Contour processes, Coalescent point processes and applications

Amaury Lambert

CIRM, Luminy, le 12 juin 2012
Jumping contour of a tree

a) Binary tree with edge lengths and b) Jumping contour process of its truncation below time $t$. 
Retrieving information from the contour

The depths of the excursions of the JCP away from $T$ are the coalescence times of consecutive extant individuals at time $T$.

- $H_i =$ coalescence time between individuals $i - 1$ and $i$
  = depth of $i$-th excursion of the contour process

- $A_i =$ age of individual $i$
  = undershoot of last jump of...

- $R_i =$ residual lifetime of individual $i$
  = overshoot of last jump of...
Splitting trees

A (time-inhomogeneous) splitting tree (Geiger & Kersting 97) is a random tree model (genealogy, epidemic, phylogeny,...), where:

- particles reproduce singly and independently
- the birth rate $\lambda(t)$ may depend on absolute time $t$ (only)
- lifetime distributions can be general and may also depend on birth time: example of a death rate $\mu(t, a)$ depending on absolute time $t$ and age $a$ of particles.

The population size process $(N_t; t \geq 0)$ is a binary Crump–Mode–Jagers process (with age-independent birth point process).
Theorem (L. (2010))

The jumping contour process of a splitting tree truncated below $T$ is a strong Markov process.

In the time-homogeneous case, it has the same law as a compound Poisson process $X$ with Lévy measure $\lambda P(V \in \cdot)$, without negative jumps and drift $-1$, reflected below $T$ and killed upon hitting $0$. 
Outline

1. Contour
2. Coalescent
3. Bottlenecks
4. Mutations
5. Epidemics
6. Phylogenies
Coalescent point process

\( H^T \) := depth of an excursion of the JCP away from \( T \).

**Corollary**

The coalescent tree (or reconstructed tree) seen from \( T \) of a splitting tree, is a coalescent point process: the coalescence times form a sequence of i.i.d. r.v. distributed as \( H^T \), killed at its first value larger than \( T \).

⇒ Notation:

\[
F_T(s) := \frac{1}{P(H^T \geq s)}.
\]

**Figure**: Illustration of a coalescent point process showing the coalescence depths $H_1, \ldots, H_6$ for each of the 6 consecutive pairs of tips. The depth $H_7$ is the first one larger than $T$. 
Three special cases

1. Time-homogeneous case (L. 2010) \( \equiv \lambda \) and \( \mu(a) \) do NOT depend on \( t \) ...And then \( F_T \) does not depend on \( T \)... 

2. Markovian case (Nee, May & Harvey 1994) \( \equiv \mu(t) \) does NOT depend on \( a \)

\[
F_T(t) = 1 + \int_{T-t}^{T} dx \lambda(x) e^{\int_x^T dy r(y)},
\]

where \( r(t) := \lambda(t) - \mu(t) \) (instantaneous growth rate).

3. Time-homogeneous + Markov (Rannala, 1997) \( \equiv \lambda \) and \( \mu \) are constant \( \equiv \) linear birth–death process

\[
F_T(t) = 1 + \frac{\lambda}{r} (e^{rt} - 1).
\]
Outline

1 Contour
2 Coalescent
3 Bottlenecks
4 Mutations
5 Epidemics
6 Phylogenies
Bottleneck : definition

- Start with a coalescent point process
- add a bottleneck with survival probability $\varepsilon$ at time $s$ backwards, i.e., all lineages crossing this time section are independently deleted with probability $1 - \varepsilon$
- Set $B_T^\varepsilon :=$ coalescence time between two consecutive survivors,
- so that $s = 0$ corresponds to sampling.
Figure: A coalescent point process with a bottleneck at time $T - s$ from present time. All lineages crossing this time are independently deleted with probability $1 - \epsilon$. 
Bottleneck : result

- With probability \( P(H^T < s) \), \( B_\varepsilon^T \) is distributed as \( H^T \) conditional on \( H^T < s \)
- With probability \( P(H^T \geq s) \),

\[
B_\varepsilon^T \overset{(d)}{=} \max\{A_1, \ldots, A_K\},
\]

where the \( A_i \)'s are i.i.d. distributed as \( H^T \) conditional on \( H^T \geq s \)

and

\[
\mathbb{P}(K = j) = \varepsilon (1 - \varepsilon)^{j-1}.
\]

- This yields

\[
F_\varepsilon(t) := \frac{1}{P(B_\varepsilon^T \geq t)} = \begin{cases} 
F_T(t) & \text{if } t < s \\
\varepsilon F_T(t) + (1 - \varepsilon) F_T(s) & \text{if } t \geq s
\end{cases}
\]
More bottlenecks

Start with a coalescent point process and add extra bottlenecks with survival probabilities $\varepsilon_1, \ldots, \varepsilon_k$ at times $T - s_1 > \ldots > T - s_k$ (where $s_1 \geq 0$ and $s_k < T$).

**Proposition (L. (2012))**

*Conditional on survival, the new reconstructed tree is again a coalescent point process with inverse tail distribution $F_\varepsilon$ given by*

$$F_\varepsilon(t) = \varepsilon_1 \cdots \varepsilon_m F_T(t) + \sum_{j=1}^{m} (1 - \varepsilon_j) \varepsilon_1 \cdots \varepsilon_{j-1} F_T(s_j) \quad t \in [s_m, s_{m+1}],$$

*for each $m \in \{0, 1, \ldots, k\}$, with $s_0 := 0$ and $s_{k+1} := T$.***
Outline

1 Contour
2 Coalescent
3 Bottlenecks
4 Mutations
5 Epidemics
6 Phylogenies
Neutral, Poissonian mutations

- **Supercritical**, time-homogeneous, splitting tree
- $N_t :=$ population size at time $t$
- $\alpha :=$ Malthusian parameter $= \lim_{t \to \infty} \frac{1}{t} \log N_t$
- $\theta :=$ mutation rate on lineages.

**Goal.** Characterize the allelic partition under the **infinitely-many alleles model**.
In (L. 2009) and (Champagnat & L. 2012a), we have characterized the clonal coalescent point process to give an explicit expression for the expectation, conditional on $N_t$, of

$$A(k, t, y) := \text{number of alleles of age in } (y, y + dy) \text{ and carried by } k \text{ alive individuals at time } t.$$  

= expected allele frequency spectrum for small families.
Clonal coalescent point process

\[ B_i^\theta = \text{distances between consecutive virgin lineages} \]

\[ H_i^\theta = \max \text{ of branch lengths between consecutive virgin lineages} \]

\( \implies \) \( (B_i^\theta, H_i^\theta) \) are i.i.d.
Largest or oldest families at time $t$

Proposition (Champagnat & L. 2012b)

Assume $\alpha \leq \theta$. The following results hold in expectation.

- If $\alpha < \theta$, there are explicit constants $b$ and $\beta := \theta/(\theta - \alpha)$, such that largest families have sizes $b(\alpha t - \beta \log(t)) + c$ and they all have age $\sim \frac{\log(t)}{\theta - \alpha}$.
  
  Oldest families have ages $\gamma t + a$ and tight sizes, where $\gamma := \alpha/\theta$.

- If $\alpha = \theta$, there are explicit constants $b$ and $\beta := 1/(2\alpha)$, such that largest families have sizes $b(\alpha t - \beta \log(t) + c)^2$ and they all have age $\sim t/2$.
  
  Oldest families have ages $t - \gamma \log(t) + a$ and tight sizes, where $\gamma := 1/\alpha$.

If $\alpha > \theta$, largest families have sizes $ce^{(\alpha-\theta)t}$ and are also the oldest ones (born at times $O(1)$).
Largest or oldest families at time $t$

Proposition (Champagnat & L. 2012b)

Assume $\alpha \leq \theta$. The following results hold in expectation.

- If $\alpha < \theta$, there are explicit constants $b$ and $\beta := \theta / (\theta - \alpha)$, such that largest families have sizes $b(\alpha t - \beta \log(t)) + c$ and they all have age $\sim \frac{\log(t)}{\theta - \alpha}$. Oldest families have ages $\gamma t + a$ and tight sizes, where $\gamma := \alpha / \theta$.

- If $\alpha = \theta$, there are explicit constants $b$ and $\beta := 1/(2\alpha)$, such that largest families have sizes $b(\alpha t - \beta \log(t) + c)^2$ and they all have age $\sim t/2$. Oldest families have ages $t - \gamma \log(t) + a$ and tight sizes, where $\gamma := 1/\alpha$.

If $\alpha > \theta$, largest families have sizes $ce^{(\alpha-\theta)t}$ and are also the oldest ones (born at times $O(1)$).
Convergence in distribution (1)

ASSUME $\alpha < \theta$.

Take the coalescent point process at time $t$, choose $s_t$ such that $s_t \to \infty$, and set

$$N'_{t-s_t} := \text{number of subtrees } (\mathcal{T}_i) \text{ grafted on branch lengths } \geq s_t$$
Convergence in distribution (2)

Set

\[ X_t^{(k)} := \text{size of the } k\text{-th largest family in the whole population} \]

\[ Y_i := \text{size of the largest family in subtree } \mathcal{T}_i. \]

With \( s_t := \log(t) / (\theta - \alpha) \), we have

- \( N_{t-s_t}' \to \infty \)
- \( (X_t^{(1)}, \ldots, X_t^{(k)}) = \text{first } k \text{ order statistics of } \{Y_1, \ldots, Y_{N_t'-s_t}\} \) W.H.P.
- With \( L_t(x) := \text{number of families larger than } x \text{ at time } t \),

\[ \mathbb{P}(Y \geq x_t + c) = \mathbb{P}(L_{s_t}(x_t + c) \geq 1) \sim \mathbb{E}(L_{s_t}(x_t + c)). \]
Convergence in distribution (3)

\[ X_t^{(k)} := \text{size of the } k\text{-th largest family in the whole population} \]

Theorem (Champagnat & L. 2012b)

There is an explicit constant \( c \in (0, 1) \), such that

\[ (X_t^{(k)} - b(\alpha t - \beta \log(t)); k \geq 1) \text{ converge (fdd) to the (ranked) atoms of a mixed Poisson point measure with intensity} \]

\[ \mathcal{E} \sum_{j \in \mathbb{Z}} c^j \delta_j, \]

where \( \mathcal{E} \) is some exponential r.v.
Convergence in distribution (4)

\[ A_t^{(k)} := \text{age of the } k\text{-th oldest family in the whole population} \]

**Theorem (Champagnat & L. 2012b)**

The sequence \( (A_t^{(k)} - (\alpha t / \theta); k \geq 1) \) converges (fdd) to the (ranked) atoms of a mixed Poisson point measure with intensity

\[ \mathcal{E} e^{-\theta a} da, \]

where \( \mathcal{E} \) is some exponential r.v.
Outline

1. Contour
2. Coalescent
3. Bottlenecks
4. Mutations
5. Epidemics
6. Phylogenies
Epidemic model

- Epidemics modelled by a splitting tree, where birth = transmission (rate $\lambda$) and lifetime = period of infectiousness
- each patient can be detected to be a carrier only after an independent exponential clock with parameter $\delta$ running from the beginning of her infection (medical exam or symptoms);
- $T :=$ detection time = first time when one these clocks rings.
Splitting tree with exponential clocks

⇒ Each individual is equipped with an exponential clock with parameter $\delta$ initialized at birth.

$$T := \text{first time when one of these clocks rings.}$$
Let $X^{(T)}$ be the JCP of the splitting tree truncated below the detection time $T$.

**Theorem (L. & Trapman 2012)**

For any $n \geq 1$, for any $t > 0$, for any càdlàg path $e$,

$$
P\left(N_T = n, T \in dt, X^{(T)} \in de\right) = \frac{\delta}{b} e^{-\delta V(e)} P\left(-I_n \in dt, Y'_n \in de\right),$$

where $V(e)$ denotes the total lifetime of a path $e$, $Y_n$ is the concatenation of $n$ i.i.d. excursions of a Lévy process, $I_n$ is its infimum and $Y'_n$ is its Vervaat transform.
Vervaat transformation

\[ H(Y_n) \]

\[ V(Y_n) \]

\[ V(Y_n) - H(Y_n) \]
Methicillin–resistant *Staphylococcus aureus*

- patients have i.i.d lengths of stay in the hospital, all distributed as some r.v. $K$ (such that $E(K) < \infty$);
- **Conditional** on infection, the length of stay of a patient is a size-biased version of $K$;
- At detection time $T$, all patients in the hospital are screened and identified.
Notation

For individual $i$, set

- $U_i :=$ time elapsed from entrance of the hospital up to infection
- $A_i :=$ time elapsed from infection up to $T$
- $R_i :=$ residual lifetime in the hospital after $T$.

Set $m := \mathbb{E}(K)$ and let $\phi$ denote the inverse of the convex function

$$x \mapsto x - \frac{\lambda}{m} \int_{(0, \infty]} (1 - e^{-xy}) \mathbb{P}(K > y) \, dy.$$
Inference from hospital data

Proposition (L. & Trapman 2012)

Conditional on $N_T = n$, the triples $(U_i, A_i, R_i)$ of the $n$ (randomly labelled) carriers at time $T$ are i.i.d., distributed as the r.v. $(U, A, R)$ (independent of $n$), where

$$\mathbb{E}(f(U, A, R)) = \frac{\lambda}{m} \frac{\phi(\delta)}{\phi(\delta) - \delta} \int_0^\infty du \int_0^\infty da \int_{u+a}^\infty \mathbb{P}(K \in dz) e^{-\phi(\delta)a} f(u, a, z-u-a),$$

In particular, the times $J_i = U_i + A_i$ spent in the hospital up to time $T$ are i.i.d., distributed as the r.v. $J$

$$\mathbb{P}(J \in dy) = \frac{\lambda / m}{\phi(\delta) - \delta} \mathbb{P}(K > y) \left(1 - e^{-\phi(\delta)y}\right) dy.$$
Temporally-spaced epidemiological data
(with Tanja Stadler)

- A sampled individual immediately leaves the infective population.
- $S_i :=$ sampling time of individual $i$
- $R_i :=$ coalescence time between individuals $i - 1$ and $i$.

By the contour technique, the $(S_i, R_i)$ is a Markov chain with explicit transitions.
$\Rightarrow$ inference of model parameters from viral phylogenies (HIV, flu).
Splitting tree with exponential clocks (2)

Black dots = sampling/detecting
Phylogenetic tree models
(with H. Morlon, R.S. Etienne, B. Haegeman)

...(statistical) work in progress...

1. **Protracted speciation** (Etienne & Rosindell 2011): New born species are **incipient**, and turn **good** after a random time.

2. **Speciation by genetic differentiation** and point mutation: two individuals are in the same species if their MRCA belongs to a geodesic without mutation.

⇒ **Infer parameters** of diversification dynamics from real phylogenetic tree shapes.
Acknowledgements

• **Stochastics & Biology group**
  ⊂ Laboratoire de Probabilités et Modèles Aléatoires
  ⊂ UPMC University Paris 06

• **Stochastic Models for the Inference of Life Evolution (SMILE)**
  ⊂ Center for Interdisciplinary Research in Biology
  ⊂ Collège de France

• **ANR Modèles Aléatoires eN Écologie, Génétique, Évolution (MANEGE)**