Fragmentation Coagulation Processes

Konstantina Palla

Reading and Communication Club
Talk based on

- *Modelling Genetic Variations using Fragmentation-Coagulation Processes*

- *Scalable Imputation of Genetic Data with a Discrete Fragmentation-Coagulation Process*
Partition-valued Processes

Mechanisms

Discrete Fragmentation Coagulation Process

Continuous Fragmentation Coagulation Process
Let:

- \([n]\) denote the natural numbers 1, \ldots, n.
- \(\Pi_{[n]}\) is the set of unlabelled partitions of \([n]\)
- Each \(\pi \in \Pi_{[n]}\) is a set of disjoint non-empty clusters indexed by a covariate (time/location); \(\pi_i\)
- Markov chain whose states are partitions of the natural numbers
- When \(t_{i+1} = t_i + dt\) and \(dt \to 0\); Markov jump process
1. Fragmentation
2. Coagulation
3. Fragmentation/Coagulation
\textbf{Fragmentation Mechanism}

- $\pi(0)$: all objects belong to the same cluster
- define: $\pi_{t+1} | \pi_t \sim FRAG_{\alpha, \theta_f}(\pi_t)$
- Partition each cluster $c_i$ of $\pi_t$ further according to $\text{CRP}_{c_i}(\alpha, \theta_f)$, where $i = 1, \ldots, K$ and $K$ is the number of clusters in $\pi_t$ and $\theta_f$ is a parameter governing the fragmentation rate.
- Then $\pi_{t+1}$ is a refinement of $\pi_t$
- Top down generative process for trees

Let $\rho_i$ be the random partitions drawn from each $\text{CRP}_{c_i}(0, \theta_f)$ at state $\pi_t$

- $P(\pi_{t+1} | \pi_t) = \prod_{i=1}^{K} p(\rho_i)$
- For a single cluster in $\pi_t$ $p(\rho_i) = \text{CRP}_{c_i}(0, \theta_f) = \frac{\Gamma(\theta_f)}{\Gamma(\theta_f+n_i)} \theta_f^{K_{i}} \prod_{j=1}^{K_{i}} \Gamma(n_{ij})$, where $K_i$ is the number of clusters in partition $\rho_i$
- Example of discrete time case: nested CRP [Blei, Griffiths, Jordan, 2003]
**Fragmentation Mechanism - Continuous Case**

- Continuous-indexed analogue of nCRP.
- Use of $\theta_f dt$, compute $P(\pi(t + dt)|\pi(t))$ as $dt \to 0$
- For a single cluster in $\pi(t)$:
  
  $$p(\rho_i) = \frac{\theta_f^{K_i-1}}{\Gamma(n_i)} \prod_{j=1}^{K_i} \Gamma(n_{ij}) \lim_{dt \to 0} dt^{K_i-1}$$

- Fragmentation rate of a single cluster to $K_i$ clusters:
  
  $$\theta_f^{K_i-1} \prod_{j=1}^{K_i} \frac{\Gamma(n_{ij})}{\Gamma(n_i)}$$

- Rate of a single cluster $c$ in $\pi(t)$ fragmenting to two clusters $a$ and $b$ ($c = a \cup b$) is for $K_i = 2$: $\theta_f \frac{\Gamma(n_a)\Gamma(n_b)}{\Gamma(n_c)}$

- Example: Dirichlet Diffusion Trees [Neal, 2001]. At a branch point:
  
  - $P($following branch $k) = \frac{n_k}{m}$
  - $P($diverging$) = \frac{\theta_f dt}{m}$
  - $n_k$: number of objects which previously took branch $k$
  - $K$: current number of branches from this branch point
  - $m = \sum_{k=1}^{K} n_k$: number of samples which previously took the current path
  - $P(\pi(t + dt)|\pi(t)) = \frac{1}{1} \frac{\theta_f dt}{2} \frac{1}{3}$
\[ \pi(0): \text{each object is in its own cluster.} \]

\[ \text{define: } \pi_{t+1} \mid \pi_t \sim \text{COAG}_{\alpha, \theta_c}(\pi_t) \]

- Partition the set of clusters of \( \pi_t \) according to a CRP \( \pi_t(\alpha, \theta_c) \), replace each cluster with the union of its elements,

- \( \pi_{t+1} \) is coarser than \( \pi_t \)

- Bottom up generative process for trees

- For \( \alpha = 0 \), \( K \) clusters in \( \pi_{t+1} \) and \( n \) in \( \pi_t \): 
  \[ P(\pi_{t+1} \mid \pi_t) = \text{CRP}_{\pi_t}(\alpha, \theta_c) = \frac{\Gamma(\theta_c)}{\Gamma(\theta_c+n)} \theta_c^K \prod_{i=1}^{K} \Gamma(n_i) \]
Continuous-indexed analogue of the discrete coagulation tree

Kingman’s coalescent [Kingman, 1982]

Define coagulation parameter $\frac{\theta_c}{dt}$ and $\pi(t + dt)|\pi(t) \sim COAG_{\alpha, \frac{\theta_c}{dt}}(\pi(t))$

$K$ and $n$ the number of clusters in $\pi(t + dt)$ and $\pi(t)$

As $dt \to 0$: $P(\pi(t + dt)|\pi(t)) = \theta_c^{K-n} [\prod_{i=1}^{K} \Gamma(n_i)] \lim_{dt \to 0} \frac{dt-K+1}{dt-n+1}$

Rate of two clusters in $\pi(t)$ coagulating to one in $\pi(t + dt)$, for $K = n - 1 : \theta_c^{-1}$
Duality between Pitman-Yor fragmentations and coagulations.

**Theorem [Pitman, 1999]**

For all $0 < \alpha < 1$, $0 \leq \beta < 1$ and $\theta > -\alpha\beta$, the following statements are equivalent

$$
\pi \sim CRP_{[n]}(\alpha\beta, \theta) \quad \text{and} \quad F|\pi \sim FRAG_{\alpha, -\alpha\beta}(c)
$$

$$
\eta \sim CRP_{[n]}(\alpha, \theta) \quad \text{and} \quad \rho|\eta \sim COAG_{\beta, \frac{\theta}{\alpha}}(\eta)
$$

Equivalent: $P(\pi = S, F = T) = P(\rho = S, \eta = T)$

- $P(\pi) = P(\rho) = CRP(\alpha\beta, \theta)$
- $P(F) = P(\eta) = CRP(\alpha, \theta)$

Fragmentation(/coagulation) is the time reversal of coagulation(/fragmentation) (with appropriately chosen parameters).
• A Markov chain over partitions.
• Transition using fragmentation followed by coagulation.
• Assume T steps on Markov chain. Define \((R_t)^{T-1}_{t=1}\); control the dependence between \(\pi_t\) and \(\pi_{t+1}\).
• Duality theorem for \(\beta = 0, \alpha = R\) at each step:

\[
\pi \sim CRP_{[n]}(0, \theta) \quad \text{and} \quad F_c|\pi \sim FRAG_{R,0}(c) \forall c \in \pi
\]

\[
\eta \sim CRP_{[n]}(R, \theta) \quad \text{and} \quad \rho|\eta \sim COAG_{0, \frac{\theta}{R}}(\eta)
\]
Stationary distribution: \( CRP(0, \theta) \) Show detailed balance
\[
P(A)P(A \rightarrow B) = P(B)P(B \rightarrow A)
\]
Hint:
\[
P(A)P(A \rightarrow B) = \sum_C P(A)P(B|C)P(C|A)
\]
and apply duality twice
Reversible markov chain; due to F/C duality
Exchangeable: each $\pi_t$ has CRP marginal
Projectivity: CRP projective
Continuous FCP

- Continuum limit of DFCP; \( \pi(t + dt)|\pi(t) \) when \( dt \to 0 \)
- Time Markov process over partitions, an exchangeable fragmentation-coalescence process [Berestycki, 2004]
- Binary events only: at most one fragmentation or one coagulation at each time
  - One cluster fragments to two, OR
  - Two clusters coagulate to one
- Parameters: \( \theta = \frac{R}{\alpha} \), where \( \alpha > 0 \) and \( R > 0 \) parameters governing the rate of coagulation and fragmentation respectively
**Continuous FCP - cont.**

Why binary events?

**Fragmentation:** for each cluster \( c_i \) in \( \pi(t) \)

- \( \theta_f = R \), probability of fragmenting to \( K_i \) clusters in \( \pi(t + dt) \):

\[
P(\rho_i) = \frac{R^{K_i-1}}{\Gamma(n_i)} \prod_{j=1}^{K_i} \Gamma(n_{ij}) \lim_{dt \to 0} dt^{K_i-1}
\]

\[
= \begin{cases} 
  \mathcal{O}(1), & K_i = 1 \\
  \mathcal{O}(dt), & K_i = 2 \\
  \mathcal{O}(dt^2), & K_i > 2 
\end{cases}
\]

**Coagulation:** \( K \) clusters in \( \pi(t + dt) \), \( n \) in \( \pi(t) \)

- Set \( \theta_c = \frac{1}{\alpha} \)

\[
P(\pi(t + dt) | \pi(t)) = \alpha^{n-K} \prod_{i=1}^{K} \Gamma(n_i) \lim_{dt \to 0} \frac{dt^{-K+1}}{dt^{-n+1}}
\]

\[
= \begin{cases} 
  \mathcal{O}(1), & K = n \\
  \mathcal{O}(dt), & K = n - 1 \\
  \mathcal{O}(dt^2), & K < n - 1 
\end{cases}
\]
• CFCP continuous-time Markov process $\pi = (\pi(t), t \in [0, T])$

• Space of partitions (states) is finite $\rightarrow$ Markov jump process

• Transition rate matrix $Q$
  
  $[q_{ij}]$: the transition rate from state $i$ to $j$

• Total transition rate from state $i$: $q_i = \sum_{j \neq i} q_{ij}$
  
  $q_i = \text{frag}_{\text{rate}} + \text{coag}_{\text{rate}} = R \sum_c H_{nc-1} + \alpha \frac{n(n-1)}{2}$

• Homogeneous Poisson Process: initial state $\pi(0) = s$, mean number of events $q_s T$

• Interarrival time $\sim \exp(q_s)$

• Equilibrium distribution: $CRP(0, \frac{R}{\alpha})$. 
Sequential Generative Process in CFCP

- Incremental construction: describe the law of the path of object $i$ given the paths of $1, \ldots, i-1$ objects
- Let $c_i(t)$ denote the cluster object $i$ belongs to at time $t$.
- $c_1(t) = \text{constant}$
- $t = 0 : \text{CRP}_n(0, \frac{R}{\alpha})$
- $t > 0$: If $i^{th}$ object is in an existing cluster $c$, $c_i(t^-) = c$:
  - if $c$ fragments to $a$ and $b$: $P(c_i(t) | c_i(t^-))$
    $$= \begin{cases} \frac{|a|}{|c| - 1}, & c_i(t) = a \\ \frac{|b|}{|c| - 1}, & c_i(t) = b \end{cases}$$
    Same as DFT’s probability of following a path.
- If $c$ coagulates at time $t$: $P(c_i(t) = c') = 1$
- If no F/C event involving $c$ takes place, $c$ fragments with rate: $\frac{R}{|c| - 1}$
  Same as DFT’s branching probability.
- $t > 0$: If $i^{th}$ object is alone in cluster, $c_i(t^-) = \emptyset$, it joins an existing cluster with rate $\alpha$ and any existing cluster with rate $\alpha |\pi|_{[i-1]}(t)|$
  Same as Kingman’s coalescent.
**Conditional distribution in CFCP - cont.**

\[ \frac{\theta}{\theta+n} \]

Coag, rate: \( \alpha \)

Frag, rate: \( \frac{R}{2} \)

(figure taken from Yee Whye Teh’s slides, MLSS 2011)
CFCP - PROPERTIES

- Rate of fragmentation same as for Dirichlet diffusion trees (with constant rate)
- Rate of coagulation is same as Kingman’s coalescent.
- Markov process is
  - reversible; fragmentation (DFT) is precisely the converse of coagulation (KC).
  - exchangeable; CRP marginals
Thank you!