Hierarchies of discrete random probabilities

Igor Prünster

Bocconi University

Conference on Bayesian Statistics in the Big Data Era

CIRM, Luminy
November 27, 2018

Outline

Beyond exchangeability

From de Finetti (1938):

> *But the case of exchangeability can only be considered as a limiting case: the case in which this "analogy" is, in a certain sense, absolute for all events under consideration. [..] To get from the case of exchangeability to other cases which are more general but still tractable, we must take up the case where we still encounter "analogies" among the events under consideration, but without attaining the limiting case of exchangeability.*

Beyond exchangeability

From de Finetti (1938):

> *But the case of exchangeability can only be considered as a limiting case: the case in which this "analogy" is, in a certain sense, absolute for all events under consideration. [..] To get from the case of exchangeability to other cases which are more general but still tractable, we must take up the case where we still encounter "analogies" among the events under consideration, but without attaining the limiting case of exchangeability.*

In applications dependence structures more general than exchangeability are required. We focus on data collected under different experimental conditions s.t.

- Homogeneity within each experimental condition
- Heterogeneity across different experimental conditions

Examples: Topic modeling, Meta-Analysis, two-sample problems, nonparametric regression (covariate–indexed data), time dependent data, change-point problems ...

Partial exchangeability

The array $(\boldsymbol{X}_1, \boldsymbol{X}_2) = (X_{1,i}, X_{2,j})_{i,j \geq 1}$ is *partially exchangeable* if

$$(X_{1,1}, \ldots, X_{1,n_1}, X_{2,1}, \ldots, X_{2,n_2}) \stackrel{d}{=} (X_{1,\pi(1)}, \ldots, X_{1,\pi(n_1)}, X_{2,\phi(1)}, \ldots, X_{2,\phi(n_2)})$$

for any $n_1, n_2 \geq 1$ and any permutations π and ϕ of $(1, \ldots, n_1)$ and $(1, \ldots, n_2)$.

de Finetti's representation theorem

$(\boldsymbol{X}_1, \boldsymbol{X}_2)$ is *partially exchangeable* if and only if

$$\mathbb{P}\left[X_{1,1} \in A_1, \ldots, X_{1,n_1} \in A_{n_1}, X_{2,1} \in B_1, \ldots, X_{2,n_2} \in B_{n_2}\right] = \int_{\mathscr{P}^2} \prod_{i=1}^{n_1} P_1(A_i) \prod_{j=1}^{n_2} P_2(B_j)\, Q(\mathrm{d}P_1, \mathrm{d}P_2).$$

Partial exchangeability

The array $(\boldsymbol{X}_1, \boldsymbol{X}_2) = (X_{1,i}, X_{2,j})_{i,j \geq 1}$ is *partially exchangeable* if

$$(X_{1,1}, \ldots, X_{1,n_1}, X_{2,1}, \ldots, X_{2,n_2}) \overset{d}{=} (X_{1,\pi(1)}, \ldots, X_{1,\pi(n_1)}, X_{2,\phi(1)}, \ldots, X_{2,\phi(n_2)})$$

for any $n_1, n_2 \geq 1$ and any permutations π and ϕ of $(1, \ldots, n_1)$ and $(1, \ldots, n_2)$.

de Finetti's representation theorem

$(\boldsymbol{X}_1, \boldsymbol{X}_2)$ is *partially exchangeable* if and only if

$$\mathbb{P}\left[X_{1,1} \in A_1, \ldots, X_{1,n_1} \in A_{n_1}, X_{2,1} \in B_1, \ldots, X_{2,n_2} \in B_{n_2}\right] = \int_{\mathscr{P}^2} \prod_{i=1}^{n_1} P_1(A_i) \prod_{j=1}^{n_2} P_2(B_j)\, Q(\mathrm{d}P_1, \mathrm{d}P_2).$$

$\Longrightarrow$ This is the same as saying that

$$\begin{aligned} (X_{1,i}, X_{2,j}) \mid \tilde{P}_1, \tilde{P}_2 &\overset{\text{iid}}{\sim} \tilde{P}_1 \times \tilde{P}_2 \qquad \forall i, j \geq 1 \\ (\tilde{P}_1, \tilde{P}_2) &\sim Q \end{aligned}$$

with $(\tilde{P}_1, \tilde{P}_2)$ a vector of dependent random probability measures and Q the prior.

Measuring dependence

Extreme cases of dependence induced by the prior Q

- Maximal dependence $\iff$ *full exchangeability*
  i.e. Q is degenerate on the diagonal $\{(P_1, P_2) \in \mathscr{P}^2 : P_1 = P_2\}$,
  namely $\tilde{P}_1 = \tilde{P}_2$ (a.s.)
- *Independence* $\iff$ $\tilde{P}_1$ and $\tilde{P}_2$ are (unconditionally) independent with respect to Q.
  $\implies$ corresponds to maximal heterogeneity: inference on each sample is not influenced by the observations from the other sample.

Measuring dependence

Extreme cases of dependence induced by the prior Q

- Maximal dependence $\iff$ *full exchangeability*
  i.e. Q is degenerate on the diagonal $\{(P_1, P_2) \in \mathscr{P}^2 : P_1 = P_2\}$,
  namely $\tilde{P}_1 = \tilde{P}_2$ (a.s.)
- *Independence* $\iff$ $\tilde{P}_1$ and $\tilde{P}_2$ are (unconditionally) independent with respect to Q.
  $\implies$ corresponds to maximal heterogeneity: inference on each sample is not influenced by the observations from the other sample.

Correlation as measure of dependence

The most popular measure of dependence is correlation: since

$$\text{corr}(\tilde{P}_1(A), \tilde{P}_2(A)),$$

typically does not depend on A it is taken as a measure of overall dependence.

Extreme cases correspond to:

- perfect linearity, which is implied by $\tilde{P}_1 = \tilde{P}_2$ a.s.
- uncorrelation, which is implied by $\tilde{P}_1 \perp\!\!\!\perp \tilde{P}_2$

$\implies$ good proxy for the desired measure of dependence!

Dependent priors via transformed random measures

Most popular approach to the definition of dependent nonparametric priors is via dependent stick-breaking constructions introduced by MacEachern (1999, 2000)

- Pros: "simple" implementation of "conditional" algorithms;
- Cons: it is almost impossible to derive analytic expressions for quantities of interest both marginal and conditional.

Dependent priors via transformed random measures

Most popular approach to the definition of dependent nonparametric priors is via dependent stick-breaking constructions introduced by MacEachern (1999, 2000)

- Pros: "simple" implementation of "conditional" algorithms;
- Cons: it is almost impossible to derive analytic expressions for quantities of interest both marginal and conditional.

Approach based on general random measures with 3 possible strategies (and combinations thereof) for creating dependence:

1. Hierarchical structures
   First proposed by Teh, Jordan, Beal & Blei (2006) for the Dirichlet process with stick-breaking representation: Hierarchical Dirichlet process (HDP)
   $\Longrightarrow$ here defined for general processes and studied as transformed random measures
2. Additive structures
   First proposed by Müller, Quintana & Rosner (2004) for the Dirichlet process with stick-breaking representation. For general normalized random measures theory developed in Lijoi, Nipoti and P. (2014).
3. Nested structures
   $\Longrightarrow$ Antonio's talk!

Completely random measures

Completely random measures (Kingman, 1967)

A random element $\tilde{\mu}$ taking values in the space of boundedly finite measures $\mathbb{M}$ such that, for any $d \geq 1$ and collection of pairwise disjoint sets $A_1, \ldots, A_d$, the random variables

$$\tilde{\mu}(A_1), \tilde{\mu}(A_2), \ldots \tilde{\mu}(A_d) \text{ are mutually independent}$$

is said to be a completely random measure (CRM).

Completely random measures

Completely random measures (Kingman, 1967)

A random element $\tilde{\mu}$ taking values in the space of boundedly finite measures $\mathbb{M}$ such that, for any $d \geq 1$ and collection of pairwise disjoint sets $A_1, \ldots, A_d$, the random variables

$$\tilde{\mu}(A_1), \tilde{\mu}(A_2), \ldots \tilde{\mu}(A_d) \text{ are mutually independent}$$

is said to be a completely random measure (CRM).

Key properties

Assume $\tilde{\mu}$ has no fixed points of discontinuity

- The realizations of a CRM are a.s. discrete i.e. $\tilde{\mu}(\,\cdot\,) = \sum_{i=1}^{\infty} J_i \delta_{Z_i}(\,\cdot\,)$
- A CRM $\tilde{\mu}$ is uniquely characterized by its *Laplace functional*

$$\mathbb{E}\left[\mathrm{e}^{-\int_{\mathbb{X}} g(x)\,\tilde{\mu}(\mathrm{d}x)}\right] = \mathrm{e}^{-\int_{\mathbb{R}^+\times\mathbb{X}}[1-\mathrm{e}^{-v\,g(x)}]\nu(\mathrm{d}v,\mathrm{d}x)}$$

with ν indicating the Lévy intensity, which characterizes the CRM $\tilde{\mu}$.

CRM–based nonparametric priors

Normalized completely random measures (Regazzini, Lijoi and P., 2003)

Let $\tilde{\mu}$ be a CRM on $\mathbb{X}$ such that $0 < \tilde{\mu}(\mathbb{X}) < \infty$ a.s. Then

$$\tilde{P}(\,\cdot\,) = \frac{\tilde{\mu}(\,\cdot\,)}{\tilde{\mu}(\mathbb{X})}$$

is a normalized completely random measure (NRMI=Normalized Random Measure with Independent increments).

In the following we will consider a.s. finite homogeneous CRMs i.e. $\nu(\mathrm{d}v, \mathrm{d}x) = \rho(\mathrm{d}v)cP(\mathrm{d}x)$ and write $\tilde{\mu} \sim CRM(\rho, c, P)$ and $\tilde{P} \sim NRMI(\rho, c, P)$.

CRM–based nonparametric priors

Normalized completely random measures (Regazzini, Lijoi and P., 2003)

Let $\tilde{\mu}$ be a CRM on $\mathbb{X}$ such that $0 < \tilde{\mu}(\mathbb{X}) < \infty$ a.s. Then

$$\tilde{P}(\,\cdot\,) = \frac{\tilde{\mu}(\,\cdot\,)}{\tilde{\mu}(\mathbb{X})}$$

is a normalized completely random measure (NRMI=Normalized Random Measure with Independent increments).

In the following we will consider a.s. finite homogeneous CRMs i.e. $\nu(\mathrm{d}v, \mathrm{d}x) = \rho(\mathrm{d}v)cP(\mathrm{d}x)$ and write $\tilde{\mu} \sim CRM(\rho, c, P)$ and $\tilde{P} \sim NRMI(\rho, c, P)$.

Pitman–Yor process (Pitman & Yor, 1997)

A Pitman–Yor process with parameters $\sigma \in (0, 1)$ and $\theta > 0$ can be defined via normalization as

$$\tilde{P} = \frac{\tilde{\mu}_{\sigma,\theta}}{\tilde{\mu}_{\sigma,\theta}(\mathbb{X})} \sim \mathrm{PY}(\sigma, \theta; P)$$

where $\tilde{\mu}_{\sigma,\theta}$ is a suitable transformation of a specific CRM (but not a CRM).

Hierarchical NRMI processes

Hierarchical NRMIs for partially exchangeable data

$$(X_{1,i}, X_{2,j}) \,|\, (\tilde{P}_1, \tilde{P}_2) \overset{\text{iid}}{\sim} \tilde{P}_1 \times \tilde{P}_2 \qquad \forall i, j \geq 1$$

$$(\tilde{P}_1, \tilde{P}_2) \,|\, \tilde{P}_0 \overset{\text{iid}}{\sim} \text{NRMI}(\rho, c, \tilde{P}_0)$$

$$\tilde{P}_0 \sim \text{NRMI}(\rho_0, c_0, P_0)$$

with P_0 a non–atomic measure on $\mathbb{X}$.

Hierarchical NRMI processes

Hierarchical NRMIs for partially exchangeable data

$$(X_{1,i}, X_{2,j}) \,|\, (\tilde{P}_1, \tilde{P}_2) \overset{\text{iid}}{\sim} \tilde{P}_1 \times \tilde{P}_2 \qquad \forall i, j \geq 1$$

$$(\tilde{P}_1, \tilde{P}_2) \,|\, \tilde{P}_0 \overset{\text{iid}}{\sim} \text{NRMI}(\rho, c, \tilde{P}_0)$$

$$\tilde{P}_0 \sim \text{NRMI}(\rho_0, c_0, P_0)$$

with P_0 a non–atomic measure on $\mathbb{X}$.

Special cases

- Hierarchical Dirichlet process (HDP)
  $\Longrightarrow$ NRMIs coincide with the Dirichlet process, i.e.

$$\rho(\mathrm{d}v) = \rho_0(\mathrm{d}v) = \frac{\mathrm{e}^{-v}}{v}\mathrm{d}v$$

- Hierarchical normalized stable process (HnstP)
  $\Longrightarrow$ NRMIs coincide with the normalized stable process, i.e.

$$\rho(\mathrm{d}v) = \frac{\sigma}{\Gamma(1-\sigma)v^{1+\sigma}}\mathrm{d}v \qquad \rho_0(\mathrm{d}v) = \frac{\sigma_0}{\Gamma(1-\sigma_0)v^{1+\sigma_0}}\mathrm{d}v$$

Correlation structure

Correlation structure for hierarchical NRMIs

$$\mathrm{corr}(\tilde{P}_1(A), \tilde{P}_2(A))$$

$$= \left\{ 1 + c_0\, c\, \frac{\int_0^\infty u\, \mathrm{e}^{-c\psi(u)}\, \tau_2(u)\, \mathrm{d}u \int_0^\infty u\, \mathrm{e}^{-c_0\psi_0(u)}\, \tau_{1,0}^2(u)\, \mathrm{d}u}{\int_0^\infty u\, \mathrm{e}^{-c_0\psi_0(u)}\, \tau_{2,0}(u)\, \mathrm{d}u} \right\}^{-1} > 0$$

with $\psi(u) = \int_0^\infty [1 - \mathrm{e}^{-us}]\rho(s)\mathrm{d}s$ and $\tau_q(u) = \int_0^\infty s^q\, \mathrm{e}^{-us}\, \rho(s)\, \mathrm{d}s$.

Correlation structure

Correlation structure for hierarchical NRMIs

$$\mathrm{corr}(\tilde{P}_1(A), \tilde{P}_2(A))$$

$$= \left\{ 1 + c_0\, c\, \frac{\int_0^\infty u\, \mathrm{e}^{-c\psi(u)}\, \tau_2(u)\, \mathrm{d}u \int_0^\infty u\, \mathrm{e}^{-c_0\psi_0(u)}\, \tau_{1,0}^2(u)\, \mathrm{d}u}{\int_0^\infty u\, \mathrm{e}^{-c_0\psi_0(u)}\, \tau_{2,0}(u)\, \mathrm{d}u} \right\}^{-1} > 0$$

with $\psi(u) = \int_0^\infty [1 - \mathrm{e}^{-us}]\rho(s)\mathrm{d}s$ and $\tau_q(u) = \int_0^\infty s^q\, \mathrm{e}^{-us}\, \rho(s)\, \mathrm{d}s$.

Special cases

- Hierarchical Dirichlet process (HDP)

$$\mathrm{corr}(\tilde{P}_1(A), \tilde{P}_2(A)) = \frac{c+1}{c+1+c_0}$$

$\Longrightarrow$ corr increasing in c and decreasing in c_0: if $c \uparrow \infty$ ($c_0 \uparrow \infty$), then corr $\uparrow 1$ (corr $\downarrow 0$).

- Hierarchical normalized stable process (HnstP)

$$\mathrm{corr}(\tilde{P}_1(A), \tilde{P}_2(A)) = \frac{1-\sigma_0}{1-\sigma\sigma_0}$$

$\Longrightarrow$ corr increasing in σ and decreasing in σ_0: if $\sigma \uparrow 1$ ($\sigma_0 \uparrow 1$), then corr $\uparrow 1$ (corr $\downarrow 0$).

Induced partition structure

- Two samples $\boldsymbol{X}_1 = \{X_{1,1}, \ldots, X_{1,n_1}\}$ and $\boldsymbol{X}_2 = \{X_{2,1}, \ldots, X_{2,n_2}\}$ from partially exchangeable sequences $(X_{1,j})_{j \geq 1}$ and $(X_{2,i})_{i \geq 1}$
- Hierarchical NRMI prior selects a.s. discrete random probabilities $(\tilde{P}_1, \tilde{P}_2)$
  $\Longrightarrow$ ties within each sample and possibly also across different samples.

Induced partition structure

- Two samples $\boldsymbol{X}_1 = \{X_{1,1}, \ldots, X_{1,n_1}\}$ and $\boldsymbol{X}_2 = \{X_{2,1}, \ldots, X_{2,n_2}\}$ from partially exchangeable sequences $(X_{1,j})_{j\geq 1}$ and $(X_{2,i})_{i\geq 1}$
- Hierarchical NRMI prior selects a.s. discrete random probabilities $(\tilde{P}_1, \tilde{P}_2)$ $\Longrightarrow$ ties within each sample and possibly also across different samples.
- Partition of $[N] = \{1, \ldots, n_1 + n_2\}$ induced by $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ into
  - k_1 distinct values specific to $\boldsymbol{X}_1$
  - k_2 distinct values specific to $\boldsymbol{X}_2$
  - k_0 distinct values shared by the two samples
  - the corresponding frequencies are best recorded as $\boldsymbol{n}_i = (n_{1,i}, n_{2,i})$ for $i = 1, \ldots, k$, with k_2 $n_{1,j}$'s and k_1 $n_{2,j}$'s being 0.

partially Exchangeable Partition Probability Function (pEPPF)

$$\Pi_k^{(N)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_k) = \mathbb{E}\int_{\mathbb{X}^k} \prod_{j=1}^{k} \tilde{P}_1^{n_{1,j}}(\mathrm{d}x_j)\, \tilde{P}_2^{n_{2,j}}(\mathrm{d}x_j)$$

where $N = n_1 + n_2$ and $k = k_1 + k_2 + k_0$.

Chinese restaurant franchise metaphor

Observable level: customers and dishes

- There are $d = 2$ restaurants sharing the same menu.
- $X_{i,j}$: label of the dish served at restaurant i to customer j
- Sample information: $N = n_1 + n_2$ customers eat $k = k_1 + k_2 + k_0$ distinct dishes with frequencies $\boldsymbol{n}_1, \ldots, \boldsymbol{n}_k$.

Chinese restaurant franchise metaphor

Observable level: customers and dishes

- There are $d = 2$ restaurants sharing the same menu.
- $X_{i,j}$: label of the dish served at restaurant i to customer j
- Sample information: $N = n_1 + n_2$ customers eat $k = k_1 + k_2 + k_0$ distinct dishes with frequencies $\boldsymbol{n}_1, \ldots, \boldsymbol{n}_k$.

Latent level: tables (governed by $\tilde{P}_0$)

- Customers eating dish j in restaurant i are further partitioned into tables.
- $\ell_{i,j}$ is the number of tables in restaurant i serving dish j whose range is $\{1, \ldots, n_{i,j}\}$ if dish j is served at restaurant i and 0 otherwise.
- $\bar{\ell}_{\bullet j} = \sum_{i=1}^{2} \ell_{i,j}$ is then the total number of tables serving dish j for $j = 1, \ldots, k$
- $|\boldsymbol{\ell}|$: total number of tables in the two restaurants

Chinese restaurant franchise metaphor

Observable level: customers and dishes

- There are $d = 2$ restaurants sharing the same menu.
- $X_{i,j}$: label of the dish served at restaurant i to customer j
- Sample information: $N = n_1 + n_2$ customers eat $k = k_1 + k_2 + k_0$ distinct dishes with frequencies $\boldsymbol{n}_1, \ldots, \boldsymbol{n}_k$.

Latent level: tables (governed by $\tilde{P}_0$)

- Customers eating dish j in restaurant i are further partitioned into tables.
- $\ell_{i,j}$ is the number of tables in restaurant i serving dish j whose range is $\{1, \ldots, n_{i,j}\}$ if dish j is served at restaurant i and 0 otherwise.
- $\bar{\ell}_{\bullet j} = \sum_{i=1}^{2} \ell_{i,j}$ is then the total number of tables serving dish j for $j = 1, \ldots, k$
- $|\boldsymbol{\ell}|$: total number of tables in the two restaurants

Augmented partition structure

- $q_{i,j,t}$: frequency of customers at restaurant i eating dish j and sitting at table t
- $\boldsymbol{q}_{i,j} = (q_{i,j,1}, \ldots, q_{i,j,\ell_{i,j}})$: frequency vector of customers in restaurant i eating dish j at each of the $\ell_{i,j}$ tables.
- By marginalizing over the tables one re-obtains the observed frequencies $n_{i,j} = |\boldsymbol{q}_{i,j}| = \sum_{t=1}^{\ell_{i,j}} q_{i,j,t}$.

Partially exchangeable partition probability function

pEPPF of a hierarchical NRMI

$$\Pi_k^{(N)}(\boldsymbol{n}_1, \cdots, \boldsymbol{n}_k) = \sum_{\boldsymbol{\ell}} \sum_{\boldsymbol{q}} \Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\bar{\ell}_{\bullet 1}, \cdots, \bar{\ell}_{\bullet k}) \prod_{i=1}^{2} \Phi_{\bar{\ell}_{i\bullet},i}^{(n_i)}(\boldsymbol{q}_{i,1}, \ldots, \boldsymbol{q}_{i,k})$$

- $\sum_{\boldsymbol{q}}$ is a sum over all partitions
- $\sum_{\boldsymbol{\ell}}$ is a sum over all compatible table configurations, i.e. over $\ell_{i,j} \in \{1, \ldots, n_{i,j}\}$ with $\ell_{i,j} = 0$ if $n_{i,j} = 0$.
- Partition probability function with the constraint $|\boldsymbol{q}_{i,j}| = n_{i,j}$

$$\Phi_{\bar{\ell}_{i\bullet},i}^{(n_i)}(\boldsymbol{q}_{i,1}, \ldots, \boldsymbol{q}_{i,k}) = \frac{\theta^{\bar{\ell}_{i\bullet}}}{\Gamma(n_i)} \int_0^\infty u^{n_i - 1} e^{-\theta\psi(u)} \prod_{j=1}^{k} \prod_{t=1}^{\ell_{i,j}} \tau_{q_{i,j,t}}(u) \mathrm{d}u.$$

- $\Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\bar{\ell}_{\bullet 1}, \cdots, \bar{\ell}_{\bullet k})$ is the EPPF associated to $\tilde{P}_0 \sim NRMI(\rho, c_0, P_0)$

$$\Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\bar{\ell}_{\bullet 1}, \cdots, \bar{\ell}_{\bullet k}) = \frac{c_0^k}{\Gamma(|\boldsymbol{\ell}|)} \int_0^\infty u^{|\boldsymbol{\ell}|-1} \, \mathrm{e}^{-c_0\psi_0(u)} \prod_{j=1}^{k} \tau_{\bar{\ell}_{\bullet j},0}(u) \, \mathrm{d}u,$$

with $\psi_0(u) = \int_0^\infty [1 - \mathrm{e}^{-uv}] \, \rho_0(\mathrm{d}v)$ and $\tau_{q,0}(u) = \int_0^\infty v^q \, \mathrm{e}^{-uv} \, \rho_0(\mathrm{d}v)$.

Special cases

pEEPF HDP

$$\Pi_k^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_k) = \frac{\theta_0^k}{\prod_{i=1}^2 (\theta)_{n_i}} \sum_{\boldsymbol{\ell}} \frac{\theta^{|\boldsymbol{\ell}|}}{(\theta_0)_{|\boldsymbol{\ell}|}} \prod_{j=1}^k (\bar{\ell}_{\bullet j} - 1)! \prod_{i=1}^2 |\mathfrak{s}(n_{i,j}, \ell_{i,j})|$$

where $|\mathfrak{s}(r, s)|$ is the absolute value of the Stirling number of the first kind.

pEEPF hierarchical normalized stable process (HnstP)

$$\frac{\sigma_0^{k-1}\,\Gamma(k)}{\prod_{i=1}^2 \Gamma(n_i)} \sum_{\boldsymbol{\ell}} \frac{\sigma^{|\boldsymbol{\ell}|-2} \prod_{i=1}^2 \Gamma(\bar{\ell}_{i\bullet})}{\Gamma(|\boldsymbol{\ell}|)} \prod_{j=1}^k (1-\sigma_0)_{\bar{\ell}_{\bullet j}-1} \prod_{i=1}^2 \prod_{j=1}^k \frac{\mathscr{C}(n_{i,j}, \ell_{i,j}; \sigma)}{\sigma^{\ell_{i,j}}}$$

where $\mathscr{C}(n, \ell; \sigma)$ is the *generalized factorial coefficient*.

$\Longrightarrow$ From the pEPPF a generalized Pólya urn scheme is obtained that can be used within MCMC samplers for density estimation, prediction problems etc.

Distribution of the number of clusters K_N

In order to derive the distribution of the distinct dishes K_N eaten by $N_1 + N_2$ customers define:

- K'_{1,N_2} and K'_{2,N_2} are the number of tables the customers seated in the two restaurants;
- $K_{0,t}$ is the number of distinct dishes (generated by $\tilde{P}_0$) served at the t tables.

Distribution of K_N for a hierarchical NRMI

$$\mathbb{P}[K_N = k] = \sum_{t=k}^{N} \mathbb{P}[K_{0,t} = k]\, \mathbb{P}\left[K'_{1,N_1} + K'_{2,N_2} = t\right]$$

Distribution of the number of clusters K_N

In order to derive the distribution of the distinct dishes K_N eaten by $N_1 + N_2$ customers define:

- K'_{1,N_2} and K'_{2,N_2} are the number of tables the customers seated in the two restaurants;
- $K_{0,t}$ is the number of distinct dishes (generated by $\tilde{P}_0$) served at the t tables.

Distribution of K_N for a hierarchical NRMI

$$\mathbb{P}[K_N = k] = \sum_{t=k}^{N} \mathbb{P}[K_{0,t} = k]\, \mathbb{P}\left[K'_{1,N_1} + K'_{2,N_2} = t\right]$$

- The distributions of $K_{0,t}$ and of K'_{1,N_1} and K'_{2,N_2} are available once the corresponding EPPFs are known.
- The law of K_N coincides with $K_{0,(K'_{1,N_1}+K'_{2,N_2})}$ i.e. the random number of dishes served at the random number of tables $K'_{1,N_1} + K'_{2,N_2}$.

Asymptotics for the number of clusters K_N

What is the growth rate of K_N as N_1 and N_2 diverge?

The notation $Y_n \simeq \lambda(n)$, for $n \to \infty$, means that $\lim_n Y_n/\lambda(n)$ almost surely exists and equals a finite random variable.

Asymptotics of K_N for a hierarchical NRMI

Suppose $K_{0,N} \simeq \lambda_0(N)$ and $K'_{i,N} \simeq \lambda(N)$ as $N \to \infty$. Then letting $N_1 = N_2 = N/2$.

$$K_N \simeq \lambda_0(\eta\, \lambda(N/2)) \qquad \text{as } N \to \infty,$$

for some positive and finite positive random variable η.

Special cases

- Hierarchical Dirichlet process (HDP):

$$K_N \simeq \log\log N$$

- Hierarchical normalized stable process (HnstP):

$$K_N \simeq N^{\sigma\sigma_0}$$

Hierarchical Pitman–Yor process

All previous results carry over to the case of hierarchical Pitman–Yor processes with minor modifications. For instance:

Correlation structure

$$\text{corr}(\tilde{P}_1(A), \tilde{P}_2(A)) = \left\{ 1 + \frac{1-\sigma}{1-\sigma_0} \frac{\theta_0 + \sigma_0}{\theta + 1} \right\}^{-1}$$

Hierarchical Pitman–Yor process

All previous results carry over to the case of hierarchical Pitman–Yor processes with minor modifications. For instance:

Correlation structure

$$\mathrm{corr}(\tilde{P}_1(A), \tilde{P}_2(A)) = \left\{ 1 + \frac{1-\sigma}{1-\sigma_0} \frac{\theta_0 + \sigma_0}{\theta + 1} \right\}^{-1}$$

Distribution of K_N

- exact distribution

$$\mathbb{P}[K_N = k] = \sum_{t=k}^{N} \mathbb{P}[K_{0,t} = k] \, \mathbb{P}\left[K'_{1,N_1} + K'_{2,N_2} = t \right]$$

$$= \sum_{t=k}^{N} \frac{\prod_{r=1}^{k-1} (\theta_0 + r\sigma_0)}{(\theta_0 + 1)_{t-1}} \frac{\mathscr{C}(t, k; \sigma_0)}{\sigma_0^k} \sum_{(\zeta_1, \zeta_2) \in \Delta_t} \prod_{i=1}^{2} \frac{\prod_{r=1}^{\zeta_i - 1} (\theta + r\sigma)}{(\theta + 1)_{N_i - 1}} \frac{\mathscr{C}(N_i, \zeta_i; \sigma)}{\sigma^{\zeta_i}}$$

- asymptotic growth $K_N \simeq N^{\sigma\sigma_0}$

Posterior characterization for hierarchical NRMIs

- $X_1^*, \ldots, X_k^*$ are the distinct dishes served and $\boldsymbol{T}$ the complete table configuration in the $d = 2$ restaurants.
- Let U_0 be a positive r.v. with density

$$f_0(u|\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}) \propto u^{|\boldsymbol{\ell}|-1} e^{-c_0 \psi_0(u)} \prod_{j=1}^{k} \tau_{\bar{\ell}_{\bullet j},0}(u).$$

Posterior characterization for hierarchical NRMIs

- $X_1^*, \ldots, X_k^*$ are the distinct dishes served and $\boldsymbol{T}$ the complete table configuration in the $d = 2$ restaurants.
- Let U_0 be a positive r.v. with density

$$f_0(u|\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}) \propto u^{|\boldsymbol{\ell}|-1} e^{-c_0\psi_0(u)} \prod_{j=1}^{k} \tau_{\bar{\ell}_{\bullet j},0}(u).$$

Posterior distribution of $\tilde{P}_0$

$$\tilde{\mu}_0 \,|\, (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}, U_0) \stackrel{\mathrm{d}}{=} \eta_0^* + \sum_{j=1}^{k} I_j \delta_{X_j^*}$$

(i) η_0^* is a CRM with intensity

$$\nu_0(\mathrm{d}s, \mathrm{d}x) = \mathrm{e}^{-U_0 s} \rho_0(s)\mathrm{d}s\, c_0\, P_0(\mathrm{d}x).$$

(ii) the I_j's are non-negative independent jumps (also independent of η_0^*) with density

$$f_j(s|\boldsymbol{X}, \boldsymbol{T}) \propto s^{\bar{\ell}_{\bullet j}} e^{-sU_0} \rho_0(s)$$

- Introduce the restaurant specific latent r.v. U_i, for $i = 1, 2$

$$f_i(u|\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}) \propto u^{n_i - 1} e^{-c\psi(u)} \prod_{j=1}^{k} \prod_{t=1}^{\ell_{i,j}} \tau_{q_{i,j,t}}(u).$$

- Introduce the restaurant specific latent r.v. U_i, for $i = 1, 2$

$$f_i(u|\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}) \propto u^{n_i - 1} e^{-c\psi(u)} \prod_{j=1}^{k} \prod_{t=1}^{\ell_{i,j}} \tau_{q_{i,j,t}}(u).$$

Posterior distribution of $(\tilde{P}_1, \tilde{P}_2)$

$$(\tilde{\mu}_1, \tilde{\mu}_2)|(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}, \boldsymbol{U}, \tilde{\mu}_0) \stackrel{d}{=} \left(\tilde{\mu}_1^* + \sum_{j=1}^{k} \sum_{t=1}^{\ell_{1,j}} J_{1,j,t} \delta_{X_j^*},\ \tilde{\mu}_2^* + \sum_{j=1}^{k} \sum_{t=1}^{\ell_{2,j}} J_{2,j,t} \delta_{X_j^*}\right)$$

(i) $(\tilde{\mu}_1^*, \tilde{\mu}_2^*)$ is a vector of hierarchical CRMs with intensity

$$\nu_i(\mathrm{d}s, \mathrm{d}x) = e^{-U_i s} \rho(s) \mathrm{d}s\, c\, \tilde{P}_0^*(\mathrm{d}x),$$

with $\tilde{P}_0^* = \tilde{\mu}_0^* / \tilde{\mu}_0^*(\mathbb{X})$;

(ii) the $J_{i,j,t}$'s are non–negative independent jumps [also independent of $(\tilde{\mu}_1, \tilde{\mu}_2)$] with density

$$f_{i,j,t}(s) \propto e^{-U_i s} s^{q_{i,j,t}}\, \rho(s).$$

$\Longrightarrow$ Based on the posterior characterization it is straightforward to set up a conditional sampling scheme (e.g. a Ferguson & Klass) for density estimation, prediction problems etc.

HDP case

By exploiting some of the special properties of the Dirichlet process one obtains a simple posterior characterization for the HDP.

Posterior distribution of the HDP

$$\tilde{P}_0|(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}) \stackrel{d}{=} \tilde{P}_0^* \sim \mathscr{D}(c_0 P_0 + \sum_{j=1}^{k} \bar{\ell}_{\bullet j} \delta_{X_j^*})$$

and for $i = 1, 2$

$$\tilde{P}_i|(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}, \tilde{P}_0^*) \stackrel{d}{=} \tilde{P}_i^* \sim \mathscr{D}(c\tilde{P}_0^* + \sum_{j=1}^{k} n_{i,j} \delta_{X_j^*})$$

Hierarchical Pitman-Yor process

- Define $V_0^{\sigma_0} \sim \mathrm{Ga}(k + \theta_0/\sigma_0, 1)$ and $V_i^{\sigma} \overset{\text{ind}}{\sim} \mathrm{Ga}(\bar{\ell}_{i\bullet} + \theta/\sigma,\ 1)$

Posterior distribution

$$\tilde{\mu}_0 \,|\, (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}, V_0) \overset{\mathrm{d}}{=} \eta_0^* + \sum_{j=1}^{k} I_j \delta_{X_j^*}$$

$$(\tilde{\mu}_1, \tilde{\mu}_2) | (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}, \boldsymbol{V}, \tilde{\mu}_0) \overset{d}{=} \left(\tilde{\mu}_1^* + \sum_{j=1}^{k} H_{1,j} \delta_{X_j^*},\ \tilde{\mu}_2^* + \sum_{j=1}^{k} H_{2,j} \delta_{X_j^*} \right)$$

(i) $\eta_0^*, \tilde{\mu}_1, \tilde{\mu}_2$ are generalized gamma CRMs (independent of the jumps) with intensities

$$\frac{\sigma_0}{\Gamma(1-\sigma_0)} \frac{\mathrm{e}^{-V_0 s}}{s^{1+\sigma_0}} \mathrm{d}s P_0(\mathrm{d}x) \qquad \frac{\sigma}{\Gamma(1-\sigma)} \frac{\mathrm{e}^{-V_i s}}{s^{1+\sigma}} \mathrm{d}s\, \tilde{P}_0^*(\mathrm{d}x) \quad i = 1, 2$$

where $\tilde{P}_0^* = \tilde{\mu}_0^* / \tilde{\mu}_0^*(\mathbb{X})$;

(ii) $I_j \overset{\text{ind}}{\sim} \mathrm{Ga}(\bar{\ell}_{\bullet j} - \sigma_0, V_0)$ and $H_{i,j} \overset{\text{ind}}{\sim} \mathrm{Ga}(n_{i,j} - \ell_{i,j}\sigma, V_i)$

Hierarchical Pitman-Yor process II

In the PY case it is possible to marginalize out the latent variables V_0, V_1, V_2 to obtain a quasi–conjugate posterior characterization

Posterior distribution II

$$\tilde{P}_0|(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}) \stackrel{d}{=} \sum_{j=1}^{k} W_j \delta_{X_j^*} + W_{k+1}\, \tilde{P}_{0,k}$$

$$\tilde{P}_i|(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}, \tilde{P}_0^*) \stackrel{d}{=} \sum_{j=1}^{k} W_{i,j}\, \delta_{X_j^*} + W_{i,k_i+1}\, \tilde{P}_{i,k}$$

with Dirichlet distributed weights

$$(W_1, \ldots, W_{k+1}) \sim \mathrm{Dir}(\bar{\ell}_{\bullet 1} - \sigma_0, \ldots, \bar{\ell}_{\bullet k} - \sigma_0, \theta_0 + k\sigma_0)$$

$$(W_{i,1}, \ldots, W_{i,k_i+1}) \sim \mathrm{Dir}(n_{i,1} - \ell_{i,1}\sigma, \ldots, n_{i,k_i} - \ell_{i,k_i}\sigma, \theta + \bar{\ell}_{i\bullet}\sigma) \qquad i = 1, 2$$

and updated PY processes

$$\tilde{P}_{0,k} \sim \mathrm{PY}(\sigma_0, \theta_0 + k\sigma_0; P_0)$$

$$\tilde{P}_{i,k} \mid \tilde{P}_0 \stackrel{\text{ind}}{\sim} \mathrm{PY}(\sigma, \theta + \bar{\ell}_{i\bullet}\sigma; \tilde{P}_0) \qquad i = 1, 2$$

Prediction with hierarchical processes

- Conditional on observed samples $\boldsymbol{X}^{(n_1)}$ and $\boldsymbol{X}^{(n_2)}$, interest in prediction of features related to additional future samples

$$X_{1,n_1+1}, \ldots, X_{1,n_1+m_1} \qquad X_{2,n_2+1}, \ldots, X_{2,n_2+m_2}$$

- Species sampling: $X_{1,j}$'s and $X_{2,j}$'s are species labels with species shared within and between samples and the goal is to estimate e.g.:
  - the number of new distinct species that will be observed
  - the probability that $(n_i + m_i + 1)$-th observation will be a new species

Prediction with hierarchical processes

- Conditional on observed samples $\boldsymbol{X}^{(n_1)}$ and $\boldsymbol{X}^{(n_2)}$, interest in prediction of features related to additional future samples

$$X_{1,n_1+1}, \ldots, X_{1,n_1+m_1} \qquad X_{2,n_2+1}, \ldots, X_{2,n_2+m_2}$$

- Species sampling: $X_{1,j}$'s and $X_{2,j}$'s are species labels with species shared within and between samples and the goal is to estimate e.g.:
  - the number of new distinct species that will be observed
  - the probability that $(n_i + m_i + 1)$-th observation will be a new species
- Exchangeable case: closed form estimators for Gibbs–type priors (reviewed in De Blasi *et al.*, 2015)
- Partially exchangeable case: it is not possible to evaluate exactly inferences
  $\Longrightarrow$ simulation algorithm based on the pEPPF

Prediction with hierarchical processes

- Conditional on observed samples $\boldsymbol{X}^{(n_1)}$ and $\boldsymbol{X}^{(n_2)}$, interest in prediction of features related to additional future samples

$$X_{1,n_1+1}, \ldots, X_{1,n_1+m_1} \qquad X_{2,n_2+1}, \ldots, X_{2,n_2+m_2}$$

- Species sampling: $X_{1,j}$'s and $X_{2,j}$'s are species labels with species shared within and between samples and the goal is to estimate e.g.:
  - the number of new distinct species that will be observed
  - the probability that $(n_i + m_i + 1)$-th observation will be a new species
- Exchangeable case: closed form estimators for Gibbs–type priors (reviewed in De Blasi *et al.*, 2015)
- Partially exchangeable case: it is not possible to evaluate exactly inferences $\Longrightarrow$ simulation algorithm based on the pEPPF
- Illustration: ESTs analyses
  - Useful tool for gene identification in organisms
  - ESTs are generated by randomly sequencing genes from a cDNA library, which consist of a large and unknown number of differentially expressed genes (typically millions) $\Longrightarrow$ potentially infinite.
  - Only a sample corresponding to a small portion of the library is typically available

Citrus clementina libraries

- Samples from two libraries of *citrus clementina* fruits:
  - FRUIT1 ('FlavFr1'):
    $n_1 = 1593$ ESTs with $k_1 = 806$ distinct genes ($k_1/n_1 \approx .51$)
  - FRUIT2 ('RindPdig24'):
    $n_2 = 900$ ESTs with $k_2 = 687$ distinct genes ($k_2/n_2 \approx .76$)

Citrus clementina libraries

- Samples from two libraries of *citrus clementina* fruits:
  - FRUIT1 ('FlavFr1'):
    $n_1 = 1593$ ESTs with $k_1 = 806$ distinct genes ($k_1/n_1 \approx .51$)
  - FRUIT2 ('RindPdig24'):
    $n_2 = 900$ ESTs with $k_2 = 687$ distinct genes ($k_2/n_2 \approx .76$)
  - The two samples share 183 genes and their frequency is 520 in the FRUIT 1 and 317 in the FRUIT 2 samples (about 1/3)

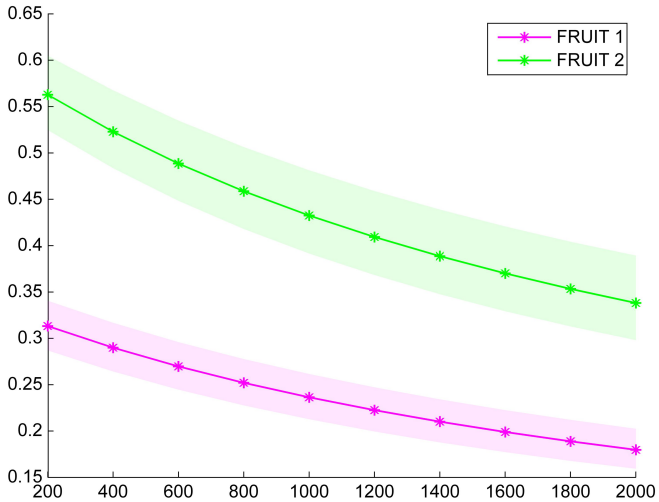
Expression level	FRUIT 1	FRUIT 2	FRUITS
1	561	549	905
2	148	99	231
3	37	20	79
4	18	12	32
5	6	4	11
6	5		9
⋮	⋮	⋮	⋮
117	1		1
n	1593	900	2493
K_n	806	687	1310

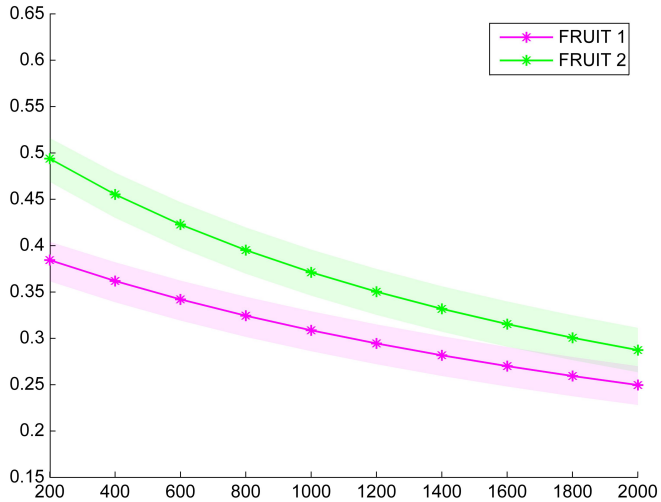
- Models: $\tilde{P}_1 \perp\!\!\!\perp \tilde{P}_2$ independent PY processes
  vs
  $(\tilde{P}_1, \tilde{P}_2)$ hierarchical PY process

- Models: $\tilde{P}_1 \perp\!\!\!\perp \tilde{P}_2$ independent PY processes
  vs
  $(\tilde{P}_1, \tilde{P}_2)$ hierarchical PY process

- Discovery probability decay: Probability that the $(n_i + m_i + 1)$–th observation is "new" as the size m_i of additional sample varies

(a) (separately) exchangeable (b) partially exchangeable





$\Rightarrow$ Borrowing of information and narrower HPD bands

Other quantities of interest:

- # "new" distinct genes identified by additional sequencing: $K^X_{m|n_1}$ & $K^Y_{m|n_2}$
- # additionally sequenced genes coinciding with "new" ones: $L^X_{m|n_1}$ & $L^Y_{m|n_2}$

	$\tilde{P}_1 \perp\!\!\!\perp \tilde{P}_2$ PY processes				$(\tilde{P}_1, \tilde{P}_2)$ HPY process			
	FRUIT 1		FRUIT 2		FRUIT 1		FRUIT 2	
m	$\hat{K}^X_{m\|n_1}$	$\hat{L}^Y_{m\|n_2}$	$\hat{K}^X_{m\|n_1}$	$\hat{L}^Y_{m\|n_2}$	$\hat{K}^X_{m\|n_1}$	$\hat{L}^Y_{m\|n_2}$	$\hat{K}^X_{m\|n_1}$	$\hat{L}^Y_{m\|n_2}$
200	65.4	68.2	117.0	122.0	79.5	82.1	103.3	108.2
400	125.6	136.2	225.5	244.0	154.1	164.2	198.3	216.6
600	181.5	204.3	326.4	366.0	224.5	246.2	286.0	324.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2000	488.1	680.8	891.3	1219.0	631.4	820.9	770.2	1083.1

Other quantities of interest:

- # "new" distinct genes identified by additional sequencing: $K^X_{m|n_1}$ & $K^Y_{m|n_2}$
- # additionally sequenced genes coinciding with "new" ones: $L^X_{m|n_1}$ & $L^Y_{m|n_2}$

	$\tilde{P}_1 \perp\!\!\!\perp \tilde{P}_2$ PY processes				$(\tilde{P}_1, \tilde{P}_2)$ HPY process			
	FRUIT 1		FRUIT 2		FRUIT 1		FRUIT 2	
m	$\hat{K}^X_{m\|n_1}$	$\hat{L}^Y_{m\|n_2}$	$\hat{K}^X_{m\|n_1}$	$\hat{L}^Y_{m\|n_2}$	$\hat{K}^X_{m\|n_1}$	$\hat{L}^Y_{m\|n_2}$	$\hat{K}^X_{m\|n_1}$	$\hat{L}^Y_{m\|n_2}$
200	65.4	68.2	117.0	122.0	79.5	82.1	103.3	108.2
400	125.6	136.2	225.5	244.0	154.1	164.2	198.3	216.6
600	181.5	204.3	326.4	366.0	224.5	246.2	286.0	324.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2000	488.1	680.8	891.3	1219.0	631.4	820.9	770.2	1083.1

In the dependent case, finer prediction is possible. For instance, considering additional sequencing for FRUIT 1:

- for $m = 600$, $\hat{K}^X_{m|n_1} = 224.5$, which is the sum of the predicted 37.6 genes new to FRUIT 1 but already observed in FRUIT 2 and 186.9 overall new genes.
- for $m = 2000$, it is predicted that 96.6 genes of the 504 originally observed only for FRUIT 2 will be detected also in the FRUIT 1 library.

Analysis of survival data

Data $\{X_{1,i}\}_{i=1}^{n_1}$ and $\{X_{2,j}\}_{j=1}^{n_2}$ take values in $\mathbb{X} = \mathbb{R}^+$ and denote survival times subject to some censoring mechanism

- $S_1 = 1 - F_1$ and $S_2 = 1 - F_2$ are the survival functions
- When F_ℓ is absolutely continuous

$$S_\ell(t) = \exp\left\{ - \int_0^t h_\ell(s)\,\mathrm{d}s \right\} \qquad h_\ell = \frac{F_\ell'}{1 - F_\ell}$$

and h_ℓ is the hazard rate function of the ℓ-th sample

Analysis of survival data

Data $\{X_{1,i}\}_{i=1}^{n_1}$ and $\{X_{2,j}\}_{j=1}^{n_2}$ take values in $\mathbb{X} = \mathbb{R}^+$ and denote survival times subject to some censoring mechanism

- $S_1 = 1 - F_1$ and $S_2 = 1 - F_2$ are the survival functions
- When F_ℓ is absolutely continuous

$$S_\ell(t) = \exp\left\{ - \int_0^t h_\ell(s)\, \mathrm{d}s \right\} \qquad h_\ell = \frac{F_\ell'}{1 - F_\ell}$$

and h_ℓ is the hazard rate function of the ℓ-th sample

Goal

Estimate S_1 and S_2 or any functional of them (mean lifetime, median lifetime, ...)

- For the exchangeable case (i.e. $S_1 = S_2$), see James (2005)
- For the partially exchangeable case, we address the issue by resorting to hCRMs

$$\text{Prior for } (S_1, S_2) \iff \text{Prior for } (h_1, h_2)$$

Hazard rates' models

Hierarchical CRMs

Let $H_0 = c_0 P_0$ where P_0 is some diffuse probability measure on $\mathbb{Y}$ and

$$(\tilde{\mu}_1, \tilde{\mu}_2) \,|\, \tilde{\mu}_0 \sim \text{CRM}(\tilde{\nu}_1) \times \text{CRM}(\tilde{\nu}_2) \qquad \tilde{\nu}_\ell(\mathrm{d}s, \mathrm{d}y) = \rho_\ell(s)\,\mathrm{d}s\,\tilde{\mu}_0(\mathrm{d}y)$$

$$\tilde{\mu}_0 \sim \text{CRM}(\nu_0) \qquad \nu_0(\mathrm{d}s, \mathrm{d}y) = \rho_0(s)\,\mathrm{d}s\,H_0(\mathrm{d}y)$$

Then $(\tilde{\mu}_1, \tilde{\mu}_2)$ is termed a *hierarchical completely random measure* (hCRM)

Hazard rates' models

Hierarchical CRMs

Let $H_0 = c_0 P_0$ where P_0 is some diffuse probability measure on $\mathbb{Y}$ and

$$(\tilde{\mu}_1, \tilde{\mu}_2) \,|\, \tilde{\mu}_0 \sim \text{CRM}(\tilde{\nu}_1) \times \text{CRM}(\tilde{\nu}_2) \qquad \tilde{\nu}_\ell(\mathrm{d}s, \mathrm{d}y) = \rho_\ell(s)\,\mathrm{d}s\,\tilde{\mu}_0(\mathrm{d}y)$$

$$\tilde{\mu}_0 \sim \text{CRM}(\nu_0) \qquad \nu_0(\mathrm{d}s, \mathrm{d}y) = \rho_0(s)\,\mathrm{d}s\,H_0(\mathrm{d}y)$$

Then $(\tilde{\mu}_1, \tilde{\mu}_2)$ is termed a *hierarchical completely random measure* (hCRM)

Kernel mixture models

- $(\tilde{\mu}_1, \tilde{\mu}_2)$ is a hCRM on $\mathbb{Y}$ and $k(t, y)$ is a kernel on $\mathbb{R}^+ \times \mathbb{Y}$
- Random hazard rates $\tilde{h}_\ell(t) = \int_{\mathbb{Y}} k(t, y)\,\tilde{\mu}_\ell(\mathrm{d}y)$ for $\ell = 1, 2$
- Partially exchangeable survival times $\{X_{1,i}\}_{i=1}^{n_1}$ and $\{X_{2,j}\}_{j=1}^{n_2}$

$$\mathbb{P}\Big[X_{i,1} > t_1,\ X_{j,2} > t_2 \,\Big|\, (\tilde{\mu}_1, \tilde{\mu}_2)\Big] = \exp\Big\{ -\sum_{\ell=1}^{2} \int_0^t \tilde{h}_\ell(s)\,\mathrm{d}s \Big\}$$

- Likelihood: with $K_\ell(y) = \sum_{j=1}^{n_i} \int_0^{X_{i,j}} k(t,y)\,\mathrm{d}t$

$$\mathscr{L}(\mu_1,\mu_2;\boldsymbol{X}) = \prod_{i=1}^{2} \exp\left\{ -\int_{\mathbb{Y}} K_i(y)\mu_i(\mathrm{d}y)\right\} \prod_{j=1}^{n_i} \int_{\mathbb{Y}} k(X_{i,j};y)\,\mu_i(\mathrm{d}y)$$

- Augmented likelihood: introduce latent variables $Y_{i,j}$ generated by a discrete random probability measure

$$\mathscr{L}^*(\mu_1,\mu_2;\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{2} \exp\left\{ -\int_{\mathbb{Y}} K_i(y)\mu_i(\mathrm{d}y)\right\} \prod_{j=1}^{n_i} k(X_{i,j};Y_{i,j})\,\mu_i(\mathrm{d}Y_{i,j})$$

- Likelihood: with $K_\ell(y) = \sum_{j=1}^{n_i} \int_0^{X_{i,j}} k(t,y)\,\mathrm{d}t$

$$\mathscr{L}(\mu_1,\mu_2;\boldsymbol{X}) = \prod_{i=1}^{2} \exp\left\{ -\int_{\mathbb{Y}} K_i(y)\mu_i(\mathrm{d}y)\right\} \prod_{j=1}^{n_i} \int_{\mathbb{Y}} k(X_{i,j};y)\,\mu_i(\mathrm{d}y)$$

- Augmented likelihood: introduce latent variables $Y_{i,j}$ generated by a discrete random probability measure

$$\mathscr{L}^*(\mu_1,\mu_2;\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{2} \exp\left\{ -\int_{\mathbb{Y}} K_i(y)\mu_i(\mathrm{d}y)\right\} \prod_{j=1}^{n_i} k(X_{i,j};Y_{i,j})\,\mu_i(\mathrm{d}Y_{i,j})$$

Posterior characterization

The posterior of $(\tilde{\mu}_1,\tilde{\mu}_2)$, given the data $\boldsymbol{X}$ and the latents $\boldsymbol{Y}$, equals the distribution of the random measure vector

$$\left(\mu_1^* + \sum_{j=1}^{k_1} J_{j,1}^*\,\delta_{Y_{j,1}^*},\ \mu_2^* + \sum_{j=1}^{k_2} J_{j,2}^*\,\delta_{Y_{j,2}^*}\right)$$

- (μ_1^*,μ_2^*) *is a hierarchical CRM with updated marginal Lévy intensities*
- *jumps $J_{j,\ell}^*$ are independent and with known density*

$\Longrightarrow$ The model can be modified so to include censored observations and also other covariates, i.e. a semiparametric Cox proportional hazards type of specification

Illustration: Estimation of the survival functions S_1 & S_2

Simulation study: 2 samples of size $n_1 = n_2 = 100$ generated from Weibull distributions with different parameters chosen s.t. the survival functions cross and do not satisfy the assumption of proportional hazards

$\Longrightarrow$ dependent hierarchical model is able to distinguish the two survival functions

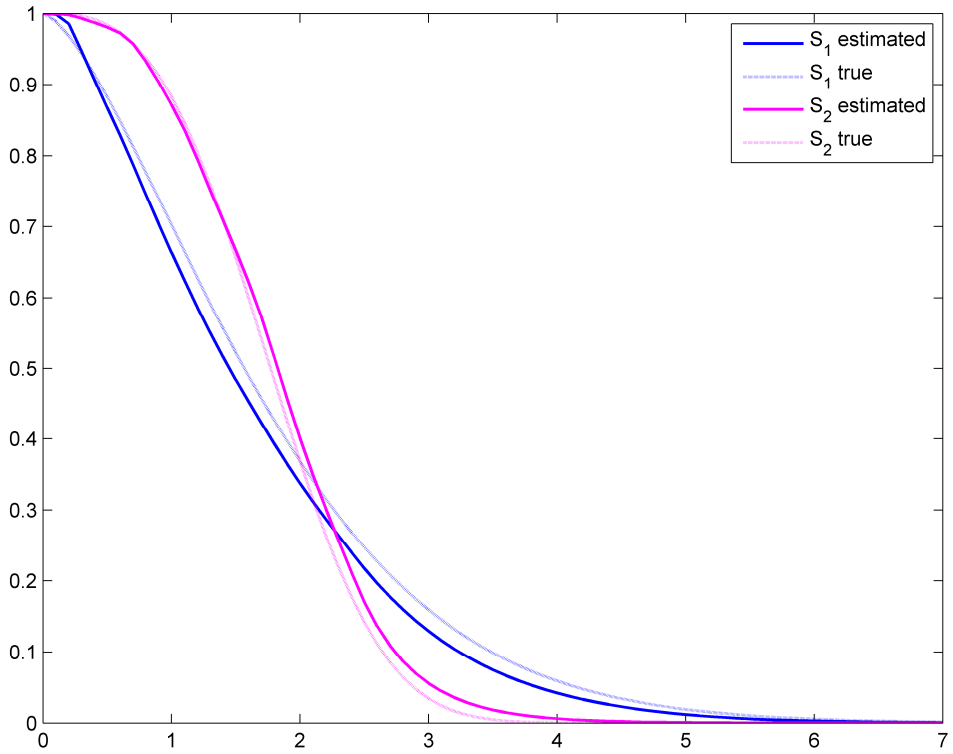


Figure: Estimated and true survival functions

Concluding remarks

- The study of nonparametric models for non–exchangeable data is analytically challenging but not impossible!
- The availability of the pEPPF allows to study important quantities such as the number of clusters K_N, which have an intuitive interpretation. Combined with closed form results on the correlation structure, methodological guidelines on the choice of the parameters can be deduced.
- The posterior characterizations for hierarchical NRMIs and hierarchical hazard rates are the first completely explicit posterior representations in the partially exchangeable case.
- Thanks to the distributional results derivation of marginal and conditional sampling schemes is quite straightforward.

Concluding remarks

- The study of nonparametric models for non–exchangeable data is analytically challenging but not impossible!
- The availability of the pEPPF allows to study important quantities such as the number of clusters K_N, which have an intuitive interpretation. Combined with closed form results on the correlation structure, methodological guidelines on the choice of the parameters can be deduced.
- The posterior characterizations for hierarchical NRMIs and hierarchical hazard rates are the first completely explicit posterior representations in the partially exchangeable case.
- Thanks to the distributional results derivation of marginal and conditional sampling schemes is quite straightforward.
- In general, CRM–based dependent priors look promising: conditionally on a suitable latent structure, they typically display distributional properties reminiscent of those available in the exchangeable case
- Technique for deriving marginal properties and posterior distributions is general and needs only adaptations depending on the specific transformations of the CRMs. For instance, it works for mixture hazards with hierarchical dependence and for nested processes (possibly including an additive layer).

Main References of the Talk

- Papers:

  • Camerlenghi, Lijoi, Orbanz & Prünster (2018). Distribution theory for Hierarchical Processes. *Annal. Statist.*, forthcoming.

  • Camerlenghi, Lijoi & Prünster (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scand. J. Statist.* **45**, 1062-1091.

  • Camerlenghi, Lijoi & Prünster (2018). Survival analysis via hierarchically dependent mixture hazards. *Tech. Report*.

  • Lijoi, Nipoti & Prünster (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20**, 1260–1291.

  • Lijoi, Prünster & Rigon (2018). Sampling hierarchies of discrete random structures. *Submitted*.

- Work in progress:

  • Camerlenghi, Lijoi & Prünster (201x). Algorithms and uncertainty quantification for hierarchical nonparametric priors.

  • Catalano, Lijoi & Prünster (201x). Measuring dependence of nonparametric processes via Wasserstein distance.

Key References

• De Blasi, Favaro, Lijoi, Mena, Prünster and Ruggiero (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE TPAMI*, 37, 212-229.

• de Finetti (1938). Sur la condition d'équivalence partielle. *Act.sci. ind.* **739**, 5-18.

• James (2005). Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *Ann. Statist.* **33**, 1771-1799.

• James, Lijoi & Prünster (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Statist.* **36**, 76-97.

• Kingman (1967). Completely random measures. *Pacific J. Math.* **21**, 59–78.

• MacEachern (1999). Dependent nonparametric processes. In *ASA Proceedings*.

• MacEachern (2000). Dependent Dirichlet processes. *OSU Tech. Report*.

• Müller, Quintana & Rosner (2004). A method for combining inference across related nonparametric Bayesian models. *J. Roy. Statist. Soc. Ser. B* **66**, 735-749.

• Pitman & Yor (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855-900.

• Regazzini, Lijoi & Prünster (2003). Distributional results for means of random measures with independent increments. *Ann. Statist.* **31**, 560-585.

• Rodríguez, Dunson & Gelfand (2008). The nested Dirichlet process. *J. Amer. Statist. Assoc.* **103**, 1131-1144.

• Teh & Jordan (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics*, 158–207. Cambridge University Press.

• Teh, Jordan, Beal & Blei (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 1566–1581.