Mixture of multinomial PCA: towards a joint analysis of histopathological texts and images

Nicolas Jouvin

Supervisors: P. Latouche, C. Bouveyron & A. Livartowski

January 2019

Modeling textual and image data

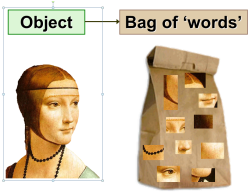
Heterogeneous data $\Longrightarrow$ Joint representation ?

Our approach: vector of **count** data.

- Text: BoW model, number of occurrences of each word,

the dog is on the table

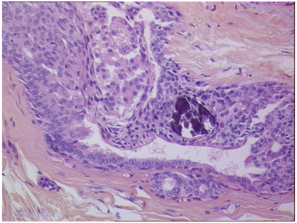
- Image: Visual codebook & quantization $\rightarrow$ number of occurrences of each visual word,

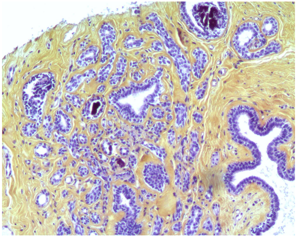


$\Longrightarrow$ text/image correspondance.

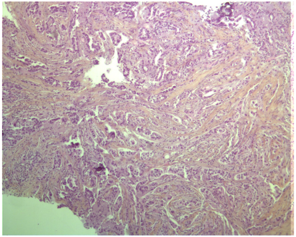
Histology images

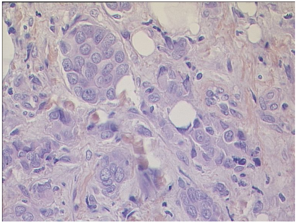
Benign





Cancer





Count data clustering

- Exploratory analysis: finding groups of medical samples (images + texts) that make sense clinically, in an unsupervised way,
- *Topic modeling*: a medical sample is a mixture of different topics, e.g. benign lesion on a portion of a cancer image.
- We propose a new model: The Mixture of multinomial PCA,
  - ➢ Based on the Latent Dirichlet Allocation [Blei et al., 2003],
  - ➢ New latent variable Y_d: cluster membership of a document,

Generative model for MMPCA

- Parameters : π, β
- $\forall q = 1, \ldots, Q, \quad \theta_q \sim \mathrm{Dir}_K(\alpha)$,
- $\forall d$ do :
  1. $Y_d \sim \mathcal{M}_Q(1;\, \pi)$
  2. $\forall n = 1, \ldots, N_d$
     (a) $Z_{dn} \mid \{Y_{dq} = 1\} \sim \mathcal{M}_K(1;\, \theta_q)$
     (b) $w_{dn} \mid \{Z_{dnk} = 1\} \sim \mathcal{M}_V(1;\, \beta_k)$

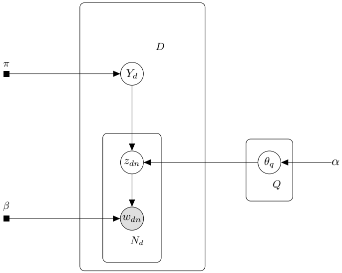


Figure: Mixture of MPCA

Inference

Complete log-likelihood :

$$\log \mathrm{p}(W, Y \mid \beta, \pi) = \underbrace{\log \mathrm{p}(W \mid Y, \theta, \beta)}_{(1)} + \log \mathrm{p}(Y \mid \pi)$$

Where:

$$(1) \iff \text{LDA on } Q \text{ aggregated meta-documents w.r.t. } Y$$

Problem: untractable posterior as in standard LDA. Solved by

- Variational Inference:

$$\mathrm{R}(\theta, Z \mid \gamma, \phi) = \mathrm{R}(\theta \mid \gamma)\ \mathrm{R}(Z \mid \phi)\,,$$

- Lower bound (Classification ELBO) maximisation:

$$\log \mathrm{p}(W, Y \mid \beta, \pi) \geq \mathcal{L}(R(\cdot); \pi, \beta, Y).$$

Classification-VEM

- Discrete optimisation in Y: Q^D solutions.
- Branch&Bound like algorithm: joint inference and clustering.

1. Initialisation avec Y, π, β.
2. Soit d t.q. $Y_d^{(k)} = \tilde{q}$. Alors, $\forall q \neq \tilde{q}$,
   (a) Try swap : $Y_d^{(k)} = q$
   (b) Temporary VE step : $\mathrm{R}_{\tilde{q}\to q} = \arg\max_{\mathrm{R}} \mathcal{L}(\mathrm{R}; \pi^{(k)}, \beta^{(k)}, Y)$
3. Choose best swap : $q^\star = \underset{q\in\{1,\dots,Q\}}{\arg\max} \mathcal{L}(\mathrm{R}_{\tilde{q}\to q}; \pi^{(k)}, \beta^{(k)}, Y_{-d}^{(k)}, Y_d^{(k)} = q)$
4. Update :

$$
\begin{aligned}
Y_d^{(k+1)} &= q^\star \\
R^{(k+1)} &= \mathrm{R}_{\tilde{q}\to q^\star} \\
(\pi^{(k+1)}, \beta^{(k+1)}) &= \underset{(\pi,\beta)}{\arg\max} \mathcal{L}(R^{(k+1)}, \pi, \beta, Y^{(k+1)}) \\
k &= k+1
\end{aligned}
$$

Thanks for your attention.

All remarks and suggestions are more than welcome !

Bibliography

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent dirichlet allocation.
Journal of machine Learning research, 3(Jan):993–1022.

Buntine, W. (2002).
Variational extensions to em and multinomial pca.
In *European Conference on Machine Learning*, pages 23–34. Springer.

sources

Images: https://machinelearnings.co
http://people.csail.mit.edu/torralba/shortCourseRLOC/ : Slides ICCV 2005
https://www.mathworks.com/help/vision/ug/image-classification-with-bag-of-visual-words.html