Count data clustering and dimension reduction with a mixture of multinomial PCA

N. Jouvin, *joint work with* G. Bataillon C. Bouveyron, P. Latouche, & A. Livartowski

AppliBugs - Jeudi 15 décembre 2022

INRAE

MIA
PARIS-SACLAY
SOLSTIS

Count data

Motivations

Collaboration with Institut Curie (CRLCC): unsupervised exploration of anatomopathological data medical reports describing cellular lesions

Text quantization: model text documents as bag-of-words

the dog is on the table

Count data also appear in:

- Image quantization: bag-of-visual-words model (Sivic et al. 2005)
- RNA-seq data: *read* counts in a gene (Rau et al. 2015)
- Ecology: abundance data (Chiquet et al. 2018)

A quick look at the data

MICROBIOPSIE SOUS ECHOGRAPHIE DU SEIN DROIT

MACROSCOPIE

Cinq fragments de 5 à 15 mm

MICROSCOPIE

Les prélèvements examinés correspondent à des fragments de tissu mammaire remanié par une prolifération tumorale dont les caractères morphologiques sont ceux d'un adénocarcinome canalaire infiltrant. Cette lésion est peu différenciée, d'architecture essentiellement trabéculaire. Les cellules néoplasiques comportent des atypies nucléaires marquées. L'index mitotique est élevé (22 mitoses sur 10 champs au grandissement 400). Deux fragments de 8 et 15 mm. Adénocarcinome mammaire de type canalaire infiltrant peu différencié. Grade histo-pronostique (EE) : III Index mitotique élevé.

...

MACROBIOPSIE DU SEIN GAUCHE

MACROSCOPIE

3 fragments de 7 à 15 mm

MICROSCOPIE

Tous les prélèvements ont un aspect histologique similaire. Ils correspondent à des fragments de tissu mammaire remanié par des lésions de mastose fibreuse commune. Présence d'un discret infiltrat inflammatoire. On retrouve également quelques microcalcifications. L'un des prélèvements cryo-préservés sera analysé histologiquement et un compte rendu complémentaire adressé ultérieurement. Trois fragments de 7 à 15 mm. Lésions de mastose fibreuse. Le prélèvement paraît peu significatif. Une analyse complémentaire sur le prélèvement cryo-préservé sera réalisée.

Document-term matrix					
Documents \ Terms	lésions	canalaire	...	lobulaire	métaplasie
"Lésions (...) carcinome canalaire"	2	1	...	0	0
"Lésions bénignes (...) métaplasie"	3	0	...	0	1

Statistical modelization and issues

Observations $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, in dimension p.

- Discrete: $\boldsymbol{y}_i \in \mathbb{N}^p$
- Positive: $\boldsymbol{y}_i \geq 0$
- Total count: $c_i \coloneqq \sum_j y_{ij}$

Some problems compared to standard frameworks:

- Statistical nature of raw data (*e.g.* no Gaussianity)
- Sparsity / zero-inflation / Over-dispersion
- Small sample scenario: $p >> n$

Solutions: use model-based approaches and integrate dimension reduction in the model.

Organization

Model-based clustering with mixture models

Greedy inference of mixture of multinomial PCA

Performances & real-data application

Model-based clustering with mixture models

Clustering in a nutshell

Goal: Find a partition $\boldsymbol{Z}$ of $\{1, \ldots, n\}$ into K groups by optimizing some criteria

$$\boldsymbol{Z}^\star \in \arg\max_{\boldsymbol{Z}} L(\boldsymbol{Y}, \boldsymbol{Z})$$

Clustering in a nutshell

Goal: Find a partition $\boldsymbol{Z}$ of $\{1, \ldots, n\}$ into K groups by optimizing some criteria

$$\boldsymbol{Z}^\star \in \arg\max_{\boldsymbol{Z}} L(\boldsymbol{Y}, \boldsymbol{Z})$$

How to design L ? Optimization ?

- Model-free heuristics: *ad hoc* criteria and optimization, *e.g.*
  - distance based
  - K-means algorithm

Clustering in a nutshell

Goal: Find a partition $\boldsymbol{Z}$ of $\{1, \ldots, n\}$ into K groups by optimizing some criteria

$$\boldsymbol{Z}^\star \in \arg\max_{\boldsymbol{Z}} L(\boldsymbol{Y}, \boldsymbol{Z})$$

How to design L ? Optimization ?

- Model-free heuristics: *ad hoc* criteria and optimization, *e.g.*
  - distance based
  - K-means algorithm
- Model-based, *e.g.*
  - maximum-likelihood or maximum a posteriori
  - EM algorithm

We seek a partition of n observations : $\boldsymbol{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$

Latent variable $\boldsymbol{z}_i \sim \mathcal{M}_K(\boldsymbol{z}_i \mid 1, \boldsymbol{\pi})$: $z_{ik} = 1$ iff $\boldsymbol{y}_i$ is in cluster k

We seek a partition of n observations : $\boldsymbol{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$

Latent variable $\boldsymbol{z}_i \sim \mathcal{M}_K(\boldsymbol{z}_i \mid 1, \boldsymbol{\pi})$: $z_{ik} = 1$ iff $\boldsymbol{y}_i$ is in cluster k

Mixture model

$$\forall i, \quad \boldsymbol{y}_i \mid \{z_{ik} = 1\} \sim p(\cdot \mid \boldsymbol{\theta}_k)$$

- Gaussian mixture models (GMM): $p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_k) = \mathcal{N}_p(\boldsymbol{y}_i \mid \boldsymbol{m}_k, \boldsymbol{\Sigma}_k)$
- Mixture of multinomials (MoM): $p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_k) = \mathcal{M}_p(\boldsymbol{y}_i \mid \theta_k)$

Marginal over $\boldsymbol{Z}$:

$$p(\boldsymbol{Y} \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{i=1}^n p(\boldsymbol{y}_i \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{i=1}^n \sum_{k=1}^K \pi_k p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_k)$$

We seek a partition of n observations : $\boldsymbol{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$

Latent variable $\boldsymbol{z}_i \sim \mathcal{M}_K(\boldsymbol{z}_i \mid 1, \boldsymbol{\pi})$: $z_{ik} = 1$ iff $\boldsymbol{y}_i$ is in cluster k

Mixture model

$$\forall i, \quad \boldsymbol{y}_i \mid \{z_{ik} = 1\} \sim p(\cdot \mid \boldsymbol{\theta}_k)$$

- Gaussian mixture models (GMM): $p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_k) = \mathcal{N}_p(\boldsymbol{y}_i \mid \boldsymbol{m}_k, \boldsymbol{\Sigma}_k)$
- Mixture of multinomials (MoM): $p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_k) = \mathcal{M}_p(\boldsymbol{y}_i \mid \theta_k)$

Marginal over $\boldsymbol{Z}$:

$$p(\boldsymbol{Y} \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{i=1}^n p(\boldsymbol{y}_i \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{i=1}^n \sum_{k=1}^K \pi_k p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_k)$$

Clustering ? Estimation of $\hat{\boldsymbol{\theta}}$ + *posterior* $p(\boldsymbol{Z} \mid \boldsymbol{Y}, \hat{\boldsymbol{\theta}})$

Dealing with high-dimensional statistical estimation

Problem: large $p \implies$ over-parameterization

- GMM: $\boldsymbol{S}_k$ involves $\mathcal{O}(p^2)$ parameters
- MoM: $K(p-1)$ parameters to estimate, what if $n << p$?

Dealing with high-dimensional statistical estimation

Problem: large $p \implies$ over-parameterization

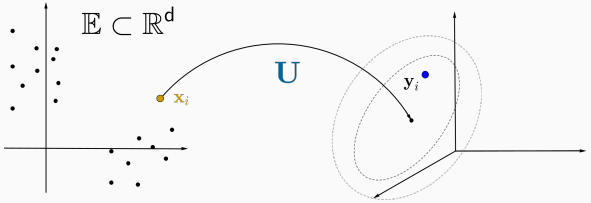
- GMM: $\boldsymbol{S}_k$ involves $\mathcal{O}(p^2)$ parameters
- MoM: $K(p-1)$ parameters to estimate, what if $n << p$?

Solution: Reduce the number of parameters

Probabilistic dimension reduction

$$\boldsymbol{x}_i \in \mathbb{E} \subset \mathbb{R}^d \longrightarrow \boldsymbol{y}_i \approx \boldsymbol{U}\boldsymbol{x}_i$$

Linear embedding $\boldsymbol{U} \in \mathbb{R}^{p \times d}$, $d \ll p$



Greedy inference of mixture of multinomial PCA

Multinomial PCA: probablistic dimension reduction for count data

Continuous data: Probabilistic PCA (pPCA, Tipping et al. 1999)

$$\boldsymbol{x}_i \sim \mathcal{N}_d(\mathbf{0}_d,\, \boldsymbol{I}_d) \qquad \text{(latent space: } \mathbb{E} = \mathbb{R}^d\text{)}$$

$$\boldsymbol{y}_i \mid \boldsymbol{x}_i \sim \mathcal{N}_p(\boldsymbol{U}\boldsymbol{x}_i + \mu,\, \sigma^2 \boldsymbol{I}_p) \qquad \text{(observation space)}$$

Multinomial PCA: probablistic dimension reduction for count data

Continuous data: Probabilistic PCA (pPCA, Tipping et al. 1999)

$$\boldsymbol{x}_i \sim \mathcal{N}_d(\mathbf{0}_d,\, \boldsymbol{I}_d) \qquad \text{(latent space: } \mathbb{E} = \mathbb{R}^d\text{)}$$
$$\boldsymbol{y}_i \mid \boldsymbol{x}_i \sim \mathcal{N}_p(\boldsymbol{U}\boldsymbol{x}_i + \mu,\, \sigma^2 \boldsymbol{I}_p) \qquad \text{(observation space)}$$

Count data: Multinomial PCA (MPCA, Buntine 2002)

$$\boldsymbol{x}_i \sim \mathcal{D}_d(\boldsymbol{\alpha}) \qquad \text{(latent space: } \mathbb{E} = \Delta_d\text{)}$$
$$\boldsymbol{y}_i \mid \boldsymbol{x}_i \sim \mathcal{M}_p(c_i, \boldsymbol{U}\boldsymbol{x}_i) \qquad \text{(observation space)}$$

- $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d] \in (\Delta_p)^d$ is called the *topic* matrix
- also known as Latent Dirichlet Allocation (Blei et al. 2003)

Not designed for clustering *per se*

Mixture of Multinomial PCA

One latent variable per cluster:

$$\boldsymbol{x} = (\boldsymbol{x}_k)_k, \quad \boldsymbol{x}_k \overset{i.i.d.}{\sim} \mathcal{D}_d(\boldsymbol{\alpha})$$

$$\forall i, \quad \boldsymbol{y}_i \mid \boldsymbol{x} \sim \sum_{k=1}^{K} \pi_k \, \mathcal{M}_p(c_i, \boldsymbol{U}\boldsymbol{x}_k)$$

Constrained MoM model: $\boldsymbol{\theta}_k = \boldsymbol{U}\boldsymbol{x}_k$ (Carel et al. 2017)

One latent variable per cluster:

$$\boldsymbol{x} = (\boldsymbol{x}_k)_k, \quad \boldsymbol{x}_k \overset{i.i.d.}{\sim} \mathcal{D}_d(\boldsymbol{\alpha})$$

$$\forall i, \quad \boldsymbol{y}_i \mid \boldsymbol{x} \sim \sum_{k=1}^{K} \pi_k \mathcal{M}_p(c_i, \boldsymbol{U}\boldsymbol{x}_k)$$

Constrained MoM model: $\boldsymbol{\theta}_k = \boldsymbol{U}\boldsymbol{x}_k$ (Carel et al. 2017)

Property

Suppose $\boldsymbol{Z}$ known and fixed, construct K meta-observations

$$\tilde{\boldsymbol{Y}}_k(\boldsymbol{Z}) = \sum_{i=1}^{n} z_{ik}\boldsymbol{y}_i$$

Then, $\boldsymbol{Y} \mid \boldsymbol{Z}$ follows a MPCA model on $\tilde{\boldsymbol{Y}}(\boldsymbol{Z})$

Also known as the probabilistic clustering-projection model (PCP, Yu et al. 2005)

Classification likelihood approach

$$\underset{\boldsymbol{Z},\boldsymbol{U},\boldsymbol{\pi}}{\arg\max} \left\{ \log p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\pi}, \boldsymbol{U}) = \log p(\boldsymbol{Z} \mid \boldsymbol{\pi}) + \underbrace{\log p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{U})}_{(*)\text{ MPCA on } \tilde{\boldsymbol{Y}}(\boldsymbol{Z})} \right\}$$

Problems:

1. Combinatorics: number of partitions exponential with n
2. $(*)$ is intractable because of marginal over $\boldsymbol{x}$

Solutions:

1. Variational inference layer on $\boldsymbol{x}$

$$\boldsymbol{x} \sim q, \qquad \log p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\pi}, \boldsymbol{U}) \geq \mathcal{J}(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{U}, q)$$

2. Greedy algorithm for joint inference and clustering

Algorithm: Explore partition space using $\mathcal{J}$ as a surrogate objective

Input: K, d, $\boldsymbol{Z}^{(0)}, \boldsymbol{\pi}^{(0)}, \boldsymbol{U}$

while $\boldsymbol{Z}$ *has not converged* **do**

For all $i = 1, \ldots, n$, try individual swaps: $z_{ik}^{(t)} = 1 \rightarrow z_{il}^{(tmp)} = 1$

`// Difference with standard greedy approaches`

Use variational inference to update q

$$(\mathcal{J}_l, q_l) = \arg\max_{q} \mathcal{J}(\boldsymbol{Z}^{(tmp)}, \boldsymbol{\pi}^{(t)}, \boldsymbol{U}, q)$$

Select $l^\star = \arg\max_l \mathcal{J}_l$

$$z_{il^\star}^{(t+1)} = 1, \qquad q^{(t+1)} = q_{l^\star} \qquad \boldsymbol{\pi}^{(t+1)} = \textstyle\sum_i \boldsymbol{z}_i^{(t+1)}/n$$

end

How to choose the pair (K, d) ?

Integrated Classification Likelihood (ICL, Biernacki et al. 2000)

$$\log p(\boldsymbol{Y}, \boldsymbol{Z}) = \log \int_{\boldsymbol{\pi}} \int_{\boldsymbol{U}} p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{U}) \, \mathrm{d}\boldsymbol{U} \, \mathrm{d}\boldsymbol{\pi}$$

ICL criterion for MMPCA

Laplace and Stirling approximations combined with a variational approximation on $p(\boldsymbol{Y} \mid \boldsymbol{Z})$ lead to

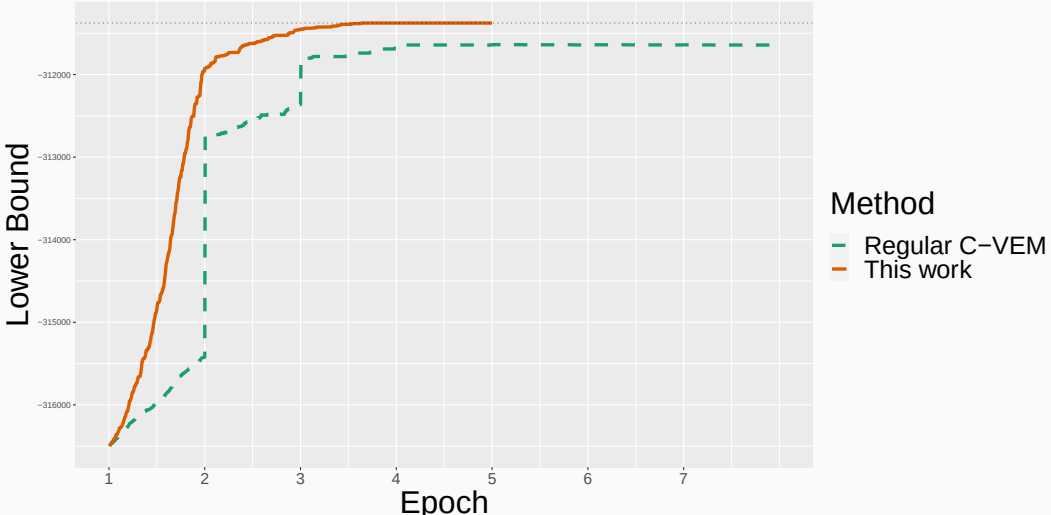
$$\begin{aligned} \mathrm{ICL}_{MMPCA}(K, d) = & \mathcal{J}(\hat{\boldsymbol{Z}}, \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{U}}, \hat{q}) \\ & - \frac{d(p-1)}{2} \log(K) - \frac{K-1}{2} \log(n) \end{aligned}$$

Performances & real-data application

Noiseless setting: Branch & Bound VS standard C-VEM

Fixed setting:

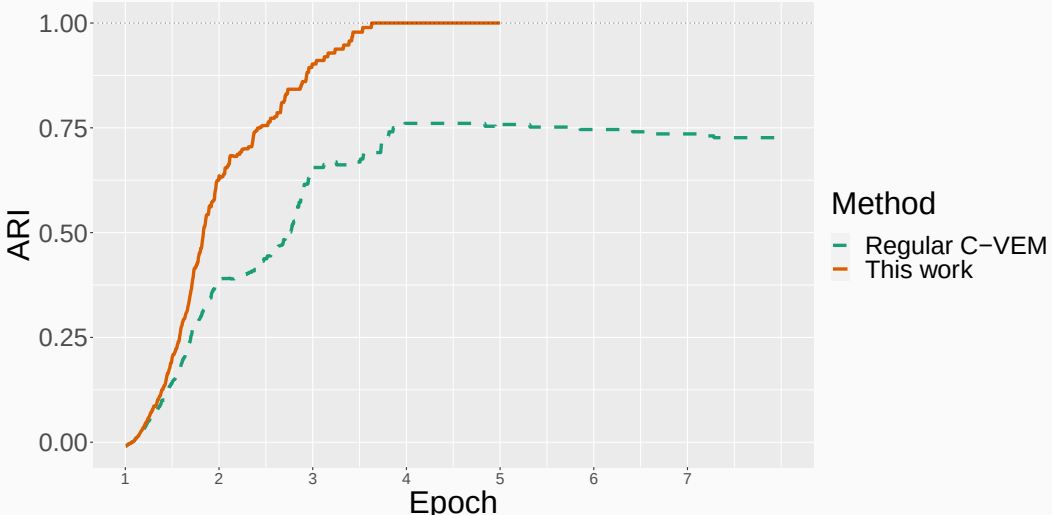
$n = 400, \quad p = 1000, \quad K = 6, \quad d = 4, \quad \boldsymbol{U}^\star, \quad \boldsymbol{x}^\star,$ metric: ARI



Noiseless setting: Branch & Bound VS standard C-VEM

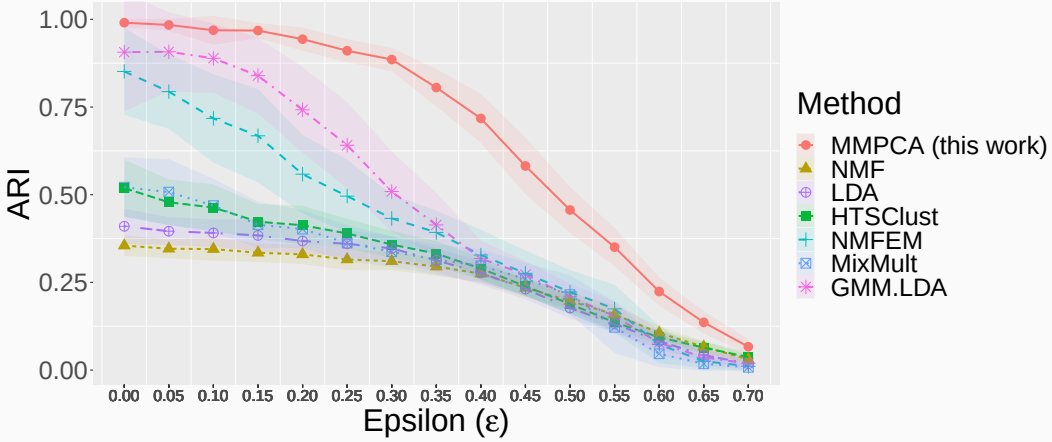
Fixed setting:

$n = 400, \quad p = 1000, \quad K = 6, \quad d = 4, \quad \boldsymbol{U}^\star, \quad \boldsymbol{x}^\star,$ metric: ARI



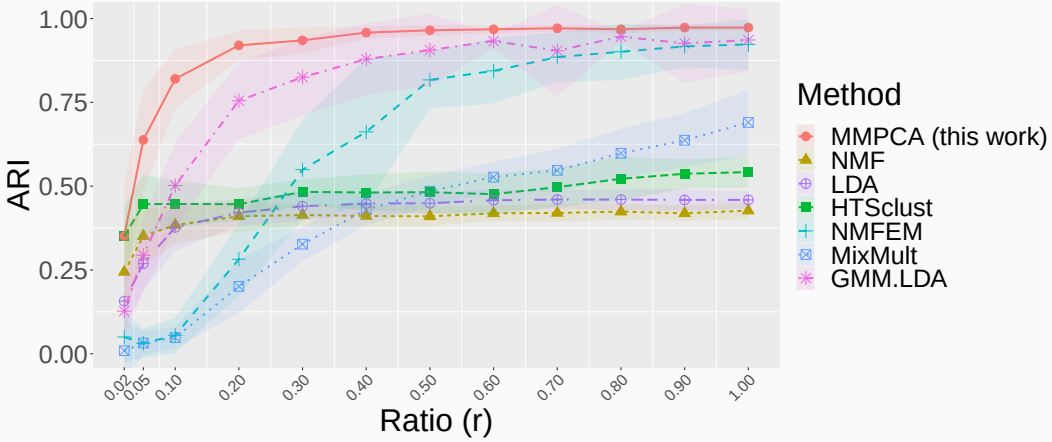
Scenario 1: noisy setting

$n = 400,$ $p = 1000,$ Noise level: $\epsilon \in [0, 1]$



Scenario 2: small-sample sizes

$p = 1000,$ $\epsilon = 0.2,$ $n = r \times p,\ r \in [0, 1]$



Application: clustering of anatomopathological reports

MICROBIOPSIE SOUS ECHOGRAPHIE DU SEIN DROIT

MACROSCOPIE

Cinq fragments de 5 à 15 mm

MICROSCOPIE

Les prélèvements examinés correspondent à des fragments de tissu mammaire remanié par une prolifération tumorale dont les caractères morphologiques sont ceux d'un adénocarcinome canalaire infiltrant. Cette lésion est peu différenciée, d'architecture essentiellement trabéculaire. Les cellules néoplasiques comportent des atypies nucléaires marquées. L'index mitotique est élevé (22 mitoses sur 10 champs au grandissement 400). Deux fragments de 8 et 15 mm. Adénocarcinome mammaire de type canalaire infiltrant peu différencié. Grade histo-pronostique (EE) : III Index mitotique élevé.

...

MACROBIOPSIE DU SEIN GAUCHE

MACROSCOPIE

3 fragments de 7 à 15 mm

MICROSCOPIE

Tous les prélèvements ont un aspect histologique similaire. Ils correspondent à des fragments de tissu mammaire remanié par des lésions de mastose fibreuse commune. Présence d'un discret infiltrat inflammatoire. On retrouve également quelques microcalcifications. L'un des prélèvements cryo-préservés sera analysé histologiquement et un compte rendu complémentaire adressé ultérieurement. Trois fragments de 7 à 15 mm. Lésions de mastose fibreuse. Le prélèvement paraît peu significatif. Une analyse complémentaire sur le prélèvement cryo-préservé sera réalisée.

Document-term matrix					
Documents \ Terms	lésions	canalaire	...	lobulaire	métaplasie
"Lésions (...) carcinome canalaire"	2	1	...	0	0
"Lésions bénignes (...) métaplasie"	3	0	...	0	1

Results

Context: textual reports describing histopathological slides

- Benign
- Lobular carcinoma
- Non Special Type (NST) carcinoma, *e.g.* ductal

Unsupervised analysis: select $K = 7$ and $d = 5$

	Benign	NST carcinoma	Lobular carcinoma
1	0	0	43
2	1	31	1
3	0	106	0
4	231	3	0
5	0	211	0
6	0	126	0
7	0	113	0

Results

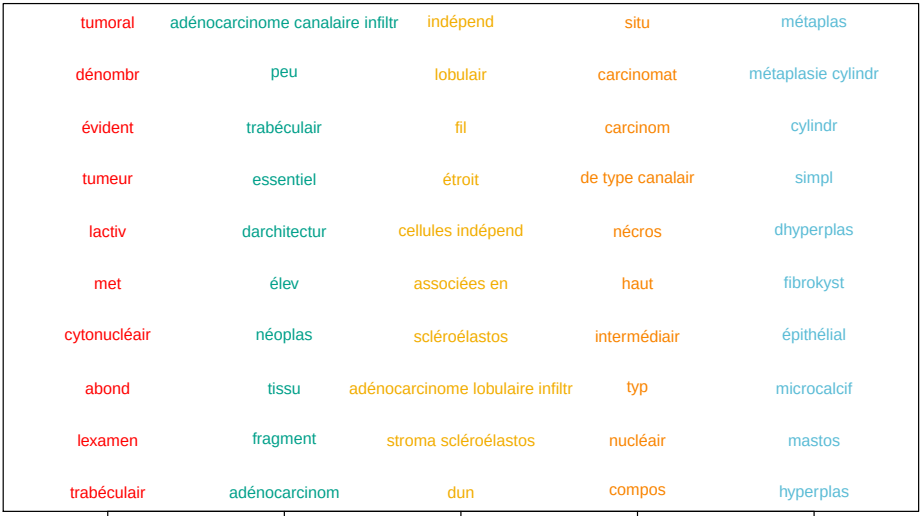
Context: textual reports describing histopathological slides

- Benign
- Lobular carcinoma
- Non Special Type (NST) carcinoma, *e.g.* ductal

Unsupervised analysis: select $K = 7$ and $d = 5$

	Benign	NST carcinoma	Lobular carcinoma
1	0	0	43
2	1	31	1
3	0	106	0
4	231	3	0
5	0	211	0
6	0	126	0
7	0	113	0

Topic space


tumoral
dénombr
évident
tumeur
lactiv
met
cytonucléair
abond
lexamen
trabéculair
adénocarcinome canalaire infiltr
peu
trabéculair
essentiel
darchitectur
élev
néoplas
tissu
fragment
adénocarcinom
indépend
lobulair
fil
étroit
cellules indépend
associées en
scléroélastos
adénocarcinome lobulaire infiltr
stroma scléroélastos
dun
situ
carcinomat
carcinom
de type canalair
nécros
haut
intermédiair
typ
nucléair
compos
métaplas
métaplasie cylindr
cylindr
simpl
dhyperplas
fibrokyst
épithélial
microcalcif
mastos
hyperplas
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5

Interpretability: representation in the topic space

	Topic1	Topic2	Topic3	Topic4	Topic5
$\boldsymbol{x}_1$	0.00	0.01	**0.98**	0.00	0.00
$\boldsymbol{x}_2$	0.19	0.11	0.04	**0.38**	**0.29**
$\boldsymbol{x}_3$	0.13	0.09	0.01	**0.76**	0.00
$\boldsymbol{x}_4$	0.01	0.00	0.01	0.01	**0.97**
$\boldsymbol{x}_5$	0.00	**1.00**	0.00	0.00	0.00
$\boldsymbol{x}_6$	0.05	**0.65**	0.03	0.26	0.01
$\boldsymbol{x}_7$	**0.74**	0.12	0.03	0.11	0.00

- Topic3: vocabulary of lobular cancer
- Topic5: vocabulary of benign lesions

Interpretability: representation in the topic space

	Topic1	Topic2	Topic3	Topic4	Topic5
$\boldsymbol{x}_1$	0.00	0.01	**0.98**	0.00	0.00
$\boldsymbol{x}_2$	0.19	0.11	0.04	**0.38**	**0.29**
$\boldsymbol{x}_3$	0.13	0.09	0.01	**0.76**	0.00
$\boldsymbol{x}_4$	0.01	0.00	0.01	0.01	**0.97**
$\boldsymbol{x}_5$	0.00	**1.00**	0.00	0.00	0.00
$\boldsymbol{x}_6$	0.05	**0.65**	0.03	0.26	0.01
$\boldsymbol{x}_7$	**0.74**	0.12	0.03	0.11	0.00

Cluster 2 contains micro-calcifications and peaked towards

- Topic4: vocabulary of *in-situ* lesions
- Topic5: vocabulary of benign lesions

Posterior explanation: all samples came from macro-biopsy exams

Conclusion

Summary

Model-based approach for high-dimensional count data clustering:

- Robust to noise and small-sample scenarios
- Handle unbalanced clusters and model selection
- Relevant results on real-data application

The clustering algorithm is essential. It would be interesting to:

- analyze its properties (greedy heuristic)
- reduce its computational costs

Sources

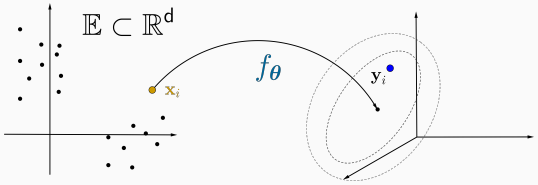
- ▶ Journal article available (link)
- ▶ CRAN package **MoMPCA**

Perspectives

Different latent spaces, *e.g.* $\boldsymbol{x}_i$ follows a Gaussian mixture

- Introduce correlation in the latent space

Non-linear transformations: $\boldsymbol{y}_i \mid \boldsymbol{x}_i \sim p\left(\cdot \mid f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right)$



- Poisson log-normal models $f_{\boldsymbol{\theta}}$ natural parametrization (Chiquet et al. 2018)
- Variational auto-encoders: $f_{\boldsymbol{\theta}}$ neural net with weights $\boldsymbol{\theta}$ (Mattei et al. 2018)

Thank you for your attention !

Questions ?

References

- Biernacki, Christophe, Gilles Celeux, and Gérard Govaert (2000). “Assessing a mixture model for clustering with the integrated completed likelihood”. In: *IEEE transactions on pattern analysis and machine intelligence* 22.7, pp. 719–725.
- Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.
- Buntine, Wray (2002). “Variational extensions to EM and multinomial PCA”. In: *European Conference on Machine Learning*. Springer, pp. 23–34.
- Carel, Léna and Pierre Alquier (2017). “Simultaneous dimension reduction and clustering via the NMF-EM algorithm”. In: *Advances in Data Analysis and Classification*, pp. 1–30.
- Chiquet, Julien, Mahendra Mariadassou, and Stéphane Robin (2018). “Variational inference for probabilistic Poisson PCA”. In: *The Annals of Applied Statistics* 12.4, pp. 2674–2698.

- Mattei, Pierre-Alexandre and Jes Frellsen (2018). “Leveraging the exact likelihood of deep latent variable models”. In: *arXiv preprint arXiv:1802.04826*.
- Rau, Andrea et al. (Jan. 2015). “Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models”. In: *Bioinformatics* 31.9, pp. 1420–1427.
- Sivic, Josef et al. (2005). “Discovering objects and their location in images”. In: *Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1*. Vol. 1. IEEE, pp. 370–377.
- Tipping, Michael E and Christopher M Bishop (1999). “Probabilistic principal component analysis”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3, pp. 611–622.
- Yu, Shipeng et al. (2005). “A probabilistic clustering-projection model for discrete data”. In: *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, pp. 417–428.

Appendix

MMPCA: two simulation scenarios

Fixed setting:

$$p = 1000, \qquad K = 6, \qquad d = 4, \qquad \boldsymbol{U}^\star, \qquad \boldsymbol{x}^\star, \qquad \forall i,\, c_i = 400$$

Scenario 1: noisy structure $n = 400$

$$\boldsymbol{x}_{\epsilon,k} = (1-\epsilon)\boldsymbol{x}_k^\star + \frac{\epsilon}{d}\underbrace{(1,\ldots,1)}_{d}^\top, \quad \epsilon \in [0,1]$$

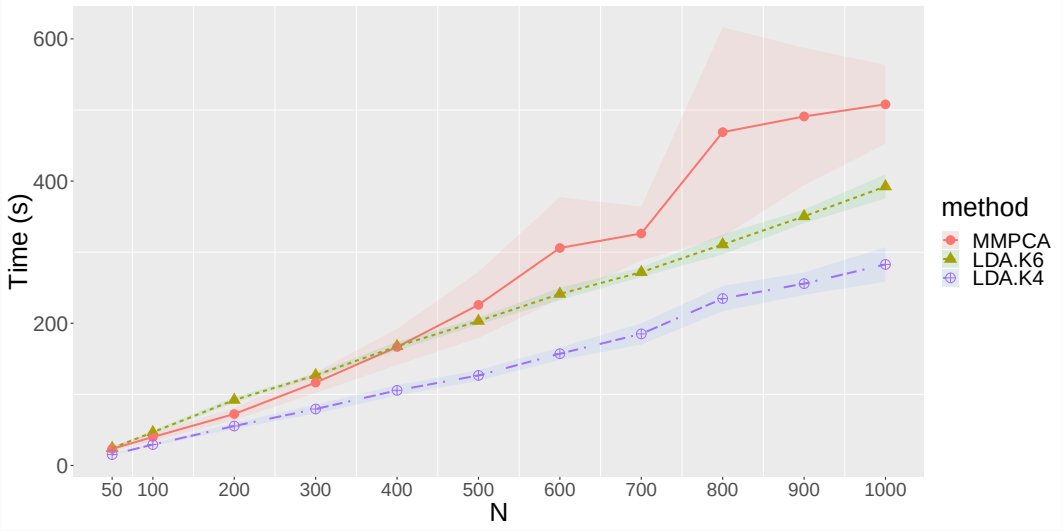
- $\epsilon = 0 \to \boldsymbol{x}_{0,k} = \boldsymbol{x}_k^\star$ distribution across topics
- $\epsilon = 1 \to \boldsymbol{x}_{1,k}$ uniform across topics (no cluster structure)

Scenario 2: small-sample size $\epsilon = 0.2$

$$n = r \times p, \quad r \in [0,1]$$

Metric: *adjusted Rand index* (ARI) the higher, the better

Time complexity



A detour via variational inference

Let $\boldsymbol{Z}$ be known and fixed

For any probabilistic distribution q on $\boldsymbol{x}$,

$$\log p(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{\pi}, \boldsymbol{U}) = \log \int_{\boldsymbol{x}} p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{x} \mid \boldsymbol{\pi}, \boldsymbol{U}) \frac{q(\boldsymbol{x})}{q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \tag{1}$$

$$\geq_{\text{(Jensen)}} \mathcal{J}(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{U}, q) \coloneqq \mathbb{E}_{\boldsymbol{x} \sim q} \left[\log p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{x} \mid \boldsymbol{\pi}, \boldsymbol{U}) \right] + \mathrm{H}(q) \tag{2}$$

Equation (1) is intractable but if $q = \prod_k q(\boldsymbol{x}_k)$

Mean-field variational inference

Maximizing Equation (2) w.r.t $(q, \boldsymbol{\pi}, \boldsymbol{U})$ is tractable by iterating over

1. (VE-step) Compute $q^{(t+1)} \in \arg\max_q \mathcal{J}(\boldsymbol{Z}, \boldsymbol{\pi}^{(t)}, \boldsymbol{U}^{(t)}, q)$.
2. (M-step) Compute $(\boldsymbol{\pi}^{(t+1)}, \boldsymbol{U}^{(t+1)}) = \arg\max_{\boldsymbol{\pi}, \boldsymbol{U}} \mathcal{J}(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{U}, q^{(t+1)})$