Bayesian Nonparametric Priors for Hidden Markov Random Fields

Florence Forbes

florence.forbes@inria.fr

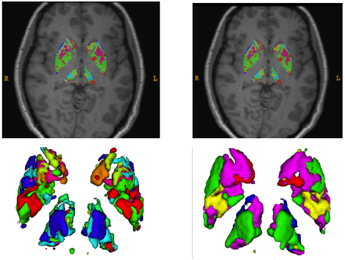
Inria Grenoble Rhône-Alpes & University Grenoble Alpes
Laboratoire Jean Kuntzmann
Statify team

joint work with Hongliang Lü, Julyan Arbel, Jean-Baptiste Durand

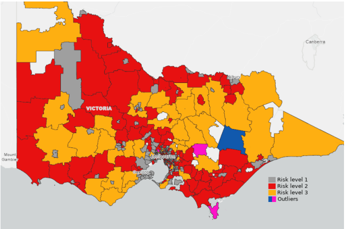
December 2020

Motivation: clustering spatial data

3D brain MRI segmentation, Risk mapping etc.



10 clusters vs **6 clusters**



Car crash risk in the region of Melbourne

Challenges for unsupervised image segmentation: blur, **noise**, color/contrast imperfection, partial volume effect (large slice thickness), anatomic variability and complexity, **number of segments**...
Challenges for spatial risk mapping: accounting for neighborhood, decide on risk level thresholds, **number of levels**...

⇒ **Design tractable BNP-MRF priors for structured data**: no commitment to an arbitrary number of clusters (BNP) and dependence modelling (Markov Random Field)

Extensions of **Dirichlet Process mixture model with spatial regularization**

Outline of the talk

1. Bayesian non parametric (BNP) priors: Dirichlet process (DP)
2. Spatially-constrained mixture model: DP-Potts mixture model
   - Finite mixture model
   - Bayesian finite mixture model
   - DP mixture model
   - DP-Potts mixture model
3. Inference using variational approximation
4. Some image segmentation results
5. Probabilistic properties of BNP-MRF priors
6. Conclusion and future work

BNP priors: Dirichlet (DP), Pitman-Yor (PY) process, etc.

The Dirichlet process (DP) is a central Bayesian nonparametric (BNP) prior[1].

Definition (Dirichlet process)

A Dirichlet process on the space $\mathcal{Y}$ is a **random process** G characterized by a concentration parameter α and a base distribution G_0 such that for any finite partition $\{A_1, \dots, A_p\}$ of $\mathcal{Y}$, the random vector $(G(A_1), \dots, G(A_p))$ is Dirichlet distributed:

$$(G(A_1), \dots, G(A_p)) \sim \mathrm{Dir}(\alpha G_0(A_1), \dots, \alpha G_0(A_p))$$

Notation: $G \sim \mathrm{DP}(\alpha, G_0)$

The DP is the infinite-dimensional generalization of the Dirichlet distribution.

[1] Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. The Annals of Statistics, 1(2):209–230.

Dirichlet process (DP) construction

A DP prior G can be constructed using three methods:

- The Blackwell-MacQueen urn scheme
- The Chinese Restaurant Process
- The Stick-Breaking construction

[2] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. Statistica Sinica, 4:639-650.

Dirichlet process (DP) construction

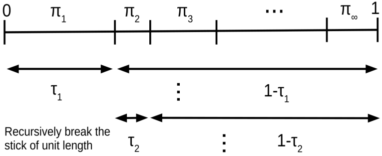
A DP prior G can be constructed using three methods:

- The Blackwell-MacQueen urn scheme
- The Chinese Restaurant Process
- The Stick-Breaking construction

The DP has almost surely discrete realizations[2]:

$$G = \sum_{k=1}^{\infty} \pi_k(\tau)\ \delta_{\theta_k^*}$$

where $\theta_k^* \stackrel{\text{iid}}{\sim} G_0$ and $\pi_k(\tau) = \tau_k\ \prod_{l<k}(1-\tau_l)$ with $\tau_k \stackrel{\text{iid}}{\sim} Beta(1,\alpha)$.



[2] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. Statistica Sinica, 4:639-650.

Spatially-constrained mixture model: DP-Potts mixture

Clustering/segmentation: **Finite mixture models** assume data are generated by a finite sum of probability distributions:

$$\mathbf{y} = (\mathbf{y_1}, ..., \mathbf{y_N}) \text{ with } \mathbf{y_i} = (y_{i1}, ..., y_{iD}) \in \mathbb{R}^D \ i.i.d$$

$$p(\mathbf{y}_i|\theta^*, \pi) = \sum_{k=1}^{K} \pi_k F(\mathbf{y}_i|\theta_k^*)$$

where

- $\theta^* = (\theta_1^*, ..., \theta_K^*)$ and $\pi = (\pi_1, ..., \pi_K)$ with θ^* class parameters and π mixture weights with $\sum_{i=1}^{K} \pi_i = 1$.
- θ^* and π can be estimated using an EM algorithm.

Equivalently

- $G = \sum_{k=1}^{K} \pi_k \ \delta_{\theta_k^*}$ non random (π_k, θ_k^*'s are unknown but fixed)
- $\theta_i \sim G$ (*ie* θ_i takes one of the θ_k^* values) and then $\mathbf{y}_i|\theta_i \sim F(\mathbf{y}_i|\theta_i)$.

Bayesian finite mixture model

In a Bayesian setting, a prior is placed over $\theta^* = (\theta_1^* \dots \theta_K^*)$ and $\pi = (\pi_1 \dots, \pi_K)$

Thus, the posterior distribution of parameters given the observations is

$$p(\theta^*, \pi | \mathbf{y}) \propto p(\mathbf{y} | \theta^*, \pi) p(\theta^*, \pi)$$

To generate a data point within a **Bayesian finite mixture model**:

- $\theta_k^* \sim G_0$
- $\pi_1, ..., \pi_K \sim \mathrm{Dir}(\alpha/K, ..., \alpha/K)$
- $G = \sum_{k=1}^K \pi_k \delta_{\theta_k^*}$ is now a random measure
- $\theta_i | G \sim G$, which means $\theta_i = \theta_k^*$ with probability π_k
- $\mathbf{y_i} | \theta_i \sim F(\mathbf{y_i} | \theta_i)$

Bayesian finite mixture model

In a Bayesian setting, a prior is placed over $\theta^* = (\theta_1^* \dots \theta_K^*)$ and $\pi = (\pi_1 \dots, \pi_K)$

Thus, the posterior distribution of parameters given the observations is

$$p(\theta^*, \pi|\mathbf{y}) \propto p(\mathbf{y}|\theta^*, \pi)p(\theta^*, \pi)$$

To generate a data point within a **Bayesian finite mixture model**:

- $\theta_k^* \sim G_0$
- $\pi_1, ..., \pi_K \sim \mathrm{Dir}(\alpha/K, ..., \alpha/K)$
- $G = \sum_{k=1}^{K} \pi_k \delta_{\theta_k^*}$ is now a random measure
- $\theta_i|G \sim G$, which means $\theta_i = \theta_k^*$ with probability π_k
- $\mathbf{y_i}|\theta_i \sim F(\mathbf{y_i}|\theta_i)$

Limitation:

Require specifying the number of components K beforehand.

Solution:

Assume an infinite number of components using BNP priors.

DP mixture model

From a Bayesian finite mixture model to a DP mixture model

To establish a DP mixture model, let G be a DP prior ($K \to \infty$), namely

$$G \sim \mathrm{DP}(\alpha, G_0)$$

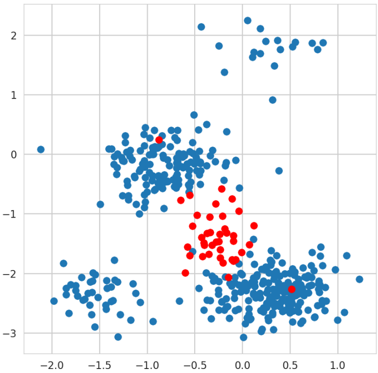
and complement it with a likelihood associated to each θ_i

To generate a data point within a **DP mixture model**:

- $G \sim \mathrm{DP}(\alpha, G_0)$
- $\theta_i | G \sim G$
- $\mathbf{y_i} | \theta_i \sim F(\mathbf{y_i} | \theta_i)$

DP mixture model

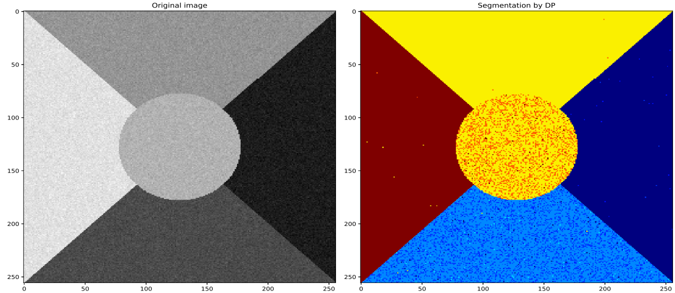
2D point clustering (unsupervised learning) based on the DP mixture model:



Let the data speak for themselves!

DP mixture model

Application to image segmentation:



Drawback:

Spatial constraints and dependencies are not considered.

Solution:

Combine the DP prior with an hidden Markov random field (HMRF).

DP-Potts mixture model

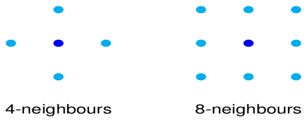
To take into account spatial information, we introduce a Potts model component:

$$M(\boldsymbol{\theta}) \propto \exp\left(\beta \sum_{i \sim j} \delta_{\theta_i = \theta_j}\right) \qquad i \text{ and } j \text{ are neighbors, eg. pixels}$$

with $\boldsymbol{\theta} = (\theta_1 \dots \theta_N)$ (associated to $\mathbf{y} = (\mathbf{y}_1 \dots \mathbf{y}_N)$) and β the interaction parameter

The DP mixture model is thus extended as:

- $G \sim \mathrm{DP}(\alpha, G_0)$
- $\boldsymbol{\theta} | M, G \sim M(\boldsymbol{\theta}) \times \prod_i G(\theta_i)$
- $\mathbf{y_i} | \theta_i \sim F(\mathbf{y_i} | \theta_i)$



DP-Potts mixture model

Other spatially-constrained BNP mixture models + inference algorithms:

- DP or PYP-Potts **partition model** + MCMC[3]
- Hemodynamic brain parcellation (DP-Potts) + PARTIAL VB[4]
- DP or PYP-Potts + Iterated Conditional Mode (ICM)[5]

Markov chain Monte Carlo (MCMC):

- Advantage: asymptotically exact
- Drawback: computationally expensive

Variational Bayes (VB):

- Advantage: much faster
- Drawback: less accurate, no theoretical guarantee

We propose **a DP-Potts mixture model** based on **a general stick-breaking construction** that allows **a natural Full VB algorithm** enabling scalable inference for large datasets and straightforward generalization to other priors (eg **PY-Potts**).

[3] Orbanz & Buhmann (2008); Xu, Caron & Doucet (2016); Sodjo, Giremus, Dobigeon & Giovannelli (2017)

[4] Albughdadi, Chaari, Tourneret, Forbes, Ciuciu (2017)

[5] Chatzis & Tsechpenakis (2010); Chatzis (2013)

DP-Potts: Stick breaking construction

Stick breaking construction of DP: $G \sim DP(\alpha, G_0)$

- $\theta_k^* | G_0 \sim G_0$
- $\tau_k | \alpha \sim \mathcal{B}(1, \alpha), k = 1, 2, \ldots$
- $\pi_k(\tau) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), k = 1, 2, \ldots$
- $G = \sum_{k=1}^{\infty} \pi_k(\tau) \delta_{\theta_k^*}$

+

- $\theta_i | G \sim G$
- $\mathbf{y_i} | \theta_i \sim F(\mathbf{y_i} | \theta_i)$

= Dirichlet Process Mixture Model (DPMM)

DP-Potts: Stick breaking construction

Stick breaking construction of DPMM

- $\theta_k^* | G_0 \sim G_0$
- $\tau_k | \alpha \sim \mathcal{B}(1, \alpha), k = 1, 2, \ldots$
- $\pi_k(\tau) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), k = 1, \ldots$
- $G = \sum_{k=1}^{\infty} \pi_k(\tau) \delta_{\theta_k^*}$
- $\theta_i | G \sim G$
- $\mathbf{y_i} | \theta_i \sim F(\mathbf{y_i} | \theta_i)$

$\Longrightarrow$

Stick breaking construction of DPMM

- $\theta_k^* | G_0 \sim G_0$
- $\tau_k | \alpha \sim \mathcal{B}(1, \alpha), k = 1, 2, \ldots$
- $\pi_k(\tau) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), k = 1, \ldots$
- $\theta_i = \theta_k^*$ with probability $\pi_k(\tau)$
- $\mathbf{y_i} | \theta_i \sim F(\mathbf{y_i} | \theta_i)$

DP-Potts: Stick breaking construction

Using assignment variables z_i defined by $z_i = k$ when $\theta_i = \theta_k^*$

DPMM view

- $\theta_k^* | G_0 \sim G_0$
- $\tau_k | \alpha \sim \mathcal{B}(1, \alpha), k = 1, 2, \ldots$
- $\pi_k(\tau) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), k = 1, \ldots$
- $\theta_i = \theta_k^*$ with probability $\pi_k(\tau)$
- $\mathbf{y_i} | \theta_i \sim F(\mathbf{y_i} | \theta_i)$ $\Longrightarrow$

Mixture/Clustering view

- $\theta_k^* | G_0 \sim G_0$
- $\tau_k | \alpha \sim \mathcal{B}(1, \alpha), k = 1, 2, \ldots$
- $\pi_k(\tau) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), k = 1, \ldots$
- $p(z_i = k | \tau) = \pi_k(\tau)$
- with $z_i = z(\theta_i) = k$ when $\theta_i = \theta_k^*$
- $\mathbf{y_i} | z_i, \theta^* \sim F(\mathbf{y_i} | \theta_{z_i}^*)$

DP-Potts: Stick breaking construction

Using assignment variables z_i

Stick breaking of DPMM

- $\theta_k^*|G_0 \sim G_0$
- $\tau_k|\alpha \sim \mathcal{B}(1, \alpha), k = 1, 2, \ldots$
- $\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l)$
- $p(z_i = k|\boldsymbol{\tau}) = \pi_k(\boldsymbol{\tau})$
- $\mathbf{y_i}|z_i, \theta^* \sim F(\mathbf{y_i}|\theta_{z_i}^*)$

Stick breaking of DP-Potts

- $\theta_k^*|G_0 \sim G_0$
- $\tau_k|\alpha \sim \mathcal{B}(1, \alpha), k = 1, 2, \ldots$
- $\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l=1}^{k-1} (1 - \tau_l)$
- $p(\mathbf{z}|\boldsymbol{\tau}, \beta) \propto \prod_i \pi_{z_i}(\boldsymbol{\tau}) \exp(\beta \sum_{i \sim j} \delta_{z_i = z_j})$

  $\mathbf{z} = \{z_1, \ldots, z_N\}$
- $\mathbf{y_i}|z_i, \theta^* \sim F(\mathbf{y_i}|\theta_{z_i}^*)$

NB: Well defined for every stick breaking construction ($\sum_{k=1}^{\infty} \pi_k = 1$) :

e.g. Pitman-Yor: $\tau_k|\alpha, \sigma \sim \mathcal{B}(1 - \sigma, \alpha + k\sigma)$

Countably infinite state space Potts model

with first and second order potentials

$$p(\boldsymbol{z}|\ \boldsymbol{\tau}, \beta) \propto \left(\prod_{i=1}^{n} \pi_{z_i}(\boldsymbol{\tau})\right) \exp\left(\beta \sum_{i\sim j} \delta_{(z_i=z_j)}\right).$$

Equivalent Gibbs representation:

$p(\boldsymbol{z}|\ \boldsymbol{\tau}, \beta,) \propto \mathrm{e}^{V(\boldsymbol{z};\boldsymbol{\tau},\beta)}$ with $V(\boldsymbol{z};\boldsymbol{\tau},\beta) = \sum_{i=1}^{n} \log \pi_{z_i}(\boldsymbol{\tau}) + \beta \sum_{i\sim j} \delta_{(z_i=z_j)}$

Hammersley–Clifford theorem still holds if we can show that $\sum_{\boldsymbol{z}} \mathrm{e}^{V(\boldsymbol{z};\boldsymbol{\tau},\beta)} < \infty$,

$$\sum_{\boldsymbol{z}} \mathrm{e}^{V(\boldsymbol{z};\boldsymbol{\tau},\beta)} \leq \left(\sum_{\boldsymbol{z}} \prod_{i=1}^{n} \pi_{z_i}\right) \mathrm{e}^{\beta\frac{n(n-1)}{2}} = \mathrm{e}^{\beta\frac{n(n-1)}{2}} < \infty$$

Inference using variational approximation

Clustering/ segmentation task:

- Estimating $\mathbf{Z}$
- while parameters $\boldsymbol{\Theta}$ unknown , eg. $\boldsymbol{\Theta} = \{\boldsymbol{\tau}, \alpha, \boldsymbol{\theta}^*\}$

Bayesian setting

Access the intractable $p(\mathbf{Z}, \boldsymbol{\Theta}|\mathbf{y}; \boldsymbol{\Phi})$ approximate as $q(\mathbf{z}, \boldsymbol{\Theta}) = q_z(\mathbf{z})q_\theta(\boldsymbol{\Theta})$

Variational Expectation-Maximization

Alternate maximization in q_z and q_θ (ϕ are hyperparameters) of the Free Energy:

$$\begin{aligned}
\mathcal{F}(q_z, q_\theta, \phi) &= E_{q_z q_\theta}\left[\log \frac{p(\mathbf{y}, \mathbf{Z}, \boldsymbol{\Theta}|\phi)}{q_z(\mathbf{z})q_\theta(\boldsymbol{\Theta})}\right] \\
&= \log p(\mathbf{y}|\phi) - KL(q_z q_\theta, p(\mathbf{Z}, \boldsymbol{\Theta}|\mathbf{y}, \phi))
\end{aligned}$$

DP-Potts Variational EM procedure

Joint DP-Potts (Gaussian) Mixture distribution

$$p(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\tau},\alpha,\boldsymbol{\theta}^*|\boldsymbol{\phi}) = \prod_{j=1}^{N} p(y_j|z_j,\boldsymbol{\theta}^*)\ p(\boldsymbol{z}|\boldsymbol{\tau},\beta)\ \prod_{k=1}^{\infty} p(\tau_k|\alpha) \prod_{k=1}^{\infty} p(\theta_k^*|\rho_k)\ p(\alpha|s_1,s_2)$$

- $p(y_j|z_j,\boldsymbol{\theta}^*) = \mathcal{N}(y_j|\mu_{z_j},\Sigma_{z_j})$ is Gaussian
- $p(\boldsymbol{z}|\boldsymbol{\tau},\beta)$ is a DP-Potts model
- $p(\tau_k|\alpha)$ is Beta $\mathcal{B}(1,\alpha)$
- $p(\theta_k^*|\rho_k) = \mathcal{NIW}(\mu_k,\Sigma_k|m_k,\lambda_k,\Psi_k,\nu_k)$ is Normal-inverse-Wishart
- $p(\alpha|s_1,s_2) = \mathcal{G}(\alpha|s_1,s_2)$ is Gamma

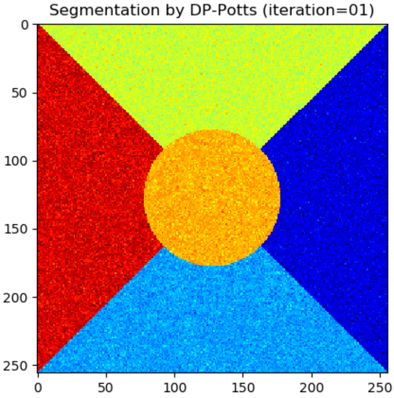
Usual truncated variational posterior, $q_{\tau_k} = \delta_1$ for $k \geq K$ (eg. $K = 40$)

$$q(\mathbf{z},\boldsymbol{\Theta}) = \prod_{j=1}^{N} q_{z_j}(z_j)\ q_\alpha(\alpha) \prod_{k=1}^{K-1} q_{\tau_k}(\tau_k) \prod_{k=1}^{K} q_{\theta_k^*}(\mu_k,\Sigma_k)$$

- E-steps: VE-Z, VE-α, VE-$\boldsymbol{\tau}$ and VE-$\boldsymbol{\theta}^*$
- M-step: $\boldsymbol{\phi}$ updating straightforward except for β

Some image segmentation results

Convergence of the VB algorithm initialized by the k-means++ clustering:



Simulated image segmentation with the PY-Potts model

- Simulated 64×64 images from a Potts model with additional Gaussian noise with varying β and K values
- Each model is simulated 100 times

(β_{true}, K_{true})	$\overline{\alpha}$	std(α)	$\overline{\sigma}$	std(σ)	$\overline{\beta}$	std(β)	cluster numbers	frequency (%)
$(0.6, 5)$	0.96	0.23	0.46	0.19	0.58	0.04	[3, 4, **5**, 6]	[1, 7, 87, 5]
$(0.8, 5)$	0.90	0.22	0.50	0.16	0.81	0.03	[4, **5**, 6]	[7, 85, 8]
$(1.0, 5)$	0.98	0.30	0.45	0.18	1.08	0.06	[4, **5**, 6]	[8, 81, 11]
$(0.6, 7)$	1.09	0.32	0.45	0.28	0.66	0.04	[6, **7**, 8]	[2, 91, 7]
$(0.8, 7)$	1.00	0.25	0.43	0.21	0.79	0.04	[4, 5, 6, **7** ,8]	[1, 3, 25, 60, 11]
$(1.0, 7)$	1.03	0.27	0.44	0.21	1.05	0.05	[5, 6, **7**, 8]	[1, 33, 61, 5]

- Variational algorithm results: parameters means $(\overline{\alpha}, \overline{\sigma}, \overline{\beta})$ and standard deviations
- Numbers of clusters found given with their frequencies (most frequent number in bold characters)

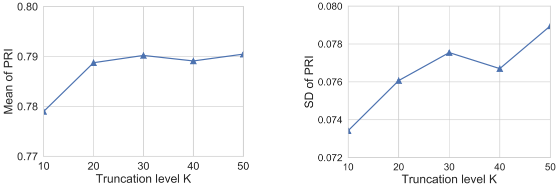
Image segmentations with the PY-Potts model

From the Berkeley segmentation data set (Arbelaez et al PAMI 2011)

Quantitative evaluation of the segmentations

Probabilistic Rand Index on 154 color (RGB) images with ground truth (several) from Berkeley dataset (1000 superpixels). But Manual ground truth segmentations are subjective !

Mean and standard deviation of the PRI as a function of the truncation level K (PY-Potts)



PRI for our PY-MRF mixture model and the approaches tested in [Chatzis 2013]

	Proposed model	Results given in [Chatzis 2013]			
PRI (%)	**PY-MRF**	DPM	iHMRF	MRF-PYP	Graph Cuts
Mean	**79.05**	74.15	75.50	76.49	76.10
Median	**80.62**	75.49	76.89	78.08	77.59
St. Dev.	**7.9**	8.4	8.2	7.9	8.3

Computation time : Berkeley 321x481 image reduced to 1000 superpixels takes **10-30 s** on a PC with CPU Intel(R) Core(TM) i7-5500U CPU 2.40GHz and 8GB RAM

Probabilistic properties of BNP-MRF priors

MRF dependencies: impact on clustering and rich-get-richer properties
eg. How does β influence the number of components?

Notation:

- K_N number of clusters in $(\theta_1, \ldots, \theta_N)$ and $(n_1, \ldots, n_{K_N})$ their size
- $\tilde{n}_\ell$ number of neighbors of θ_{N+1} which belong to cluster ℓ
- $\delta_{\mathcal{N}_{N+1}}(\ell)$ is 1 when ℓ is a label present in the neighborhood of θ_{N+1} and 0 otherwise.

Usual Gibbs-type prior predictive with $V_{N,k} = (N - \sigma k) V_{N+1,k} + V_{N+1,k+1}$ and $V_{1,1} = 1$:

$$p(\theta_{N+1} \mid \theta_1 \ldots \theta_N) = \frac{V_{N+1,K_N+1}}{V_{N,K_N}} G_0 + \frac{V_{N+1,K_N}}{V_{N,K_N}} \sum_{\ell=1}^{K_N} (n_\ell - \sigma) \delta_{\theta^\star_\ell}$$

Predictive distribution of a Gibbs-MRF prior:

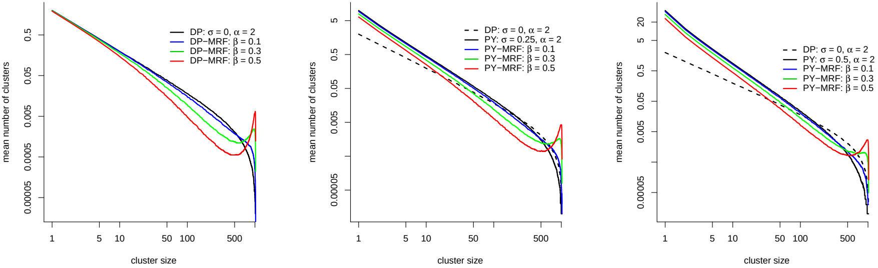
$$p(\theta_{N+1} | \theta_1 \ldots \theta_N) = \frac{V_{N+1,K_N+1}}{V_{N,K_N} + V_{N+1,K_N} \boldsymbol{\eta}_{N+1}} G_0 + \frac{V_{N+1,K_N}}{V_{N,K_N} + V_{N+1,K_N} \boldsymbol{\eta}_{N+1}} \sum_{\ell=1}^{K_N} \boldsymbol{\lambda}_{N+1,\ell} \, \delta_{\theta^\star_\ell}$$

$$\text{with } \boldsymbol{\eta}_{N+1} = \sum_{\ell \in z_{\mathcal{N}_{N+1}}} (n_\ell - \sigma)(\mathrm{e}^{\beta \tilde{n}_\ell} - 1) \text{ and } \boldsymbol{\lambda}_{N+1,\ell} = (n_\ell - \sigma) \mathrm{e}^{\beta \tilde{n}_\ell \delta_{\mathcal{N}_{N+1}}(\ell)}$$

$\Longrightarrow$ the probability of a new draw reduces as β increases.

Empirical cluster sizes

Empirical cluster sizes, over 10^5 32×32 images, 4 neighbors, Pitman-Yor with $\alpha = 2$, $\sigma \in \{0, 0.25, 0.5\}$ and Potts interaction parameter $\beta \in \{0, 0.1, 0.3, 0.5\}$.



Monte Carlo approximations of the expected number of clusters of size j with an additional smoothing over j

- BNP priors ($\beta = 0$): the probability of a large cluster decreases to 0 with its size
- BNP-MRF priors tends to favor large clusters: for larger j the number of configurations with clusters of size j decreases but their probability is much higher

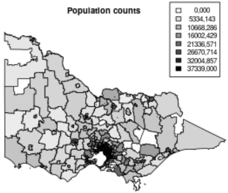
Conclusion and future work

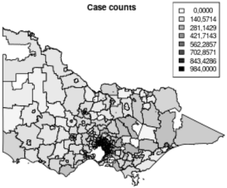
- A general scheme based on stick-breaking was proposed to build **spatial BNP priors** that can model dependencies (Markov random field).
- The **stick-breaking representation** was further exploited to derive clustering properties and to provide a variational inference algorithm based on a standard truncation.
- Illustration on a challenging unsupervised image segmentation task

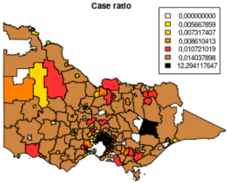
Conclusion and future work

- A general scheme based on stick-breaking was proposed to build **spatial BNP priors** that can model dependencies (Markov random field).
- The **stick-breaking representation** was further exploited to derive clustering properties and to provide a variational inference algorithm based on a standard truncation.
- Illustration on a challenging unsupervised image segmentation task

- Try other variational approximations (truncation-free), other Gibbs-type priors, stick breaking representations with dependent weights, etc.
- Other possible applications include community detection or risk mapping (extension to count data)







References

1. **This work:** H. Lü, J. Arbel, F. Forbes, Bayesian nonparametric priors for hidden Markov random fields, Statistics and Computing,30, p.1015–1035, 2020
2. **Extension to risk mapping (in prep.):** *JB. Durand, F. Forbes, H. Nguyen, CD. Phan, L. Truong, Bayesian non parametric spatial prior for car crash risk mapping*
3. M. Albughdadi, L. Chaari, J.-Y. Tourneret, F. Forbes, P. Ciuciu. A Bayesian nonparametric hidden Markov random model for hemodynamic brain parcellation. Signal Processing, 135:132-146, 2017.
4. S. P. Chatzis. A Markov random field-regulated Pitman-Yor process prior for spatially constrained data clustering. Pattern Recognition, 46(6): 1595-1603, 2013.
5. S. P. Chatzis and G. Tsechpenakis. The infinite hidden Markov random field model. IEEE Trans. Neural Networks, 21(6):1004-1014, 2010.
6. F. Forbes and N. Peyrard. Hidden Markov Random Field Selection Criteria based on Mean Field-like approximations. IEEE PAMI, 25(9):1089-1101, 2003
7. P. Orbanz and J. M. Buhmann. Nonparametric Bayesian image segmentation. International Journal of Computer Vision, 77(1-3):25-45, 2008.
8. J. Sodjo, A. Giremus, N. Dobigeon, J.F. Giovannelli, A generalized Swendsen-Wang algorithm for Bayesian nonparametric joint segmentation of multiple images, ICASSP, 2017.
9. R. Xu, F. Caron, and A. Doucet. Bayesian nonparametric image segmentation using a generalized Swendsen-Wang algorithm. ArXiv e-prints, February 2016.

Thank you for your attention!

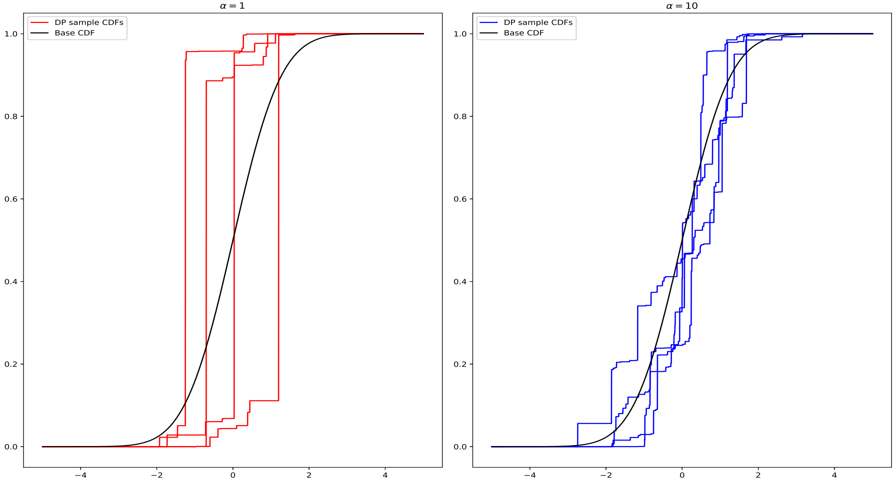
contact: florence.forbes@inria.fr

Une page de publicité

- Workshop series around ABC methods: Svalbard, Norway, 12-13 April 2021
- Mirror meetings: Brisbane, Coventry and Grenoble
- Live talks by local speakers, live interaction with Svalbard (time zone permitting)
- Mirror website: https://sites.google.com/view/abcinsvalbard-grenoble-mirror/home
- Registration free but mandatory

Annex

Stick breaking construction



DP simulations with G_0 being a standard normal distribution $\mathcal{N}(0, 1)$ and $\alpha = 1, 10$ using the Stick-Breaking representation.

Variational EM

General formulation, at iteration (r)

E-Z $q_z^{(r)}(\mathbf{z}) \propto \exp\left(E_{q_\theta^{(r-1)}}[\log p(\mathbf{y}, \mathbf{z}, \boldsymbol{\Theta}|\boldsymbol{\phi}^{(r-1)})]\right)$

E-$\boldsymbol{\Theta}$ $q_\theta^{(r)}(\boldsymbol{\Theta}) \propto \exp\left(E_{q_z^{(r)}}[\log p(\mathbf{y}, \mathbf{Z}, \boldsymbol{\Theta}|\boldsymbol{\phi}^{(r-1)})]\right)$

M-ϕ $\boldsymbol{\phi}^{(r)} = arg\max_{\boldsymbol{\phi}} E_{q_z^{(r)} q_\theta^{(r)}}[\log p(\mathbf{y}, \mathbf{Z}, \boldsymbol{\Theta}|\boldsymbol{\phi})]$

VE-Z, VE-α, VE-$\boldsymbol{\tau}$, and VE-$\boldsymbol{\theta}^*$

e.g. VE-Z step divides into N VE-Z_j steps ($q_{z_j}(z_j) = 0$ for $z_j > K$)

$$q_{z_j}(z_j) \propto \exp\left(\mathrm{E}_{q_{\theta^*_{z_j}}}\left[\log p(y_j|\theta^*_{z_j})\right] + \mathrm{E}_{q_\tau}\left[\log \pi_{z_j}(\boldsymbol{\tau})\right] + \beta \sum_{i \sim j} q_{z_i}(z_j)\right)$$

Estimation of β

M-β step: involves $p(\boldsymbol{z}|\boldsymbol{\tau},\beta) = \mathcal{K}(\beta,\boldsymbol{\tau})^{-1}\exp(V(\boldsymbol{z};\boldsymbol{\tau},\beta))$

with $V(\boldsymbol{z};\boldsymbol{\tau},\beta) = \sum_i \log \pi_{z_i}(\boldsymbol{\tau}) + \beta \sum_{i\sim j} \delta_{(z_i=z_j)}$

$$\begin{aligned}\hat{\beta} &= \arg\max_{\beta} \mathrm{E}_{q_z q_\tau}\left[\log p(\boldsymbol{z}|\boldsymbol{\tau};\beta)\right] \\ &= \arg\max_{\beta} \mathrm{E}_{q_z q_\tau}\left[V(\boldsymbol{z};\boldsymbol{\tau},\beta)\right] - \mathrm{E}_{q_\tau}\left[\log \mathcal{K}(\beta,\boldsymbol{\tau})\right]\end{aligned}$$

Two difficulties

(1) $p(\boldsymbol{z}|\boldsymbol{\tau},\beta)$ is intractable (normalizing constant $\mathcal{K}(\beta,\boldsymbol{\tau})$, typical of MRF)

(2) it depends on $\boldsymbol{\tau}$ (typical of DP)

Two approximations

(1) "standard" Mean Field like approximation[a]

(2) Replace the random $\boldsymbol{\tau}$ by a fixed $\tilde{\boldsymbol{\tau}} = E_{q_\tau}[\boldsymbol{\tau}]$

[a] Forbes & Peyrard 2003

Approximation of $p(\boldsymbol{z}|\boldsymbol{\tau};\beta)$

$$p(\boldsymbol{z}|\boldsymbol{\tau},\beta) \approx \tilde{q}_z(\boldsymbol{z}|\beta) = \prod_{j=1}^{N} \tilde{q}_{z_j}(z_j|\beta)$$

$$\tilde{q}_{z_j}(z_j = k|\beta) \quad = \quad \frac{\exp(\log \pi_k(\tilde{\boldsymbol{\tau}}) + \beta \sum_{i \in N(j)} q_{z_i}(k))}{\sum_{l=1}^{\infty} \exp(\log \pi_l(\tilde{\boldsymbol{\tau}}) + \beta \sum_{i \in N(j)} q_{z_i}(l))} \quad \text{and} \quad \tilde{\boldsymbol{\tau}} = E_{q_\tau}[\boldsymbol{\tau}]$$

β is estimated at each iteration by setting the approximate gradient to 0

$$\mathrm{E}_{q_z q_\tau}\left[\nabla_\beta V(\boldsymbol{z};\boldsymbol{\tau},\beta)\right] \quad = \quad \sum_{k=1}^{K} \sum_{i \sim j} q_{z_j}(k)\; q_{z_i}(k)$$

$$\nabla_\beta \mathrm{E}_{q_\tau}\left[\log \mathcal{K}(\beta,\boldsymbol{\tau})\right] \quad = \quad \mathrm{E}_{p(\boldsymbol{z}|\boldsymbol{\tau},\beta)q_\tau}\left[\nabla_\beta V(\boldsymbol{z};\boldsymbol{\tau},\beta)\right] \approx \sum_{k=1}^{K} \sum_{i \sim j} \tilde{q}_{z_j}(k|\beta)\; \tilde{q}_{z_i}(k|\beta)$$