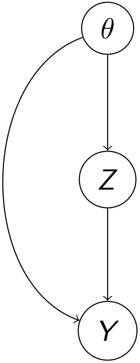
Présentation des travaux en statistiques bayésiennes

- Membres impliqués à Lille 3
  - Ophélie Guin
  - Aurore Lavigne
- Thèmes de recherche
  - Modélisation hiérarchique temporelle, spatiale, et spatio-temporelle
  - Classification
  - Modélisation non paramétrique
- Domaines d'applications
  - Environnement
  - Santé - Epidémiologie
  - Social

Modèles hiérarchiques temporels, spatiaux et spatio-temporels

Définis par une succession de distributions conditionelles

1. Le modèle d'observation $[Y|Z]$
2. Le processus spatio-temporel $[Z|\theta]$
3. Les priors $[\theta]$



Paramètres

Processus ST

Données

Publications

- Guin, O., Bel, L., Parent, E., Eckert, N., (2017) Bayesian non parametric modelling for extreme avalanches. *(En préparation)*
- Lavigne, A., Eckert, N., Bel, L., Deschâtre, M., et Parent, E. (2016). Modelling the spatio-temporal repartition of right-truncated data: an application to avalanche runout altitudes in Hautes-Savoie. SERRA, 1-16.

Classification bayésienne

- Elicitation de prior en classification et composante spatiale
  Lavigne, A., Eckert, N., Bel, L. et Parent, E., (2015). Adding expert contributions to the spatiotemporal modelling of avalanche activity under different climatic influences, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(4), 651–671.

- Classification d'unités spatiales
  Liverani, S., Lavigne, A., et Blangiardo, M. (2016). Modelling collinear and spatially correlated data. *Spatial and Spatio-temporal Epidemiology.*

- Classification de bureaux de votes selon leur comportement
  Guin, O., Bar-Hen, A., *Travail en cours*

Modélisation non paramétrique

- Processus temporel : splines de lissage traitées comme GMRF
  Guin, O., Naveau, P., et Boreux, J-J. (2017). Extracting a common signal in tree ring widths with a semi-parametric Bayesian hierarchical model. *JABE*

- Processus de Dirichlet
  cf présentation

Uncertainty quantification of a partition coming from a DPMM

Aurore Lavigne^1 et Silvia Liverani^2

1. Université Lille 3, LEM CNRS UMR 8179, 2. Brunel University London

1ère Journée lilloise de Proba-Stat - Vendredi 6 Janvier 2016

Bayesian clustering

- Clustering: group together subjects which are close from each other (k-means, hierarchical clustering).

Bayesian clustering

- Clustering: group together subjects which are close from each other (k-means, hierarchical clustering).

- Model-based clustering: data are modelled with a mixture distribution
  - take benefits of selection model tools to tackle sensitive questions (number of clusters)
  - probabilistic framework for dealing with uncertainty

Bayesian clustering

- Clustering: group together subjects which are close from each other (k-means, hierarchical clustering).
- Model-based clustering: data are modelled with a mixture distribution
  - take benefits of selection model tools to tackle sensitive questions (number of clusters)
  - probabilistic framework for dealing with uncertainty
- Bayesian model-based clustering:
  - Prior on the mixture distribution
  - Dirichlet process: infinite number of components

Outline

1. Introduction
2. Dirichlet process for Bayesian clustering
   - Dirichlet process
   - Dirichlet process in a hierarchical model : Dirichlet process mixture model
3. Selection of an estimate partition
4. Uncertainty of a given a partition
5. Conclusion

Outline

1. Introduction

2. Dirichlet process for Bayesian clustering
   - Dirichlet process
   - Dirichlet process in a hierarchical model : Dirichlet process mixture model

3. Selection of an estimate partition

4. Uncertainty of a given a partition

5. Conclusion

Dirichlet process

Let $(\Omega, \mathcal{F}, \mathbf{P})$ a probability space. The probability measure $\mathbf{P}$ follows a dirichlet process with concentration parameter α and base distribution $\mathbf{P}_{\Theta_0}$ parametrised by Θ_0, $\mathbf{P} \sim DP(\alpha, \mathbf{P}_{\Theta_0})$ if

$$(\mathbf{P}(A_1), \mathbf{P}(A_2), \cdots, \mathbf{P}(A_r)) \sim Dirichlet(\alpha \mathbf{P}_{\Theta_0}(A_1), \cdots, \alpha \mathbf{P}_{\Theta_0}(A_r))$$

for all $A_1, A_2, \cdots, A_r \in \mathcal{F}$ such that $A_i \cap A_j = \emptyset$ and $\cup_{j=1}^{r} A_j = \Omega$.

DP prediction rule (Blackwell and MacQueen 1973)

Let $(\tilde{\Theta}_1, \tilde{\Theta}_2, \cdots, \tilde{\Theta}_n)$ be i.i.d. sampled from probability $\mathbf{P} \sim DP(\alpha, \mathbf{P}_{\Theta_0})$. The conditional distribution of $\tilde{\Theta}_n$ given $(\tilde{\Theta}_1, \tilde{\Theta}_2, \cdots, \tilde{\Theta}_{n-1})$ is given by

$$(\tilde{\Theta}_n | \tilde{\Theta}_1, \tilde{\Theta}_2, \cdots, \tilde{\Theta}_{n-1}) \sim \left(\frac{\alpha}{\alpha + n - 1}\right) \mathbf{P}_{\Theta_0} + \left(\frac{1}{\alpha + n - 1}\right) \sum_{i=1}^{n-1} \delta_{\tilde{\Theta}_i}$$

where $\delta_{\tilde{\Theta}_i}$ is the Dirac measure at $\tilde{\Theta}_i$.

- Probability measure $\mathbf{P}$ is discrete.
- α tunes the number of distincts $\tilde{\Theta}_i$.

The DP process defines a partition of $\mathbb{N}$.

- Let $k_n \leq n$ be the number of distincts values in $(\tilde{\Theta}_1, \tilde{\Theta}_2, \cdots, \tilde{\Theta}_n)$.
- Rename $(\Theta_1, \Theta_2, \cdots, \Theta_{k_n})$ the k_n distincts values of $(\tilde{\Theta}_1, \tilde{\Theta}_2, \cdots, \tilde{\Theta}_n)$.
- Let Z_i be the allocation variable, meaning that Z_i is the label of observation i, $\tilde{\Theta}_i = \Theta_{Z_i}$.

The DP process defines a partition of $\mathbb{N}$.

- Let $k_n \leq n$ be the number of distincts values in $(\tilde{\Theta}_1, \tilde{\Theta}_2, \cdots, \tilde{\Theta}_n)$.
- Rename $(\Theta_1, \Theta_2, \cdots, \Theta_{k_n})$ the k_n distincts values of $(\tilde{\Theta}_1, \tilde{\Theta}_2, \cdots, \tilde{\Theta}_n)$.
- Let Z_i be the allocation variable, meaning that Z_i is the label of observation i, $\tilde{\Theta}_i = \Theta_{Z_i}$.

Relying on the DP prediction rule, the conditional distribution of Z_n given $\mathbf{Z}_{-n} = (Z_1, Z_2, \cdots, Z_{n-1})$ is multinomial with probabilities

$$P(Z_n = c|\mathbf{Z}_{-n}) = \begin{cases} \frac{\sum_{i=1}^{n-1} 1_{Z_i = c}}{\alpha + n - 1} \text{ if } c = 1, \cdots, k_{n-1} \\ \\ \frac{\alpha}{\alpha + n - 1} \text{ if } c = k_{n-1} + 1 \end{cases}$$

Stick breaking construction, *Sethuraman (1994)*

If

$$\mathbf{P} = \sum_{c=1}^{\infty} \psi_c \delta_{\Theta_c}, \text{ where}$$

$$\Theta_c \sim \mathbf{P}_{\Theta_0} \text{ i.i.d. for } c \in \mathbb{Z}^+, \text{ and}$$

$$\begin{aligned} &\psi_c = V_c \prod_{l<c}(1 - V_l) \text{ for } c \in \mathbb{Z}^+ \setminus \{1\}, \\ &\psi_1 = V_1, \text{ and} \\ &V_c \sim \mathrm{Beta}(1, \alpha) \text{ i.i.d. for } c \in \mathbb{Z}^+ \end{aligned}$$

then, $\mathbf{P} \sim DP(\alpha, \mathbf{P}_{\Theta_0})$.

Dirichlet Process mixture model

- Conditionally to latent variables $(\tilde{\Theta}_1, \tilde{\Theta}_2, \cdots, \tilde{\Theta}_n)$, observed data $\mathbf{D} = (D_1, D_2, \ldots, D_n)$ follow a parametric density of the form $f(\cdot|\Theta)$

$$D_i|\tilde{\Theta}_i \sim f(D_i|\tilde{\Theta}_i) \text{ i.i.d.}.$$

- Latent variables $(\tilde{\Theta}_1, \tilde{\Theta}_2, \cdots, \tilde{\Theta}_n)$ follow a Dirichlet process

$$(\tilde{\Theta}_1, \tilde{\Theta}_2, \cdots, \tilde{\Theta}_n) \sim DP(\alpha, \mathbf{P}_{\Theta_0})$$

Dirichlet Process mixture model (2)

- $\mathbf{D} = (D_1, D_2, \dots, D_n)$ follow an infinite mixture distribution where component c of the mixture is $f(\cdot|\Theta_c)$

$$D_i|\Theta_1, \Theta_2, \dots \sim \sum_{c=1}^{\infty} \psi_c f(D_i|\Theta_c) \text{ for } i = 1, 2, \dots, n$$

- Weights ψ_c are sampling from the stick breaking process

$$\begin{aligned} \psi_c &= V_c \prod_{l<c} (1 - V_l) \text{ for } c \in \mathbb{Z}^+ \setminus \{1\}, \\ \psi_1 &= V_1, \text{ and} \\ V_c &\sim \text{Beta}(1, \alpha) \text{ i.i.d. for } c \in \mathbb{Z}^+ \end{aligned}$$

- Location parameters Θ_c are sampled from the baseline distribution

$$\Theta_c \sim P_{\Theta_0} \text{ i.i.d. for } c \in \mathbb{Z}^+$$

Latent variable

We introduce Z_i the latent allocation variable of observation D_i.

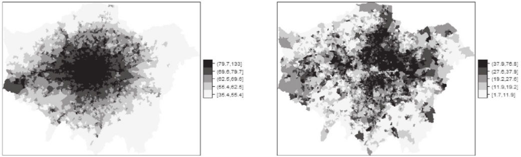
- $Z_i = c$ if observation i belongs to cluster c.
- We denote $\mathbf{\Theta} = (\Theta_1, \Theta_2, \cdots)$

$$D_i | Z_i, \mathbf{\Theta} \sim f(D_i | \Theta_{Z_i}) \text{ for } i = 1, 2, \ldots, n$$

- and

$$\mathbf{P}(Z_i = c) = \psi_c$$

Example: clustering London areas from multiple deprivation index and air pollution measurements, (Liverani et al, 2015)



(a) NO$_x$ concentration for 2003-2010

(b) IMD score, 2010

Fig. 1. Quintilesof the NO$_x$ concentration (average 2003–2010) and of IMD score (2010) at LSOA level in Greater London.

11 clusters

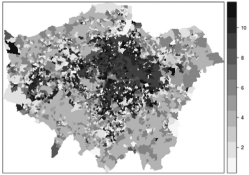


Fig. 3. Geographical representation of the eleven clusters of the areas in Greater London identified by profile regression. The colours reflect the mean of the observed pollution levels, with dark grey identifying the most polluted clusters and light grey the least polluted clusters.

Outline

1. Introduction

2. Dirichlet process for Bayesian clustering
   - Dirichlet process
   - Dirichlet process in a hierarchical model : Dirichlet process mixture model

3. **Selection of an estimate partition**

4. Uncertainty of a given a partition

5. Conclusion

Inference

- Gibbs algorithms are developed: Escobar and West (1995), Neal (2000), Papaspiliopoulos and Roberts (2008)
  - Difficulty to explore the space of partitions
  - Difficulty to assess convergence
- They provide a sample of partitions, and location parameters $\mathbf{\Theta}$, drawn in their posterior distribution
- Various methods to choose one 'optimal' partition from this set:
  1. Maximum a posteriori
  2. Based on loss criteria within or without the sample set.
  3. Clustering from distance within pairs.

Maximum a posteriori

$$\hat{\mathbf{Z}} = \underset{Z \in \mathcal{Z}}{\mathrm{argmax}} p(Z|\mathbf{D}) = \underset{Z \in \mathcal{Z}}{\mathrm{argmax}} p(\mathbf{D}|Z)p(Z)$$

where $\mathcal{Z}$ is the set of partitions of $\{1, 2 \cdots, n\}$.

Difficulties

- Cardinal of $\mathcal{Z}$ is huge, given by the Bell's number.
- $\hat{\mathbf{Z}}$ is not necessary in the sampled set.
- Various methods for selecting the MAP : trade off between computational time and approximation.

Methods based on loss criteria

AIM: Selecting $\hat{\mathbf{Z}}$ which minimizes the expected posterior loss

$$\hat{\mathbf{Z}} = \underset{Z \in \mathcal{Z}}{\operatorname{argmin}} \sum_{Z' \in \mathcal{Z}} \ell(Z, Z') p(Z' | \mathbf{D})$$

where $\ell(Z, Z')$ is a given loss between partitions Z and Z'.

Classical losses

- 0-1 loss: $\ell(Z, Z') = \mathbf{1}_{\mathbf{Z} \neq \mathbf{Z}'}$ => MAP
- Binder's loss : $\ell(\mathbf{Z}, \mathbf{Z}') = \sum_{i<j} \mathbf{1}_{Z_i = Z_j} \mathbf{1}_{Z'_i \neq Z'_j} + \mathbf{1}_{Z_i \neq Z_j} \mathbf{1}_{Z'_i = Z'_j}$
- Variation of information: $\ell(\mathbf{Z}, \mathbf{Z}') = H(\mathbf{Z}) + H(\mathbf{Z}') - 2I(\mathbf{Z}, \mathbf{Z}')$ (Wade and Ghahramani, 2015)

Clustering from distance within pairs

Distance within pairs d_{ij} is defined as

$$d_{ij} = 1 - \pi_{ij} = 1 - p(Z_i = Z_j | \mathbf{D})$$

A deterministic classification is post-processed using distances d_{ij}

- Hierarchical clustering (Complete linkage) (Medvedovic et al., 2004)
- Partioning Aroung Medoids (Molitor et al., 2010)

Drawbacks: Two clusterings.

Problems:

- Estimated partitions may be very different according the method chosen.
- Interpretation is based on the estimated partition only.
- Is the number of components in the mixture is relevant a relevant estimator of the number of clusters?
- All of the variability of the partitions is forgotten

Objective:

- Propose a method to represent and take into account the uncertainty about the partition.

Outline

1. Introduction

2. Dirichlet process for Bayesian clustering
    - Dirichlet process
    - Dirichlet process in a hierarchical model : Dirichlet process mixture model

3. Selection of an estimate partition

4. Uncertainty of a given a partition

5. Conclusion

Uncertainty measure in an estimated finite mixture model

Consider the case where:

- the number of clusters K is known and finite,
- $\Theta_1, \Theta_2, \cdots, \Theta_K$ and $\psi_1, \psi_2, \cdots, \psi_K$ are known and fixed.

The certainty that an observation D_{n+1} belongs to cluster c, is given by the Bayes theorem:

$$\mathbf{P}(Z_{n+1} = c | D_{n+1}) = \frac{\psi_c f(D_{n+1} | \Theta_c)}{\sum_{c=1}^{K} \psi_c f(D_{n+1} | \Theta_c)}.$$

However, because of label switching and the infinite number of clusters, we cannot estimate Θ_c and ψ_c.

Objective

We consider a estimated partition $\mathbf{Z}^*$ given by one of the post-process methods. This partition has k clusters.

Objective: write the predictive distribution $P(D_{n+1}|\mathbf{D}_n)$ as a finite mixture of k densities.

Predictive distribution in a DPMM

Escobar et West, (1995) : given a partition $\mathbf{Z}$, and the location of parameters in the groups *i.e.* $\mathbf{\Theta} = (\Theta_1, \Theta_2, \ldots)$, the predictive distribution of a new observation D_{n+1} is

$$P(D_{n+1}|\mathbf{Z}, \mathbf{\Theta}) = \frac{\alpha}{\alpha + n} \overbrace{\int f(D_{n+1}|\Theta_{n+1}) G_0(\Theta_{n+1}) d\Theta_{n+1}}^{f_0(D_{n+1})} + \frac{1}{\alpha + n} \sum_{c: n_c > 0} n_c f(D_{n+1}|\Theta_c).$$

where n_c is the number of individuals belonging to cluster c.
A new data can be clustered

- in the clusters defined by observations $\mathbf{D}_n = (D_1, \ldots, D_n)$
- in a new cluster in which the parameter $\Theta_{n+1} \sim G_0()$.

Then to derive the uncertainty on a partition, we consider the predictive distribution

$$\begin{aligned} P(D_{n+1}|\mathbf{D}_n) &= \frac{\alpha}{\alpha+n} f_0(D_{n+1}) \\ &+ \frac{1}{\alpha+n} \sum_{\mathbf{Z}\in\mathcal{Z}} \sum_{c:n_c(\mathbf{Z})>0} n_c f(D_{n+1}|\mathbf{Z}, \{D_i : Z_i = c\}) p(\mathbf{Z}|\mathbf{D}_n) \end{aligned}$$

where $f(D_{n+1}|\mathbf{Z}, \{D_i : Z_i = c\})$ is the predictive distribution in the cluster c of partition Z.

$$f(D_{n+1}|\mathbf{Z}, \{D_i : Z_i = c\}) = \int f(D_{n+1}|\Theta_c) p(\Theta_c|\mathbf{Z}, \{D_i : Z_i = c\}) d\Theta_c$$

We then make appeared the weights n_l/n:

$$\begin{aligned} P(D_{n+1}|\mathbf{D}_n) &= \sum_{l=1}^{k} \frac{n_l}{n}[\frac{\alpha}{\alpha+n} f_0(D_{n+1}) \\ &\quad + \frac{1}{\alpha+n} \sum_{\mathbf{Z}\in\mathcal{Z}} \sum_{c:n_c(\mathbf{Z})>0} n_{cl}(\mathbf{Z}) f(D_{n+1}|\mathbf{Z}, \{D_i : Z_i = c\}) p(\mathbf{Z}|\mathbf{D}_n)] \end{aligned}$$

- n_l the number of subjects belonging to cluster l of $\mathbf{Z}^*$
- $n_{cl}(\mathbf{Z})$ the number of subjects belonging to cluster l of $\mathbf{Z}^*$ and to cluster c of $\mathbf{Z}$

By denoting $\tilde{f}_l(D_{n+1})$ the component l of the mixture,

$$\begin{aligned}\tilde{f}_l(D_{n+1}) &= \frac{\alpha}{\alpha+n} f_0(D_{n+1}) \\ &+ \frac{1}{\alpha+n} \sum_{\mathbf{Z}\in\mathcal{Z}} \sum_{c:n_c(\mathbf{Z})>0} n_{cl}(\mathbf{Z}) f(D_{n+1}|\mathbf{Z}, \{D_i : Z_i = c\}) p(\mathbf{Z}|\mathbf{D}_n)\end{aligned}$$

we have

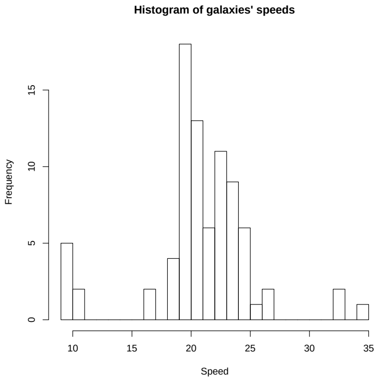
$$P(D_{n+1}|\mathbf{D}_n) = \sum_{l=1}^{k} \frac{n_l}{n} \tilde{f}_l(D_{n+1})$$

We note that:

- $\tilde{f}_l(D_{n+1})$ is not parametric.
- $\tilde{f}_l(D_{n+1})$ do not depend only on observations clustered in cluster l of partition $\mathbf{Z}^*$.

Application to velocity galaxy data (Rouder, 1990)

- $n = 82$ measures of galaxy velocity
- Is the distribution is multimodal?
- How many modes they are? Clusters?



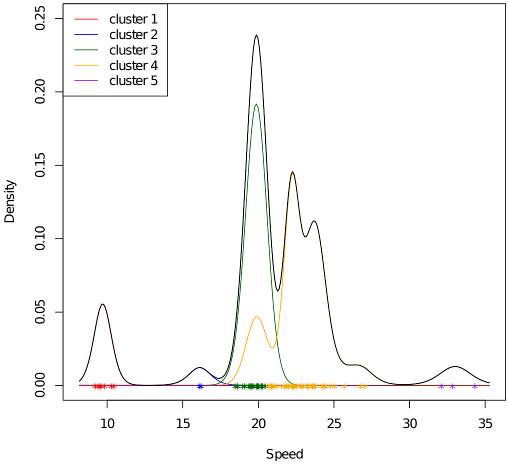
Case of Gaussian multivariate mixture

- $D_i|Z_i, \mathbf{\Theta} \sim \mathcal{N}(\mu_{Z_i}, \Sigma_{Z_i})$ with $\Theta_c = (\mu_c, \Sigma_c)$
- Use of the conjugated normal inverse Wishart prior $(NIW(\mu_0, \nu_0, \kappa_0, R_0))$

$$\begin{cases} \Sigma_c \sim \mathcal{IW}(\kappa_0, R_0) \\ \mu_c|\Sigma_c \sim \mathcal{N}(\mu_0, \frac{1}{\nu_0}\Sigma_c) \end{cases}$$

- 300000 iterations of a Gibbs algorithm (Rpackage PReMiuM), one out of 30 is recorded
- the 'optimal' partition $\mathbf{Z}^*$ is obtained with algorithm PAM.

Case study

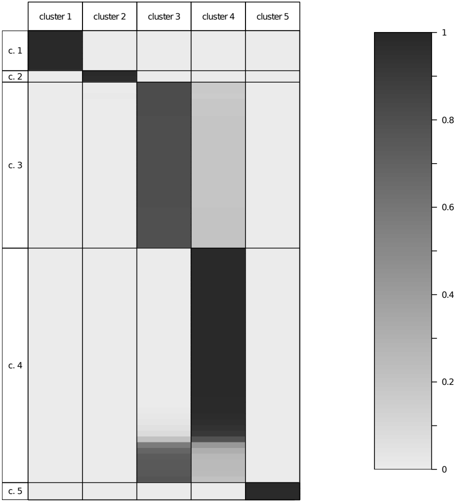

cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
Density
0.00
0.05
0.10
0.15
0.20
0.25
10
15
20
25
30
35
Speed

Uncertainty about the clusters

p_i^l probability that observation i belong to cluster l

$$p_i^l = \frac{n_l \tilde{f}_l(D_i)}{\sum_{l=1}^{k} n_l \tilde{f}_l(D_i)}$$

Graphical representation of clusters uncertainty



Conclusion

We provide a method for assessing uncertainty

- given an optimal partition
- based on a mixture of non parametric densities

Others methods to assess the uncertainty of a partition in DPMM

- Matrix of similarity within pairs
- Marginal partition posterior are only useful for comparing partitions
- Credible balls (Wade and Ghahramani, 2014)

On going work: influence of an observation on a cluster of the optimal partition.

References

- Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixture, *Journal of the American Statistical Association 90*(430), 577– 588.
- Liverani, S., D. I. Hastie, and S. Richardson (2014). PReMiuM: An R Package for Profile Regression Mixture Models using Dirichlet Processes, *Forthcoming in the Journal of Statistical Software. Preprint available at arXiv:1303.2836.*
- Liverani, S., A. Lavigne, and M. Blangiardo (2016). Modelling collinear and spatially correlated data, *Spatial and Spatio-temporal Epidemiology*.
- Neal, R. M. (2000, June). Markov chain sampling methods for Dirichlet process mixture models, *Journal of Computational and Graphical Statistics 9*(2), 249.
- Papaspiliopoulos, O. and G. O. Roberts (2008, January). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models, *Biometrika 95*(1), 169–186.
- Wade, S. and Z. Ghahramani (2015). Bayesian cluster analysis: Point estimation and credible balls, *arXiv preprint arXiv:1505.03339*.