[1]Bocconi University, Milano

Trees of random probability measures and Bayesian nonparametric modelling.

Ascolani, F. [1] Lijoi, A. [1] Prünster, I. [1]

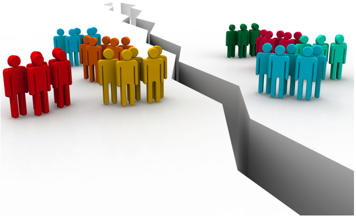
July, 2022

Outline

A crucial question in Statistics is how to combine data from different sources:

- Homogeneity **within** each group.
- Heterogeneity **across** groups.

$\Rightarrow$ **partial exchangeability**.



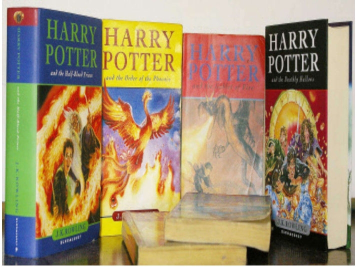
Patients coming from **different hospitals**.

Outline

A crucial question in Statistics is how to combine data from different sources:

- Homogeneity **within** each group.
- Heterogeneity **across** groups.

$\Rightarrow$ **partial exchangeability**.



Books belonging to the **same corpus**.

Outline

A crucial question in Statistics is how to combine data from different sources:

- Homogeneity **within** each group.
- Heterogeneity **across** groups.

$\Rightarrow$ **partial exchangeability**.

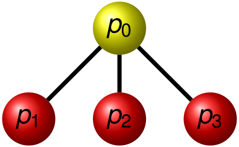
Images of **similar subjects**.

A first example

A popular model for incorporating heterogeneous information is given by **hierarchical Dirichlet processes**.

The model:

$$
\begin{aligned}
X_{i,j} \,|p_i &\overset{\text{i.i.d.}}{\sim} p_i, \\
p_i \mid p_0 &\sim \mathrm{DP}(\theta, p_0), \\
p_0 &\sim \mathrm{DP}(\theta, P_0),
\end{aligned}
$$



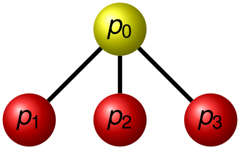
with P_0 diffuse measure.

A first example

A popular model for incorporating heterogeneous information is given by **hierarchical Dirichlet processes**.

The model:

$$X_{i,j} \mid p_i \overset{\text{i.i.d.}}{\sim} p_i,$$
$$p_i \mid p_0 \sim \mathrm{DP}(\theta, p_0),$$
$$p_0 \sim \mathrm{DP}(\theta, P_0),$$



with P_0 diffuse measure.

- Each node is a **discrete** random measure.
- Often convolved with a suitable **kernel** $k(y \mid x)$.
- In **topic modelling**, the clusters correspond to different **topics**.

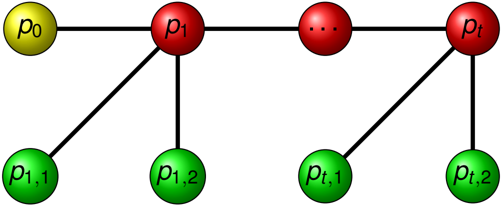
A more complex problem...

- *Data*: abstracts accepted at the NeurIPS conference from 2009 to 2019.
- *Goal*: find suitable **topics** (i.e. probability distribution on words) that describe the abstracts.
- *Problem*: incorporate the **temporal dynamics**.

$\Rightarrow$ hierarchical structure is no more appropriate!

...with a more complex structure.

- p_0 = **root** of the tree.
- p_j = node associated to year $2008 + j$.
- $p_{j,i}$ = node associated to abstract i in year $2008 + j$.



$\Rightarrow$ **tree** structure!

Another example

- A book can be seen as a **sequence of chapters**.
- Later chapters may have **similar topics** than the previous ones.
- How to introduce dependence between chapters?

Another example

- A book can be seen as a **sequence of chapters**.
- Later chapters may have **similar topics** than the previous ones.
- How to introduce dependence between chapters?

$\Rightarrow$ another special tree!

Existing literature

Many proposals for such structure:

- Extensions of Hierarchical Dirichlet processes (Teh et al., 2006, Caron et al., 2007, 2017)
- Other stick breaking priors (Qi et al, 2008)
- Other classess of priors, e.g. Pólya trees (Wang et al., 2021)

Existing literature

Many proposals for such structure:

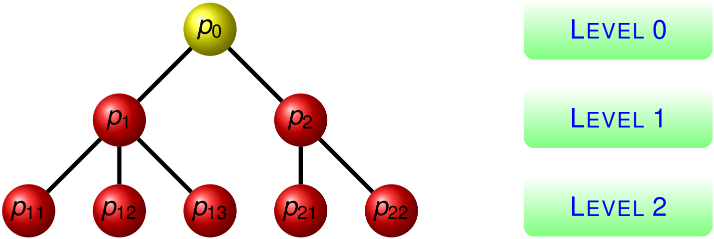
- Extensions of Hierarchical Dirichlet processes (Teh et al., 2006, Caron et al., 2007, 2017)
- Other stick breaking priors (Qi et al, 2008)
- Other classess of priors, e.g. Pólya trees (Wang et al., 2021)

What's new?

- **General framework** for constructing trees of random probability measures.
- Impact of the **shape of the tree**.
- Induced **clustering properties**.

Setting and terminology

The tree is described as follows:



- $X_{\mathbf{i},j}$ denotes the j-th observation at node $\mathbf{i}$ and $X_{\mathbf{i},j} \mid p_{\mathbf{i}} \overset{\text{i.i.d.}}{\sim} p_{\mathbf{i}}$.
- We call MRCA($\mathbf{i}, \mathbf{j}$) the Most Recent Common Ancestor of nodes $\mathbf{i}$ and $\mathbf{j}$.
- When talking about a specific level, we may omit the subscript $\mathbf{k}$.

Setting and terminology

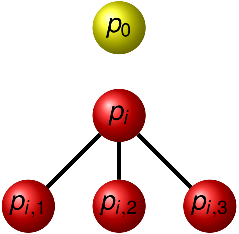
As a building block we consider Discrete RPMs, that is

$$p \overset{\text{a.s.}}{=} \sum_{k\geq 1} W_k \delta_{Z_k}, \quad \text{with} \quad \begin{cases} Z_k \overset{\text{i.i.d.}}{\sim} Q & \text{random atoms} \\ W_k & \text{random weights} \end{cases}$$

where Q is a probability distribution. We say $p \sim \mathrm{DRPM}(Q)$.

1. **Root**: $p_0 \sim \mathrm{DRPM}(P_0)$, with P_0 diffuse measure.

2. **Edges**: $p_{\mathbf{i},k} \mid p_{\mathbf{i}} \overset{\text{i.i.d.}}{\sim} \mathrm{DRPM}(p_{\mathbf{i}})$.



Completely random measures

It is a notion due to Kingman (1967).

Definition

A random variable μ is a **completely random measure** (CRM) if for any $A_1, \dots, A_n$ measurable, with $A_i \cap A_j = \emptyset$ for any $i \neq j$, the random variables $\mu(A_1), \dots, \mu(A_n)$ are mutually independent.

Completely random measures

It is a notion due to Kingman (1967).

Definition

A random variable μ is a **completely random measure** (CRM) if for any $A_1, \dots, A_n$ measurable, with $A_i \cap A_j = \emptyset$ for any $i \neq j$, the random variables $\mu(A_1), \dots, \mu(A_n)$ are mutually independent.

Key property (**Lévy–Khintchine** representation):

$$E\left[e^{-\lambda\mu(A)}\right] = e^{-\theta P_0(A)\psi(\lambda)}, \quad \psi(\lambda) = \int_{\mathbb{R}_+} (1 - e^{-\lambda s})\, \rho(\mathrm{d}s),$$

where $\theta > 0$, P_0 is a probability distribution and ρ is a measure on $\mathbb{R}_+$ such that

$$\int_{\mathbb{R}_+} \min\{1, s\}\, \rho(\mathrm{d}s) < \infty.$$

Normalized version

Under some technical conditions a CRM can be normalized (Regazzini et al. (2003)).

Definition

Let μ be a completely random measure, identified by (ρ, θ, P_0). Then $p(\cdot) = \mu(\cdot)/\mu(\mathbb{X})$ is called **normalized random measure with independent increments** (NRMI) and we say $p \sim \text{NRMI}(\rho, \theta, P_0)$.

Normalized version

Under some technical conditions a CRM can be normalized (Regazzini et al. (2003)).

Definition

Let μ be a completely random measure, identified by (ρ, θ, P_0). Then $p(\cdot) = \mu(\cdot)/\mu(\mathbb{X})$ is called **normalized random measure with independent increments** (NRMI) and we say $p \sim \text{NRMI}(\rho, \theta, P_0)$.

Notable examples:

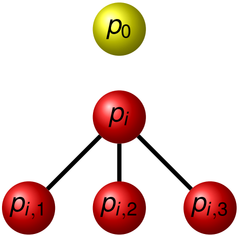
1. **Dirichlet process** (DP): $\rho(\mathrm{d}s) = s^{-1} e^{-s} \mathrm{d}s$.
2. **Normalized stable process** (NSP): $\rho(\mathrm{d}s) = \frac{\sigma}{\Gamma(1-\sigma)} s^{-1-\sigma}$, with $\sigma \in (0, 1)$.

Final definition

The model becomes:

1. **Root**: $p_0 \sim \text{NRMI}(\rho, \theta, P_0)$, with P_0 diffuse measure.

2. **Edges**: $p_{\mathbf{i},k} \mid p_{\mathbf{i}} \overset{\text{i.i.d.}}{\sim} \text{NRMI}(\rho, \theta, p_{\mathbf{i}})$.

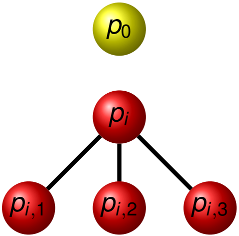


Final definition

The model becomes:

1. **Root**: $p_0 \sim \mathrm{NRMI}(\rho, \theta, P_0)$, with P_0 diffuse measure.

2. **Edges**: $p_{\mathbf{i},k} \mid p_{\mathbf{i}} \overset{\text{i.i.d.}}{\sim} \mathrm{NRMI}(\rho, \theta, p_{\mathbf{i}})$.



- Main advantage: **analytical tractability**, many prior and posterior results available.
- Probability of a tie:

$$\gamma = -\theta \int_{\mathbb{R}^+} u \left\{ \frac{\mathrm{d}^2}{\mathrm{d}u^2} \psi(u) \right\} e^{-\theta\psi(u)} \, \mathrm{d}u \in (0, 1).$$

Prior properties: correlation structure

Proposition

Let $\boldsymbol{i}$ and $\boldsymbol{j}$ be two nodes at level i and j, with MRCA at level k. Then

$$Corr\left(p_{\boldsymbol{i}}(A), p_{\boldsymbol{j}}(A)\right) = \frac{1-(1-\gamma)^{k+1}}{\sqrt{1-(1-\gamma)^{i+1}}\sqrt{1-(1-\gamma)^{j+1}}}$$

and

$$Corr\left(X_{\boldsymbol{i}}, X_{\boldsymbol{j}}\right) = 1-(1-\gamma)^{k+1}.$$

- As we move along the tree, the random measures become more and more correlated.
- As regards the DP and NSP we have:

$$\mathrm{Corr}_{DP}\left(X_{\mathbf{i}}, X_{\mathbf{j}}\right) = 1-\left(\frac{\theta}{1+\theta}\right)^{k+1}, \quad \mathrm{Corr}_{NSP}\left(X_{\mathbf{i}}, X_{\mathbf{j}}\right) = 1-\sigma^{k+1}.$$

Prior properties: asymptotic clusters

- Call K_n the number of distinct clusters in n observations. For simplicity, we focus on DP and NSP.

Proposition

Let $\boldsymbol{k}$ be a node at level k at which we collect n observations. Then

$$K_{n,DP} \approx \underbrace{\log \ldots \log}_{k+1 \text{ times}} n, \quad K_{n,NSP} \approx n^{\sigma^{k+1}},$$

as $n \to \infty$.

Prior properties: asymptotic clusters

- Call K_n the number of distinct clusters in n observations. For simplicity, we focus on DP and NSP.

Proposition

Let $\mathbf{k}$ *be a node at level* k *at which we collect* n *observations. Then*

$$K_{n,DP} \approx \underbrace{\log \ldots \log}_{k+1 \ \textit{times}} n, \quad K_{n,NSP} \approx n^{\sigma^{k+1}},$$

as $n \to \infty$*.*

Proposition

Assume to collect $m \geq 1$ *observations at each level. Then*

$$\limsup K_{n,DP} < \infty, \quad \limsup K_{n,NSP} < \infty,$$

as the number of levels diverges.

Consequences

- We can obtain the whole range of **clustering behaviours**.
- The **shape of the tree** is very relevant.

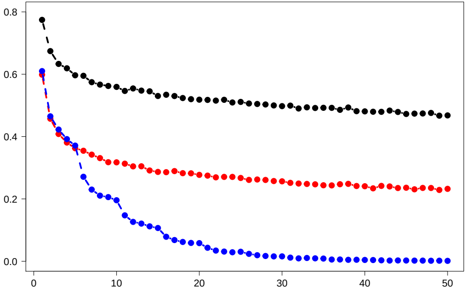


Figure: Average proportion of distinct values for groups of 20 observations.

In black: NSP with $\sigma = 0.9$.
In red: hierarchical NSP with $\sigma = 0.9$.
In blue: sequence of NSPs with $\sigma = 0.9$.

Predictive structure

How are the observations generated at a generic level k?

1. At level k, n observations are sampled from
   $p_k \mid p_{k-1} \sim NRMI(\rho, \theta, p_{k-1})$.
   Since $p_k \mid p_{k-1}$ is **discrete**, they are **grouped** in $l_k \leq n$ clusters.

Predictive structure

How are the observations generated at a generic level k?

1. At level k, n observations are sampled from
   $p_k \mid p_{k-1} \sim NRMI(\rho, \theta, p_{k-1})$.
   Since $p_k \mid p_{k-1}$ is **discrete**, they are **grouped** in $l_k \leq n$ clusters.
2. At level $k-1$, l_k observations are sampled from
   $p_{k-1} \mid p_{k-2} \sim NRMI(\rho, \theta, p_{k-2})$.
   Since $p_{k-1} \mid p_{k-2}$ is **discrete**, they are **grouped** in $l_{k-1} \leq l_k$ clusters.

Predictive structure

How are the observations generated at a generic level k?

1. At level k, n observations are sampled from $p_k \mid p_{k-1} \sim NRMI(\rho, \theta, p_{k-1})$.
   Since $p_k \mid p_{k-1}$ is **discrete**, they are **grouped** in $l_k \leq n$ clusters.
2. At level $k-1$, l_k observations are sampled from $p_{k-1} \mid p_{k-2} \sim NRMI(\rho, \theta, p_{k-2})$.
   Since $p_{k-1} \mid p_{k-2}$ is **discrete**, they are **grouped** in $l_{k-1} \leq l_k$ clusters.

$$\vdots$$

3. At level 0, l_1 observations are sampled from p_0, with $l_0 \leq l_1$ unique values from P_0.

Predictive structure

How are the observations generated at a generic level k?

1. At level k, n observations are sampled from $p_k \mid p_{k-1} \sim NRMI(\rho, \theta, p_{k-1})$.
   Since $p_k \mid p_{k-1}$ is **discrete**, they are **grouped** in $l_k \leq n$ clusters.
2. At level $k-1$, l_k observations are sampled from $p_{k-1} \mid p_{k-2} \sim NRMI(\rho, \theta, p_{k-2})$.
   Since $p_{k-1} \mid p_{k-2}$ is **discrete**, they are **grouped** in $l_{k-1} \leq l_k$ clusters.

$$\vdots$$

3. At level 0, l_1 observations are sampled from p_0, with $l_0 \leq l_1$ unique values from P_0.

- All the unique values come from the root.
- We have a **hidden clustering structure**. We only observe l_0.

Predictive structure

In terms of the **Chinese Restaurant metaphor**:

- At each level the observations are subdivided in different **tables** (i.e. clusters).
- The **dishes** (i.e. unique values) come from the previous levels.
- Different tables may **share** the same dish.

Predictive structure

In terms of the **Chinese Restaurant metaphor**:

- At each level the observations are subdivided in different **tables** (i.e. clusters).
- The **dishes** (i.e. unique values) come from the previous levels.
- Different tables may **share** the same dish.

Thus:

- $l_k =$ number of tables at level k.
- The levels share the same dishes.
- The root p_0 becomes the **common menu**.

Posterior properties: the internal nodes

We consider a sample $\mathbf{X} = \{X_{\mathbf{i},j}\}$, with $\mathbf{i}$ in the tree.

- Let $X_1^*, \dots, X_r^*$ denote the distinct observations in sample $\mathbf{X}$.
- We call $\mathbf{T}$ the **(latent) labels** of the tables.

For a fixed level k, we can then define

$l_{i,j}$ = number of tables at node i with dish j.

$q_{i,j,t}$ = number of customers at node i in table t eating dish j.

Conditional on $\mathbf{T}$, the posterior distribution becomes **accessible**!

Posterior properties: the root

Let **U** be a positive random vector with density depending on **T**.

Theorem

We have

$$\mu_0 \mid (\boldsymbol{X}, \boldsymbol{T}, U_0) \overset{d}{=} \hat{\mu}_0 + \sum_{j=1}^{r} J_{0,j} \delta_{X_j^*},$$

where

1. $\hat{\mu}_0$ *is a CRM with intensity*

$$\hat{\rho}_0(\mathrm{d}s) = e^{-U_0 s} \rho(s) \mathrm{d}s.$$

2. *The* $J_{0,j}$*'s are independent and non-negative jumps with density*

$$f_{0,j}(s \mid \boldsymbol{X}, \boldsymbol{T}) \propto s^{l_{1,j}} e^{-sU_0} \rho(s).$$

Posterior properties: the internal nodes

Theorem

At level k, with ancestor p^*, we have

$$(\mu_{\boldsymbol{k},1},\ldots,\mu_{\boldsymbol{k},d}) \mid (p^*,\boldsymbol{X},\boldsymbol{T},\boldsymbol{U}) \stackrel{d}{=} (\hat\mu_1,\ldots,\hat\mu_d) + \left(\sum_{j=1}^{r}\sum_{t=1}^{l_{1,j}} J_{1,j,t}\delta_{X_j^*},\ldots,\sum_{j=1}^{r}\sum_{t=1}^{l_{d,j}} J_{d,j,t}\delta_{X_j^*}\right),$$

where

1. $\hat\mu_i$ *is a CRM with baseline distribution p^* and intensity*

$$\hat\rho_i(\mathrm{d}s) = e^{-U_i s}\rho(s)\mathrm{d}s.$$

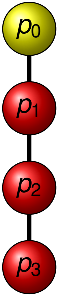
2. *The $J_{i,j,t}$'s are independent and non-negative jumps with density*

$$f_{i,j,t}(s \mid \boldsymbol{X},\boldsymbol{T}) \propto s^{q_{i,j,t}} e^{-sU_i}\rho(s).$$

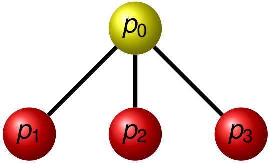
Alice in Wonderland

- We want to incorporate **chapters' specific information**.
- Two different structures.

Tree



Hierarchy



p_i is the random measure associated to chapter i.

Specifications

The **model**:

- Each p is a Dirichlet process, whose baseline distribution is given by the hierarchical structure.
- The hyperparameters at each node are endowed with vague priors.
- The root has the Dirichlet distribution as a baseline distribution, with common rate $\alpha = 50/V$.

Specifications

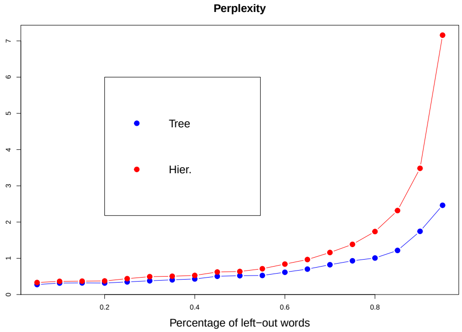
The **model**:

- Each p is a Dirichlet process, whose baseline distribution is given by the hierarchical structure.
- The hyperparameters at each node are endowed with vague priors.
- The root has the Dirichlet distribution as a baseline distribution, with common rate $\alpha = 50/V$.

The **data**:

- We consider the **initial three chapters**: standard pre-processing is applied.
- We randomly eliminate words from the second chapter and see whether the two methods are able to **recover** them.
- We measure goodness of fit in terms of the **perplexity** associated to the held-out words.

Results



- The **lower** the perplexity the **better**.
- Results are averaged over 20 runs.
- **The tree always behaves better** and has good performances even with a high proportion of missing data.

Conclusions

Summary:

- Many BNP models can be described using **trees**.
- If the nodes are given by NRMI, prior and posterior properties are available.
- We can use trees to make our learning process **explicit**.

Conclusions

Summary:

- Many BNP models can be described using **trees**.
- If the nodes are given by NRMI, prior and posterior properties are available.
- We can use trees to make our learning process **explicit**.

What's next?

- Construct a tree based on **covariates**.
- Study the asymptotic properties.
- Develop **efficient samplers** for posterior inference.

Bibliography

- *Camerlenghi, F.*, *Lijoi, A.* and *Prünster, I.* (2019). Distribution theory for hierarchical processes. *Ann. Statist.* **47**, 67–92.
- *Kingman, J. F. C.* (1967). Completely random measures. *Pacific J. Math.* **21**, 59–78.
- *Regazzini, E.*, *Lijoi, A.* and *Prünster, I.* (2003). Distributional results for means of normalized random measures with independent increments. *Ann. Statist.* **31**, 560–585.