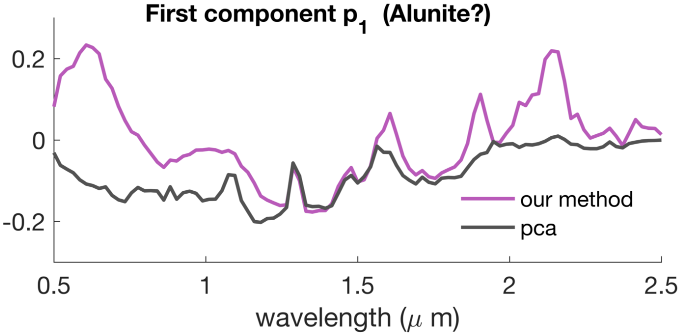


Application 1 : Hyperspectral subspace identification

$\boldsymbol{y} \in \mathbb{R}^{\#\text{wavelength}}$ $\quad \widehat{K} = 14$ ground truth $\simeq 10$



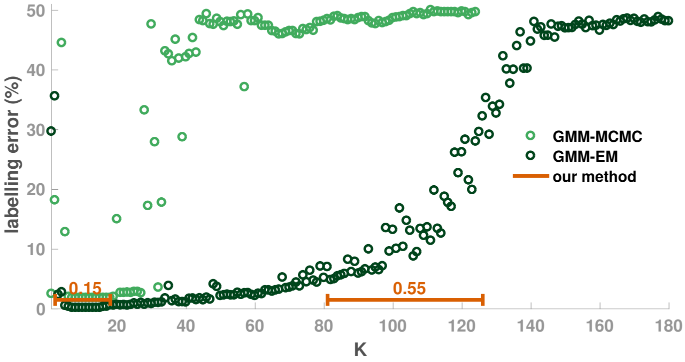
$|\hat{\mathbf{p}}_1^t Y|$ $\quad |\hat{\mathbf{p}}_2^t Y|$ $\quad |\hat{\mathbf{p}}_3^t Y|$



Application 2 : Coupling with clustering

$$\boldsymbol{x}_n \sim \pi_1 \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\delta}_1^2 \sigma^2) + (1-\pi_1)\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\delta}_2^2 \sigma^2)$$

2 digits (6 and 7) of MNIST dataset



Conclusion

- sparse Bayesian nonparametric PCA
- Metropolis within Gibbs for inference
- $K|\boldsymbol{Y}$ inconsistent $\rightarrow$ new estimator
- validation on simulated data
- 2 applications on real data.

- consistence of the new estimator?
- hyperspectral subspace identification $\rightarrow$ Hyperspectral unmixing
- extension to non linear methods?

`Preprint soon available`
`http://c-elvira.github.io`

- M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. Roy. Stat. Soc. Ser. B*, vol. 61, pp. 611–622, 1999.
- C. M. Bishop, "Bayesian pca," in *Adv. in Neural Information Processing Systems* (M. J. Kearns, S. A. Solla, and D. A. Cohn, eds.), pp. 382–388, MIT Press, 1999.
- T. P. Minka, "Automatic choice of dimensionality for pca," *Adv. in Neural Information Processing Systems*, vol. 13, p. 514, 2000.
- Z. Zhang, K. L. Chan, J. T. Kwok, and D.-Y. Yeung, "Bayesian inference on principal component analysis using reversible jump markov chain monte carlo," in *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, pp. 372–377, AAAI Press, 2004.
- V. Šmídl and A. Quinn, "On bayesian principal component analysis," *Comput. Stat. Data Anal.*, vol. 51, no. 9, pp. 4101 – 4123, 2007.
- O. Besson, N. Dobigeon, and J.-Y. Tourneret, "Minimum mean square distance estimation of a subspace," *IEEE Trans. Signal Process.*, vol. 59, pp. 5709–5720, Dec. 2011.
- 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.
- P.-A. Mattei, C. Bouveyron, and P. Latouche, "Globally sparse probabilistic PCA," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 976–984, 2016.

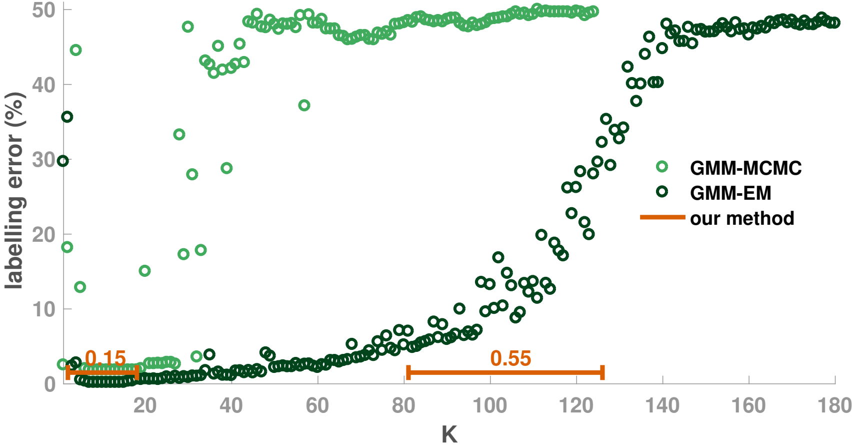
Annexe 1 :

Result with a generative model

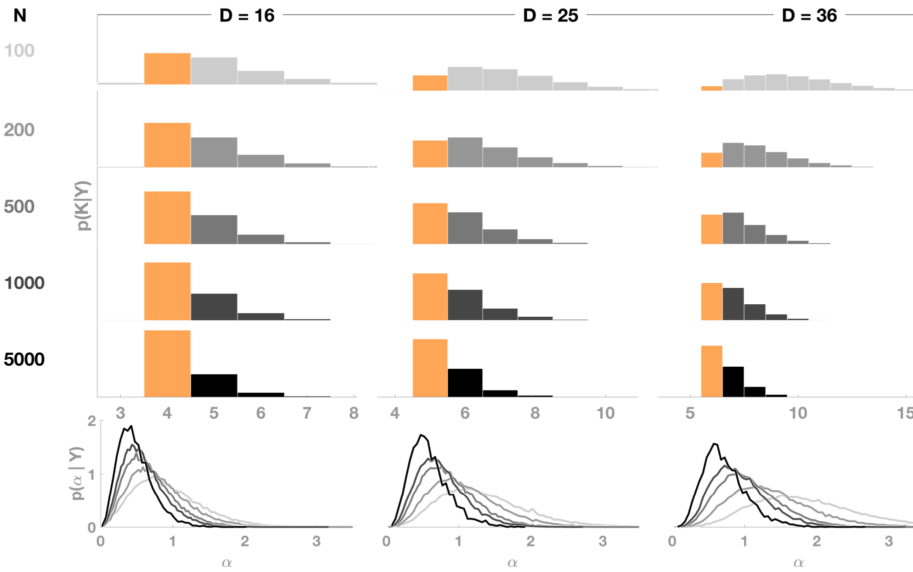
$$\text{If } \boldsymbol{y} \sim \mathcal{N}(0, \sigma^2 \mathrm{I}) \quad \mathrm{p}\left[K_N = 0 | \boldsymbol{y}_1 \ldots \boldsymbol{y}_N, \alpha, \sigma^2\right] \underset{N \to +\infty}{\overset{a.s.}{\to}} 0$$

Summary

General result	$\arg\max\ \mathrm{p}\left(K \mid \boldsymbol{Y}, \alpha\right)$	mass never tends to 1
Specific setting $\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$	$\arg\max\ \mathrm{p}\left(K \mid \boldsymbol{Y}, \alpha, \sigma^2\right)$	Only noise as input We expect $\mathrm{p}(K = 0) \to 0$ but $\mathrm{p}(K = 0) \to 1$
Empirical results various settings	$\arg\max\ \mathrm{p}\left(K \mid \boldsymbol{Y}\right)$	Behaves as expected but no guaranties How large N should be?



labelling error (%)
50
40
30
20
10
0
20
40
60
80
100
120
140
160
180
K
GMM-MCMC
GMM-EM
our method
0.15
0.55



N
D = 16
D = 25
D = 36
100
200
500
1000
5000
p(K|Y)
3 4 5 6 7 8
4 6 8 10
5 10 15
p(α | Y)
0 1 2
0 1 2 3
α
α
α

Update existing directions

A. $\mathbf{v}_k | \mathbf{P}_{\setminus k} \sim$ Bingham