Statistical inference in transport-fragmentation models

Marc Hoffmann

Paris-Dauphine University

Van Dantzig Seminar, 6 March 2015

Acknowledgements

This talk is based on joint projects (some are still in progress!) with

- M. Doumic (INRIA)
- N. Krell (University of Rennes)
- A. Olivier (Paris-Dauphine University)
- P. Reynaud-Bouret (CNRS)
- V. Rivoirard (Paris-Dauphine University)
- L. Robert (INRA)

Context (1/4)

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

- We consider (simple) **branching processes** with **deterministic evolution** between jump times.
- Such models appear as toy models for population growth in cellular biology.
- We wish to **statistically estimate** the parameters of the model, in order to ultimately **discriminate** between different hypotheses related to the mechanisms that trigger cell division.

Context (2/4)

- We structure the model by state variables for each individual like **size, age**, growth rate, DNA content and so on.
- The evolution of the particle system is described by a common mechanism:
  1. Each particle grows by "ingesting a common nutrient" = **deterministic evolution**.
  2. After some time, depending on a structure variable, each particle gives rise to $k = 2$ offsprings by cell division = **branching event**.
- Our goal in this talk: estimate the **branching rate** as a function of age or size (or both).
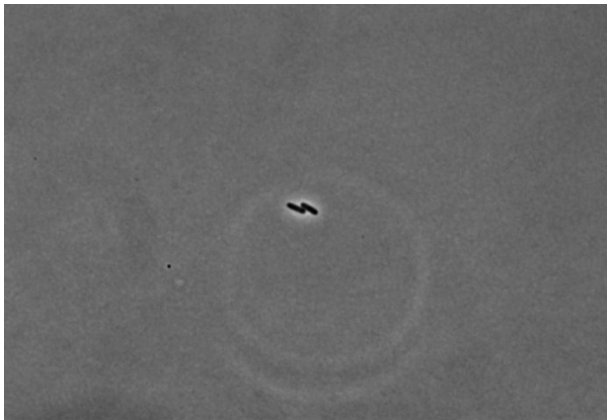
Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Figure : Evolution of a *E. Coli* population.

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Figure : Evolution of a *E. Coli* population.

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Figure : Evolution of a *E. Coli* population.

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
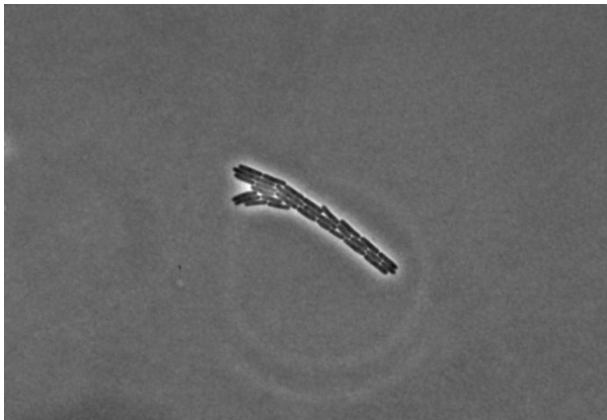DIVISION RATE

Figure : Evolution of a *E. Coli* population.

Figure : Evolution of a *E. Coli* population.

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Figure : Evolution of a *E. Coli* population.

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

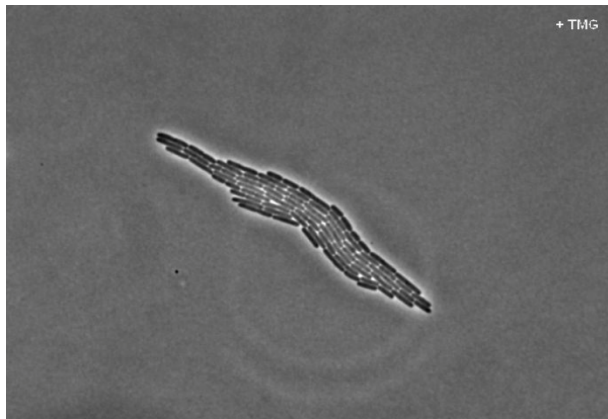Figure : Evolution of a *E. Coli* population.

Figure : Evolution of a *E. Coli* population.

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

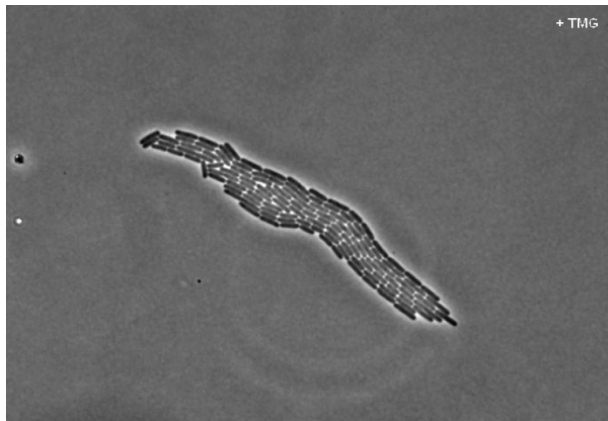ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Figure : Evolution of a *E. Coli* population.

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Figure : Evolution of a *E. Coli* population.

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

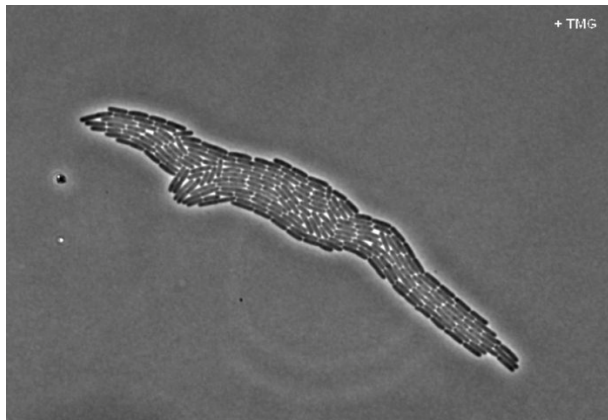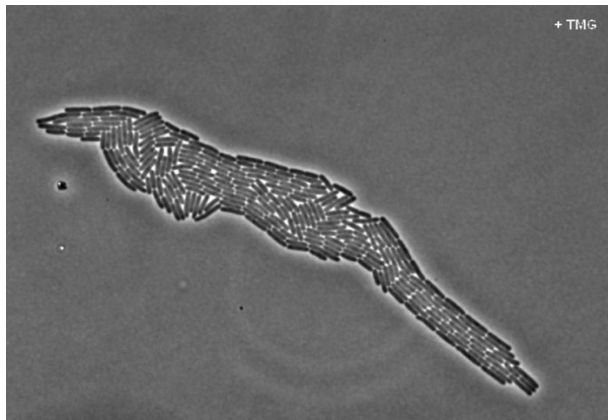ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Figure : Evolution of a *E. Coli* population.

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Context (3/4)

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

- Deterministically the density of structured state variables evolves according to a so-called **fragmentation-transport PDEs**

- Stochastically, the particles evolve according to a **piecewise deterministic Markov process** that evolves along a branching tree.

- We study **nonparametric inference of the division rate**, with the concern of matching deterministic and stochastic approaches.

- I will follow a "pedestrian route" by reviewing some of the results we progressively obtained by "trial-and-error".
- In particular, the results are **highly sensitive** to the choice of the observation schemes (genealogical versus temporal).
- Our control experiments are data sets extracted from the observation of 88 microcolonies of *E. Coli* bacteria cultures (a colony is followed from a single ancestor up to a few hundreds descendants).

Outline

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Genealogical representation

- In the talk we focus on structuring variables that are either age or size.
- The population evolution is associated with an infinite marked binary tree

$$\mathcal{U} = \bigcup_{n=0}^{\infty} \{0,1\}^n \quad \text{with} \quad \{0,1\}^0 := \emptyset.$$

- To each cell or node $u \in \mathcal{U}$, we associate a cell with size at birth given by ξ_u and lifetime ζ_u.
- To each $u \in \mathcal{U}$, we associate a birth time b_u and a time of death d_u so that $\zeta_u = d_u - b_u$.

Observation scheme I: temporal data

- Fix a (large) $T > 0$. Define

$$\mathcal{U}_T = \big\{ u \in \mathcal{U}, b_u \leq T \big\}.$$

- We have $\mathcal{U}_T = \mathring{\mathcal{U}}_T \cup \partial \mathcal{U}_T$, with

$$\mathring{\mathcal{U}}_T = \big\{ u, d_u \leq T \big\} \ \text{ and } \ \partial \mathcal{U}_T = \big\{ u, b_u \leq T < d_u \big\}$$

- We observe

$$\boxed{\big\{ \zeta_u^T \ \text{and/or} \ \xi_u^T, u \in \mathcal{U}_T \big\}}$$

where $\zeta_u^T = \min\{d_u, T\} - b_u$, and $\xi_u^T = \xi_u$ if $d_u \leq T$ and the "size of u at time T" otherwise.

Observation scheme II: genealogical data

- $|u| = n$ if $u = (u_1, \ldots, u_n) \in \mathcal{U}$,
  $uv = (u_1, \ldots, u_n, v_1, \ldots, v_m)$ if $v = (v_1, \ldots, v_m) \in \mathcal{U}$.
- **Sparse tree case** Given $u^{(n)} \in \mathcal{U}$, with $|u^{(n)}| = n$, let

$$\mathcal{U}_{u^{(n)}} = \left\{ u \in \mathcal{U}, uw = u^{(n)} \text{ for some } w \in \mathcal{U} \right\}.$$

We observe

$$\boxed{\left\{ \zeta_u \text{ and/or } \xi_u, u \in \mathcal{U}_{u^{(n)}} \right\}.}$$

- **Full tree case** For $n = 2^{k_n}$, define

$$\mathcal{U}_{[n]} = \{ u \in \mathcal{U}, \ |u| \leq k_n \}.$$

We observe

$$\boxed{\left\{ \xi_u \text{ and/or } \zeta_u, u \in \mathcal{U}_{[n]} \right\}.}$$

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Temporal data

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
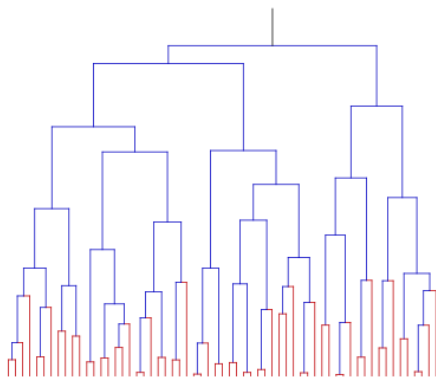THE AGE
DEPENDENT
DIVISION RATE

Figure : *Genealogical tree observed up to $T = 7$ for a time-dependent division rate $B(a) = a^2$ (60 cells). In blue: $\mathring{\mathcal{U}}_T$. In red: $\partial \mathcal{U}_T$.*

Genealogical data

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

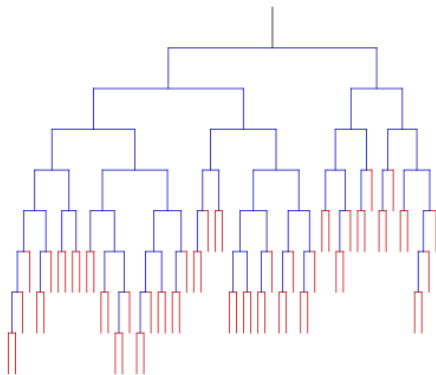ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Figure : *The same outcome organised at a genealogical level.*

Outline

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

Estimation at a
(large) fixed
time in a proxy
model
Estimation
through
genealogical
data

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Size dependent division rate (1/2)

- Perthame, *Transport equations in biology*, Birkäuser, 2006.
- $n(t, x)$: density of cells of size x.
- Parameter of interest: Division rate $B(x)$.
- 1 cell of size x gives birth to 2 cells of size $x/2$.
- The growth of the cell size by nutrient uptake is given by a growth rate $g(x) = \tau x$ in this talk: it follows the deterministic evolution

$$\frac{dX(t)}{dt} = g(X(t))dt$$

Size dependent division rate (2/2)

- The deterministic model: transport-fragmentation equation

$$\partial_t n(t,x) + \partial_x\big(\tau x n(t,x)\big) + B(x)n(t,x) = 4B(2x)n(t,2x)$$

$n(t, x = 0) = 0,\ t > 0$ and $n(0,x) = n^{(0)}(x),\ x \geq 0$.

- obtained by mass conservation law:
  - LHS: density evolution + growth by nutrient + division of cells of size x.
  - RHS: division of cells of size $2x$.

Nonparametric estimation of B: First approach

- Represent the solution of the transport-fragmentation equation in a stationary regime.

- Obtain a reconstruction formula for $B(x)$ via this representation in terms of the steady-state or stationary density of the model.

- Postulate a proxy model where one observes exactly a drawn from the stationary density.

- Transfer standard nonparametric estimation techniques in this setting.

Solution by stable distribution

- Start with the transport-fragmentation equation

$$\partial_t n(t,x) + \partial_x\big(\tau x n(t,x)\big) + B(x)n(t,x) = 4B(2x)n(t,2x)$$

- Ansatz $n(t,x) = e^{\lambda t}N(x)$.

$$\boxed{\partial_x\big(\tau x N(x)\big) + \big(\lambda + B(x)\big)N(x) = 4B(2x)N(2x)}$$

- $N(0) = 0$, $N(x) > 0$ for $x > 0$ and $\int_{[0,\infty)} N(x)dx = 1$.
- Perthame *et al.* (2005) prove $n(t,x) \approx e^{\lambda t}N(x)$ with explicit (fast) rates of convergence (steady-state) under fairly general conditions.

A proxy statistical model (1/4)

- Yields a strategy for the nonparametric estimation of B.
- At time T, the data approximately behave like drawn from $N(x)dx$.
- Recover B through the representation

$$L(N, \lambda) = \mathfrak{L}(BN),$$

with

$$L(f, \lambda)(x) = \partial_x\big(\tau x f(x)\big) + \lambda f(x),$$
$$\mathfrak{L}(f)(x) = 4f(2x) - f(x).$$

- The operator $L(\cdot, \lambda)$ has ill-posedness degree of order 1.

A proxy statistical model (2/4)

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

Estimation at a
(large) fixed
time in a proxy
model

Estimation
through
genealogical
data

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

- We postulate the observation of outcomes of cell size $X_1, \ldots, X_n$ in a stationary regime and that are independent:

$$\mathbb{P}(X_1 \in dx_1, \ldots, X_n \in dx_n) := \prod_{i=1}^{n} N(x_i) dx_i.$$

- We can take advantage of kernel methods in nonparametric estimation.

- τ and λ assumed to be known (or λ_n proxy of λ given within sufficient accuracy).

A proxy statistical model (3/4)

- Reconstruction method:
  1. Construct an estimator $\widehat{L}_n(x)$ of the action

  $$L(N, \lambda)(x) = \partial_x(\tau x N(x)) + \lambda N(x),$$

  2. Build an approximate inverse $\mathfrak{L}_k^{-1}$ of the inverse of $\mathfrak{L}(f)(x) = 4f(2x) - f(x)$.
  3. Use representation

  $$L(N, \lambda) = \mathfrak{L}(BN)$$

  and take as final estimator

  $$\boxed{\widehat{B}_n(x) := \frac{\mathfrak{L}_{k_n}^{-1}(\widehat{L}_n(x))}{\widehat{N}_n(x)}}$$

  where $\widehat{N}_n(x) = n^{-1} \sum_{i=1}^n h_n^{-1} K(h_n^{-1}(x - X_i))$ kernel estimator of $N(x)$ for an approriate bandwidth $h_n > 0$.

A proxy statistical model (4/4)

- In Doumic, H, Rivoirard and Reynaud-Bouret (2011), we construct an approximate inverse $\mathfrak{L}_k^{-1}$ such that

$$\|\mathfrak{L}_k^{-1}(\varphi) - \mathfrak{L}^{-1}(\varphi)\|_{L^2(\mathcal{D})} \lesssim k^{-1/2}\|\varphi\|_{H^1}$$

and reconstruct $L(N, \lambda)(x)$ by kernel methods. We obtain an estimator $\widehat{B}_n$ s.t.

$$\left(\mathbb{E}\left[\|\widehat{B}_n - B\|_{L^2(\mathcal{D})}^2\right]\right)^{1/2} \lesssim n^{-s/(2s+3)}$$

uniformly in B over Sobolev balls (over the compact $\mathcal{D} \subset (0, \infty)$).

- The result is compatible with previous deterministic results by Perthame and collaborators.

Limitations of the deterministic based approach

- We implicitly assume a stationary regime (the steady-state approximation).
- We do not take advantage of richer available observation schemes. I particular, if we have access of the finer structure of the tree, can we beat the ill-posedness imposed by our approach?
- And more: constant growth rate, assuming two (sibling) offsprings are of the same size at birth, etc.

The stochastic (cell level) approach (1/3)

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

- We start with a singe cell of size x_0. The cell grows exponentially according to a constant rate τ.
- The mother cell gives rize to two offsprings, at a rate $B(x)$ that depend on its size x.
- The two offsprings have initial size $x_1/2$, where x_1 is the size of the mother at division.
- The two offsprings start independent growth according to the rate τ and divide according to the rate $B(x)$.

The stochastic (cell level) approach (2/3)

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

- To each node $u \in \mathcal{U}$, we associate a cell with size at birth given by ξ_u and lifetime ζ_u.
- u^- denotes the parent of u. Thus

$$2\xi_u = \xi_{u^-} \exp\left(\tau \zeta_{u^-}\right).$$

- $X(t) = \left(X_1(t), X_2(t), \dots\right)$ process of the sizes of the population at time t.

The stochastic approach (3/3)

- $X(t) \leftrightarrow$ finite point measure valued process $\sum_{i=1}^{\sharp X(t)} \delta_{X_i(t)}$
- Identity between point measures

$$\sum_{i=1}^{\infty} \mathbf{1}_{\{X_i(t)>0\}} \delta_{X_i(t)} = \sum_{u \in \mathcal{U}} \delta_{\xi_u e^{\tau(t-b_u)}} \mathbf{1}_{\{b_u \leq t < b_u + \zeta_u\}}.$$

- In particular, observing $(X(t), t \in [0, T])$ is equivalent to observing $\{\xi_u^T, \zeta_u^T, u \in \mathcal{U}_T\}$.

Matching det. and stoch. approaches (1/3)

- We can relate $X(t)$ and $n(t, x)$ via so-called many-to-one formulae.
- Classical technique for fragmentation and branching processes (see *e.g.* Bansaye *et al.* 2009, Bertoin, 2006, Cloez 2011): Pick a cell at random at each division and follow its size $\chi(t)$ through time. For $\xi_\varnothing = x$

$$\boxed{\chi(t) = x \frac{e^{\tau t}}{2^{N_t}}}$$

where N_t is the number of divisions of the tagged fragment up to time t.

- **Step 1** for every (regular compactly supported) f:

$$\mathbb{E}\Big[\sum_{i=1}^{\infty} f\big(X_i(t)\big)\Big] = \mathbb{E}\Big[\sum_{u\in\mathcal{U}} f\big(\xi_t^u\big)\Big]$$

- **Step 2** : many-to-one formula

$$\mathbb{E}\Big[f\big(\chi(t)\big)\Big] = \mathbb{E}\Big[\sum_{u\in\mathcal{U}} \xi_t^u \frac{e^{-\tau t}}{x} f\big(\xi_t^u\big)\Big]$$

- **Step 3** Finally

$$\mathbb{E}\Big[\frac{f\big(\chi(t)\big)}{\chi(t)} x e^{\tau t}\Big] = \mathbb{E}\Big[\sum_{i=1}^{\infty} f\big(X_i(t)\big)\Big].$$

Transport-fragmentation equation

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL
Estimation at a
(large) fixed
time in a proxy
model
Estimation
through
genealogical
data

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

- Set, for (regular compactly supported) f

$$\langle n(t,\cdot), f \rangle := \mathbb{E}\Big[\sum_{i=1}^{\infty} f\big(X_i(t)\big) \Big].$$

We have (in a weak sense)

$$\partial_t n(t,x) + \partial_x \big(\tau x \, n(t,x)\big) + B(x)n(t,x) = 4B(2x)n(t,2x).$$

- Therefore the mean empirical distribution of $X(t)$ satisfies the deterministic transport-fragmentation equation.

Statistical estimation of $B(x)$

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL
Estimation at a
(large) fixed
time in a proxy
model
Estimation
through
genealogical
data

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

- **Observation scheme**: genealogical data from two possible schemes:

  - Sparse tree: we observe, for some $u^{(n)}$ with $|u^{(n)}| = n$,

    $$\{\xi_u, uw = u^{(n)} \text{ for some } w \in \mathcal{U}\}$$

  - Full tree: we observe, for $n = 2^{k_n}$,

    $$\{\xi_u, |u| \leq k_n\}$$

- **Asymptotics**: $n \to \infty$.

Statistical estimation: identifying $B(x)$

- We have

$$\mathbb{P}(\zeta_u \in [t, t+dt] \,|\, \zeta_u \geq t, \xi_u = x) = B(xe^{\tau t})dt$$

from which we obtain the density of the lifetime ζ_{u^-}
conditional on $\xi_{u^-} = x$:

$$\boxed{t \rightsquigarrow B(xe^{\tau t}) \exp\Big(-\int_0^t B(xe^{\tau s})ds\Big).}$$

Toward a Markov kernel

- Using $2\xi_u = \xi_{u^-}\exp\left(\tau\zeta_{u^-}\right)$, we further infer

$$\mathbb{P}\left(\xi_u \in dx' \,\middle|\, \xi_{u^-} = x\right)$$
$$= \frac{B(2x')}{\tau x'}\mathbf{1}_{\{x' \geq x/2\}}\exp\left(-\int_{x/2}^{x'}\frac{B(2s)}{\tau s}ds\right)dx'.$$

- We thus obtain a simple an explicit representation for the transition kernel $\mathcal{P}_B\left(x, dx'\right) = \mathcal{P}_B\left(x, x'\right)dx'$:

$$\boxed{\mathcal{P}_B\left(x, x'\right) = \frac{B(2x')}{\tau x'}\mathbf{1}_{\{x' \geq x/2\}}\exp\left(-\int_{x/2}^{x'}\frac{B(2s)}{\tau s}ds\right).}$$

Assumptions on B

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

Estimation at a
(large) fixed
time in a proxy
model

Estimation
through
genealogical
data

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

- Under appropriate conditions on B, the Markov chain on $(0, \infty)$ is geometrically ergodic: there exists a unique invariant probability $\nu_B(dx) = \nu_B(x)dx$ on $[0, \infty)$ such that

$$\nu_B \mathcal{P}_B = \nu_B.$$

(the chain is however not reversible.)

- More precisely, we have the contraction property

$$\sup_{|g| \leq V} \left| \mathcal{P}_B^k g(x) - \int_{\mathcal{S}} g(z) \nu_B(z) dz \right| \leq R V(x) \gamma^k$$

for an appropriate Lyapunov function V and some (explicitly computable) $\gamma < 1$.

Identifying $B(x)$ through the invariant measure

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL
Estimation at a
(large) fixed
time in a proxy
model
Estimation
through
genealogical
data

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

- Expand the equation $\nu_B \mathcal{P}_B = \nu_B$:

$$\nu_B(y) = \int_0^\infty \nu_B(x) \mathcal{P}_B(x, y) \, dx$$

$$= \frac{B(2y)}{\tau y} \int_0^{2y} \nu_B(x) \exp\left(-\int_{x/2}^y \frac{B(2s)}{\tau s} \, ds\right) dx$$

$$= \frac{B(2y)}{\tau y} \int_0^\infty \int_0^\infty \mathbf{1}_{\{x \le 2y, s \ge y\}} \nu_B(x) \, \mathcal{P}_B(x, s) \, ds dx.$$

- This yields the key representation

$$\boxed{\nu_B(y) = \frac{B(2y)}{\tau y} \, \mathbb{P}_{\nu_B}\left(\xi_{u^-} \le 2y, \, \xi_u \ge y\right).}$$

Key representation

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL
Estimation at a
(large) fixed
time in a proxy
model
Estimation
through
genealogical
data

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

- We conclude

$$B(y) = \frac{\tau y}{2} \frac{\nu_B(y/2)}{\mathbb{P}_{\nu_B}\left(\xi_u^- \leq y, \xi_u \geq y/2\right)}.$$

- This yields the estimator

$$\widehat{B}_n(y) \;\; = \frac{\tau y}{2} \frac{n^{-1} \sum_{u \in \mathcal{U}_{[n]}} K_{h_n}(\xi_u - y/2)}{n^{-1} \sum_{u \in \mathcal{U}_{[n]}} \mathbf{1}_{\{\xi_{u^-} \leq y, \xi_u \geq y/2\}} \bigvee \varpi_n},$$

where the kernel $K_{h_n}(y) = h^{-1}K\left(h_n^{-1}y\right)$ is specified with an appropriate bandwidth (and technical thershold ϖ_n).

- Under the previous assumptions ($+$ the additional condition $\gamma < \frac{1}{2}$ for the geometric ergodicity decay in the full tree case), we have

$$\mathbb{E}_\mu \left[\| \widehat{B}_n - B \|^2_{L^2(\mathcal{D})} \right]^{1/2} \lesssim (\log n)^{1/2} n^{-s/(2s+1)}$$

uniformly in B over s-smooth Hölder balls intersected with "nice geometrically ergodic classes".

- Here, μ is any initial condition so that V^2 is μ-integrable.

Statistical inference in transport-fragmentation models

Marc Hoffmann

GENEALOGICAL VERSUS TEMPORAL DATA

THE SIZE DEPENDENT DIVISION RATE MODEL

Estimation at a (large) fixed time in a proxy model

Estimation through genealogical data

ESTIMATING THE AGE DEPENDENT DIVISION RATE

Remarks and extensions

- Smoothness adaptation (by means of appropriate concentration inequalities on trees)
- The rate are minimax (which is of course no surprise).
- (Possible extension: variability in the growth rate: extension to a cell-dependent $\tau = \tau_u$ drawn via a Markov kernel $\kappa(\tau_{u^-}, d\tau)$.)
- (Possible extension: the cell mother divides into offsprings of different sizes.)

Effect of variability (sparse tree case)

n=2047, B(x)=x^2, the growth rate is uniform on [0.5, 1.5], sparse tree

- ⋯⋯ distribution of all cell sizes
- —— distribution of size at division
- —— true division rate
- —— estimated division rate with variability
- —— estimated division rate without variability

Statistical
inference in
transport-
fragmentation
models

Marc
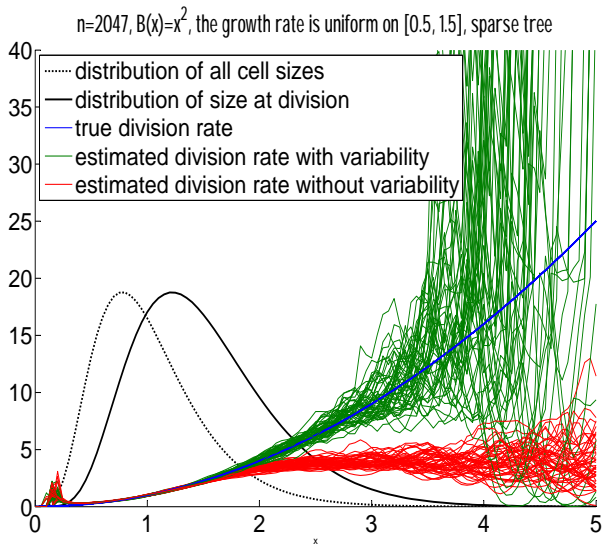Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

Estimation at a
(large) fixed
time in a proxy
model

Estimation
through
genealogical
data

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Effect of variability (dense tree case)

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann
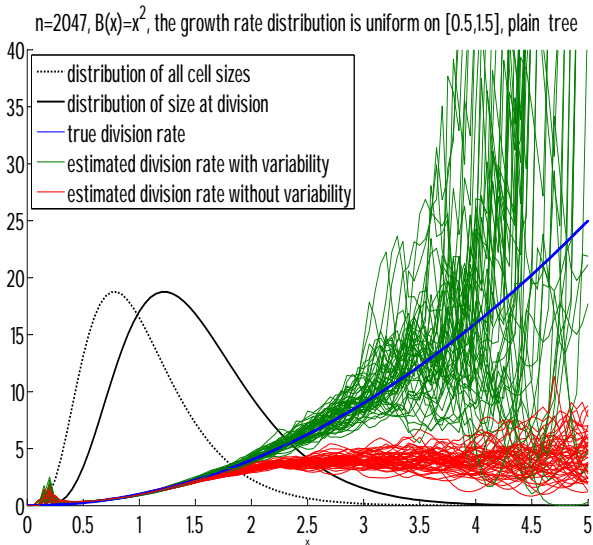
GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

Estimation at a
(large) fixed
time in a proxy
model

Estimation
through
genealogical
data

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

n=2047, B(x)=x^2, the growth rate distribution is uniform on [0.5,1.5], plain tree

- .... distribution of all cell sizes
- —— distribution of size at division
- —— true division rate
- —— estimated division rate with variability
- —— estimated division rate without variability

Une légère surprise (1/3)

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL
Estimation at a
(large) fixed
time in a proxy
model
Estimation
through
genealogical
data

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

- Revisit the representation formula

$$B(y) = \frac{\tau y}{2} \frac{\nu_B(y/2)}{\mathbb{P}_{\nu_B}\left(\xi_{u^-} \leq y, \ \xi_u \geq y/2\right)}.$$

- We always have $\{\xi_{u^-} \geq y\} \subset \{\xi_u \geq y/2\}$, hence

$$\mathbb{P}_{\nu_B}\left(\xi_{u^-} \leq y, \xi_u \geq y/2\right) = \mathbb{P}_{\nu_B}\left(\xi_u \geq y/2\right) - \mathbb{P}_{\nu_B}(\xi_{u^-} \geq y)$$

$$= \int_{y/2}^{\infty} - \int_{y}^{\infty}$$

$$= \int_{y/2}^{y} \nu_B(x) dx \quad (!).$$

Une légère surprise (2/3)

- Finally (for constant growth rate) we have

$$B(y) = \frac{\tau y}{2} \frac{\nu_B(y/2)}{\int_{y/2}^{y} \nu_B(x)dx}$$

- We have a "gain": rate $n^{-s/(2s+1)}$ versus $n^{-s/(2s+3)}$ in the proxy model based on the transport-fragmentation equation...
- But it only comes from the fact that we estimate the invariant measure "at division", versus the invariant measure "at fixed time" in the proxy model.

Une légère surprise (3/3)

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

- There seems to be more "nonparametric statistical information" in data extracted from $\mathring{\mathcal{U}}_T$ rather than $\partial \mathcal{U}_T$
- However $|\mathring{\mathcal{U}}_T| \approx |\partial \mathcal{U}_T|$ (supercritical branching processes).
- Can we make that argument more precise (up to changing the model)?

Outline

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Age dependent division rate $B(a)$

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

- $n(t, a)$ is now solution to

$$\partial_t n(t, a) + \partial_a \big(an(t, a)\big) + B(a)n(t, a) = 0,$$

$$n(t, a = 0) = 2 \int_0^\infty B(a)n(t, a)da \quad n(t = 0, a) = n^{(0)}(a).$$

- This translates into the stochastic model as

$$\mathbb{P}(\zeta_u \in [a, a + da] \,\big|\, \zeta_u \geq a) = B(a)da.$$

- Here, the ζ_u are i.i.d. We have nothing but a renewal process on a tree.

Observation scheme

- The ζ_u are i.i.d.: the case of genealogical data is readily embedded into standard density estimation.

- **Temporal data**: we observe, for some (large) $T > 0$

$$\left\{ \zeta_u^T, u \in \mathcal{U}_T \right\}$$

which can be split into two data sets

$$\left\{ \zeta_u, u \in \mathring{\mathcal{U}}_T \right\} \cup \left\{ T - b_u, u \in \partial \mathcal{U}_T \right\}.$$

- Analogue of what we did for the size dependent $B(x)$ in the sense that we have (empirical) access to the time at division.

- Additional difficulty: bias selection (small lifetimes are observed more often than large lifetimes).

- Strategy: many-to-one formulae (Bansaye *et al.*, 2009, Cloez, 2012)

- Many-to-one formula (Cloez, 2012): we have, for a nice test function g:

$$\mathbb{E}\Big[\sum_{u\in\mathring{\mathcal{U}}_T} g(\zeta_u)\Big] = \int_0^T e^{\lambda_B s}\,\mathbb{E}\big[g(\chi(s))H_B\big(\chi(s)\big)\big]\,ds.$$

where $\chi(t)$ is a tagged branch picked at random on the tree, and $H_B(a)$ an explicit function.

- Also $\mathbb{E}[|\mathring{\mathcal{U}}_T|] \sim \kappa_B e^{\lambda_B T}$.

- All the ingredients needed for a law of large numbers.

Estimation of $B(a)$ from $\mathring{\mathcal{U}}_T$ (3/4)

- Let $f_B(a) = B(a) \exp\left(- \int_0^\infty B(s)ds \right)$.
- We have

$$\frac{1}{|\mathring{\mathcal{U}}_T|} \sum_{u \in \mathring{\mathcal{U}}_T} g(\zeta_u) \overset{\mathbb{P}}{\to} 2 \int_0^\infty g(a) e^{\lambda_B a} f_B(a) da.$$

- We even obtain a rate of convergence (in probability)

$$\left(\exp(\lambda_B T) \right)^{1/2}$$

with some uniformity in $B \in \mathcal{B}$ (in a "neighbourhood" of constant functions B).

- Proof: rates of convergence in the many-to-one formula for $g(\zeta_u, \zeta_v)$ for $u, v \in \mathring{\mathcal{U}}_T$ + geometric ergodicity.

- We derive kernel estimators that achieve the rate

$$\big(\exp(\lambda_B T) \big)^{s/(2s+1)}$$

uniformly over $\mathcal{B} \cap \mathcal{H}(s, M)$.

- The rate is nearly minimax (use likelihood expansions established by Löcherbach in the early 2000's).

What if data are taken from $\partial \mathcal{U}_T$ solely?

- We now have (using Cloez's many-to-one formulae), for a test function g

$$|\partial \mathcal{U}_T|^{-1} \sum_{u \in \partial \mathcal{U}_T} g(\zeta_u) \xrightarrow{\mathbb{P}} 2\lambda_B \int_0^\infty g(a) e^{\lambda_B a} \frac{f_B(a)}{B(a)} da$$

$$= 2\lambda_B \int_0^\infty g(a) e^{\lambda_B a} e^{-\int_0^a B(s) ds} da.$$

- We have a rate of convergence (in probability) $\left(\exp(\lambda_B T) \right)^{1/2}$ uniformly in $B \in \mathcal{B}$.

- We retrieve an ill-posed problem of order 1, leading to concergence rate $\left(\exp(\lambda_B T) \right)^{s/(2s+3)}$.

The age dependent model, simulated data

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL
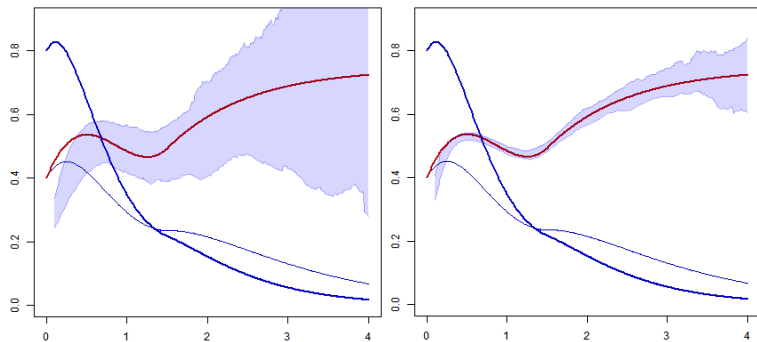
ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Figure : *Reconstruction of B over $\mathcal{D} = [0.1, 4]$ with 95%-level confidence bands constructed over $M = 100$ Monte-Carlo trees. In bold red line: $x \rightsquigarrow B(x)$; in bold blue line: f_{H_B}; in blue line: f_B. Left: $T = 15$. Right: $T = 23$.*

Conclusion/Overall picture

data	Size model	Age model
proxy model	$n^{-s/(2s+3)}$ + adaptation	irrelevant
$\partial \mathcal{U}_T$	?	$(e^{\lambda_B T})^{-s/(2s+3)}$
genealogical	$n^{-s/(2s+1)}$ + adaptation	irrelevant
$\mathring{\mathcal{U}}_T$	?	$(e^{\lambda_B T})^{-s/(2s+1)}$

Statistical
inference in
transport-
fragmentation
models
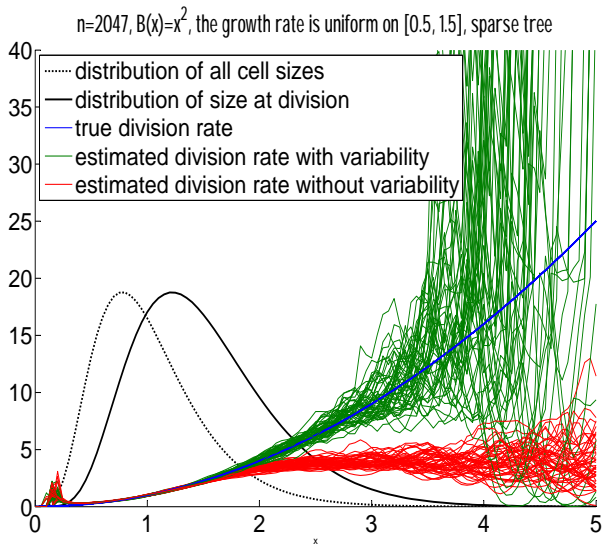
Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

THANK YOU FOR YOUR ATTENTION!

- Doumic, M., H. M., Reynaud-Bouret, P. and Rivoirard, V. (2012) *Nonparametric estimation of the division rate of a size-structured population*. SIAM Journal on Numerical Analysis. **50**, 25pp.

- Doumic, M., H.,M., Krell, N. and Robert, L. (2013) *Statistical estimation of a growth-fragmentation model observed on a genealogical tree. Bernoulli*, in press.

- L Robert, M.H., N. Krell, S. Aymerich, J. Robert and M. Doumic. (2014) *Division control in Escherichia coli is based on a size-sensing rather than a timing mechanism*. BMC Biology, 02/2014 10pp.

- M.H., Olivier, A. (2014) *Nonparametric estimation of the division rate of an age dependent branching process*. arXiv:1412.5936. 32pp.

Effect of variability (sparse tree case)

Statistical
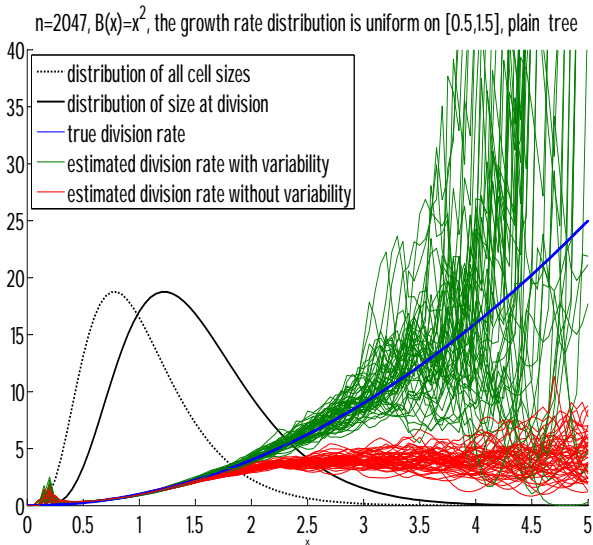inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

$n=2047$, $B(x)=x^2$, the growth rate is uniform on $[0.5, 1.5]$, sparse tree

- $\cdots\cdots$ distribution of all cell sizes
- —— distribution of size at division
- —— true division rate
- —— estimated division rate with variability
- —— estimated division rate without variability

Effect of variability (dense tree case)
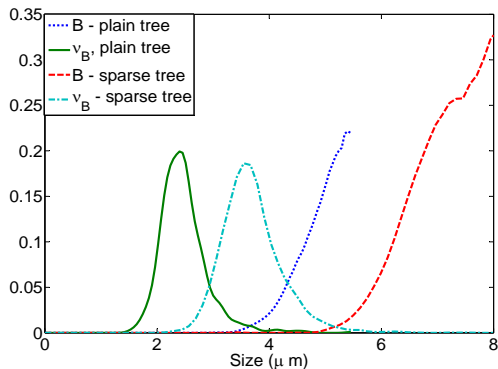
Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

n=2047, B(x)=x^2, the growth rate distribution is uniform on [0.5,1.5], plain tree

- distribution of all cell sizes
- distribution of size at division
- true division rate
- estimated division rate with variability
- estimated division rate without variability

Exploration on real data (*E. Coli*, sparse and dense tree case)
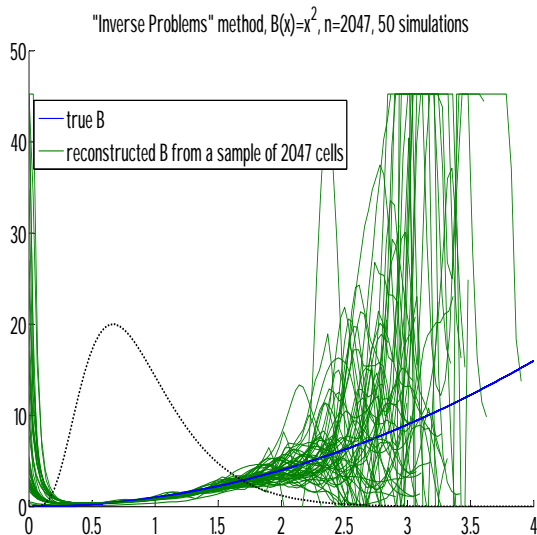
Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Figure : Implementation on real data

Comparison with the inverse problem approach

"Inverse Problems" method, $B(x)=x^2$, n=2047, 50 simulations

true B
reconstructed B from a sample of 2047 cells

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE
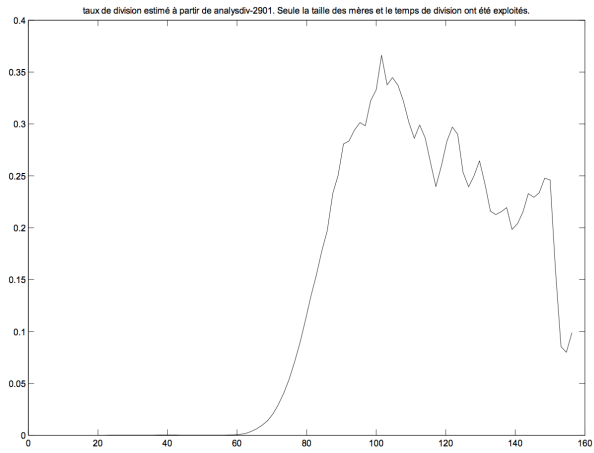
Numerical implementation

Statistical
inference in
transport-
fragmentation
models

Marc
Hoffmann

GENEALOGICAL
VERSUS
TEMPORAL
DATA

THE SIZE
DEPENDENT
DIVISION RATE
MODEL

ESTIMATING
THE AGE
DEPENDENT
DIVISION RATE

Figure : Exploration on real-data. Sparse tree, $n \approx 3000$.