Limit theorems for Bifurcating Auto-Regressive processes with missing data

Benoîte de Saporta Anne Gégout-Petit Laurence Marsalle

Université de Bordeaux and INRIA Bordeaux Sud-Ouest Université de Lille 1

APS 2011, Stockholm

Outline

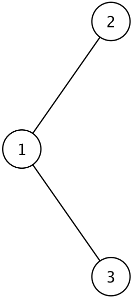
Modelisation of cell lineage data

Escherichia coli

1

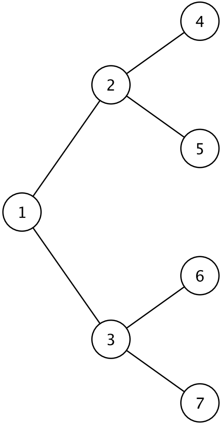
Modelisation of cell lineage data

Escherichia coli



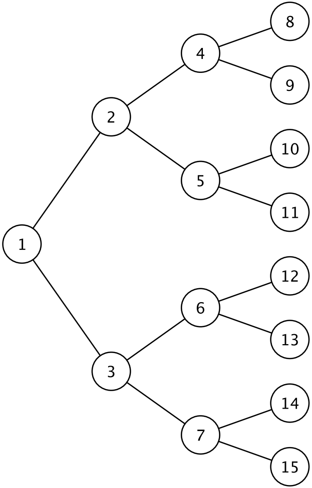
Modelisation of cell lineage data

Escherichia coli

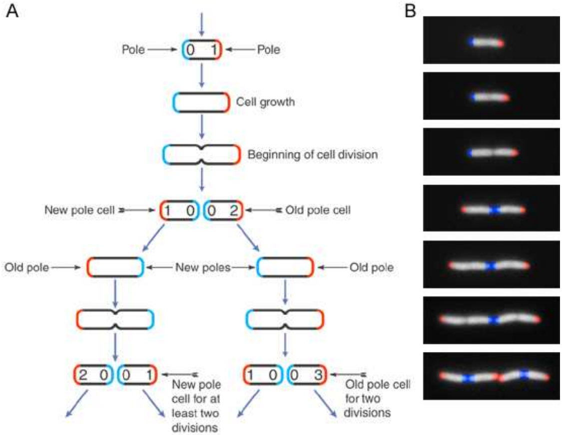


Modelisation of cell lineage data

Escherichia coli

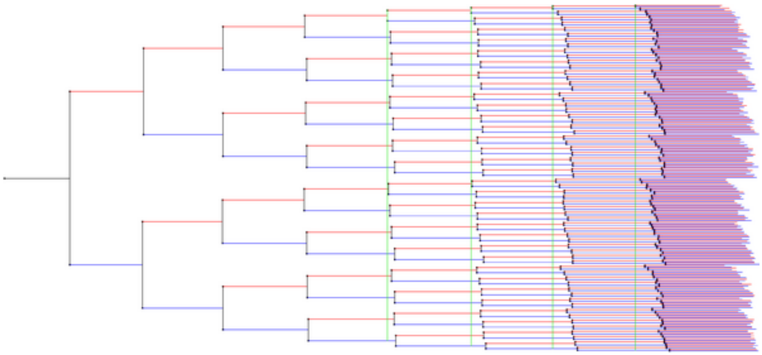


Division of Escherichia coli



From Stewart et al. *PLoS Biol.* 2005.

Do single cell organisms age?



Mean growth rate of E. coli for 94 genealogies up to 9 generations

From Stewart et al. *PLoS Biol.* 2005.

Aim of the talk

Aims

- propose a bifurcating auto-regressive model to study lineages one by one
- take possibly missing data into account
- study the asymptotic properties of the estimators of the parameters
- present symmetry tests
- investigate simulated and real data

Asymmetric BAR process

Bifurcating autoregressive process BAR = autoregressive process indexed by a binary tree.

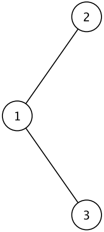
$$\begin{cases} X_{2n} = a + b\,X_n + \varepsilon_{2n} \\ X_{2n+1} = c + d\,X_n + \varepsilon_{2n+1} \end{cases}$$

- X_1 ancestor
- $(\varepsilon_{2n}, \varepsilon_{2n+1})$ noise
- $0 < \max(|b|, |d|) < 1$

Asymmetric BAR process

Bifurcating autoregressive process BAR = autoregressive process indexed by a binary tree.

$$\begin{cases} X_{2n} = a + b\,X_n + \varepsilon_{2n} \\ X_{2n+1} = c + d\,X_n + \varepsilon_{2n+1} \end{cases}$$

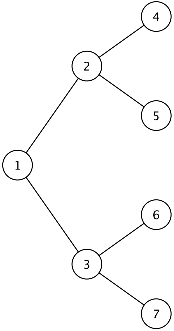


- X_1 ancestor
- $(\varepsilon_{2n}, \varepsilon_{2n+1})$ noise
- $0 < \max(|b|, |d|) < 1$

Asymmetric BAR process

Bifurcating autoregressive process BAR = autoregressive process indexed by a binary tree.

$$\begin{cases} X_{2n} = a + b\, X_n + \varepsilon_{2n} \\ X_{2n+1} = c + d\, X_n + \varepsilon_{2n+1} \end{cases}$$

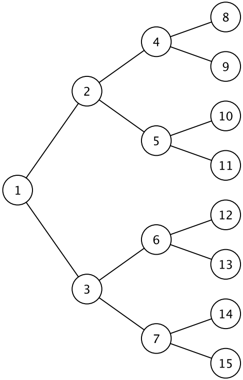


- X_1 ancestor
- $(\varepsilon_{2n}, \varepsilon_{2n+1})$ noise
- $0 < \max(|b|, |d|) < 1$

Asymmetric BAR process

Bifurcating autoregressive process BAR = autoregressive process indexed by a binary tree.

$$\begin{cases} X_{2n} = a + b\,X_n + \varepsilon_{2n} \\ X_{2n+1} = c + d\,X_n + \varepsilon_{2n+1} \end{cases}$$

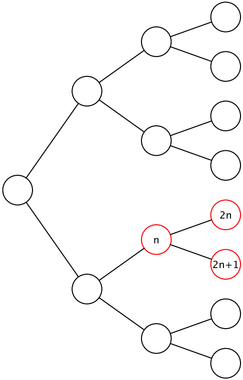


- X_1 ancestor
- $(\varepsilon_{2n}, \varepsilon_{2n+1})$ noise
- $0 < \max(|b|, |d|) < 1$

Asymmetric BAR process

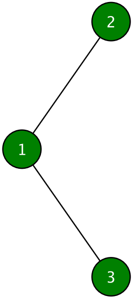
Bifurcating autoregressive process BAR = autoregressive process indexed by a binary tree.

$$\begin{cases} X_{2n} = a + b\,X_n + \varepsilon_{2n} \\ X_{2n+1} = c + d\,X_n + \varepsilon_{2n+1} \end{cases}$$



- X_1 ancestor
- $(\varepsilon_{2n}, \varepsilon_{2n+1})$ noise
- $0 < \max(|b|, |d|) < 1$

Missing data

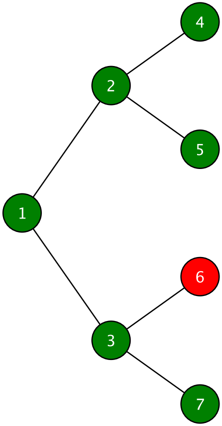


Observation process: $(\delta_n)_{n>0}$, taking values in $\{0, 1\}$

X_n observed if $\delta_n = 1$, else $\delta_n = 0$

If a cell is not observed, all its descendants are not observed.

Missing data

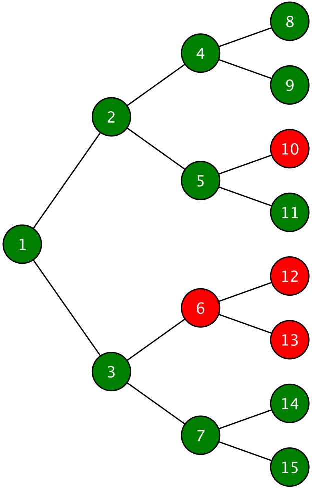


Observation process: $(\delta_n)_{n>0}$, taking values in $\{0, 1\}$

X_n observed if $\delta_n = 1$, else $\delta_n = 0$

If a cell is not observed, all its descendants are not observed.

Missing data

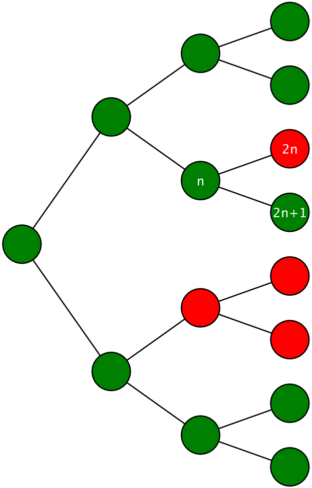


Observation process: $(\delta_n)_{n>0}$, taking values in $\{0,1\}$

X_n observed if $\delta_n = 1$, else $\delta_n = 0$

If a cell is not observed, all its descendants are not observed.

Missing data



Observation process: $(\delta_n)_{n>0}$, taking values in $\{0,1\}$

X_n observed if $\delta_n = 1$, else $\delta_n = 0$

If a cell is not observed, all its descendants are not observed.

Two-type Galton-Watson process

Reproduction laws

- all individuals reproduce independently
- the reproduction laws depend on the type of the mother and daughter
- possible asymmetry in the reproduction laws

$p^0(0,0)$: proba that an even mother has 0 even and 0 odd daughter
$p^0(1,0)$: proba that an even mother has 1 even and 0 odd daughter
$p^0(0,1)$: proba that an even mother has 0 even and 1 odd daughter
$p^0(1,1)$: proba that an even mother has 1 even and 1 odd daughter

Two-type Galton-Watson process

Reproduction laws

- all individuals reproduce independently
- the reproduction laws depend on the type of the mother and daughter
- possible asymmetry in the reproduction laws

$p^1(0,0)$: proba that an odd mother has 0 even and 0 odd daughter
$p^1(1,0)$: proba that an odd mother has 1 even and 0 odd daughter
$p^1(0,1)$: proba that an odd mother has 0 even and 1 odd daughter
$p^1(1,1)$: proba that an odd mother has 1 even and 1 odd daughter

Extinction

Descendance matrix

$$P = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$$

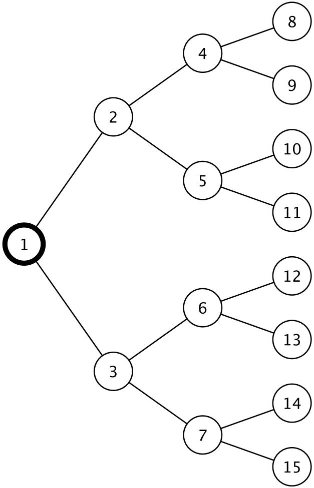
$p_{i0} = p^i(1,0) + p^i(1,1)$: expected number of even daughters of an individual of type i
$p_{i1} = p^i(0,1) + p^i(1,1)$: expected number of odd daughters of an individual of type i

Criteria for extinction

π spectral radius of P

- if $\pi < 1$, extinction is almost sure
- if $\pi \geq 1$, extinction has probability <1

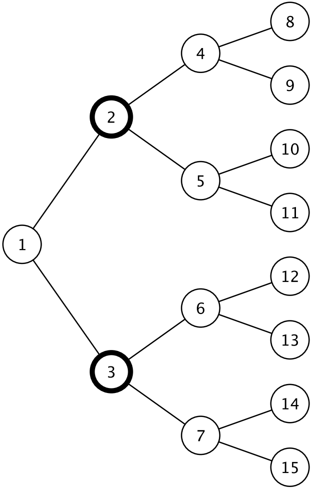
Generations



Fully observed generation 0:

$$\mathbb{G}_0 = \{1\}$$

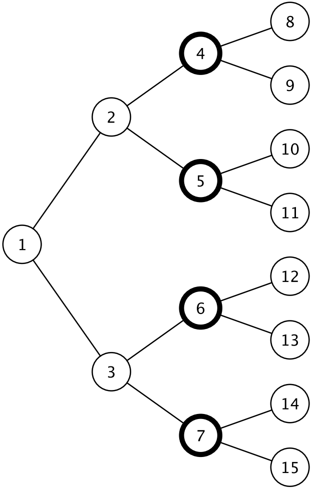
Generations



Fully observed generation 1:

$$\mathbb{G}_1 = \{2, 3\}$$

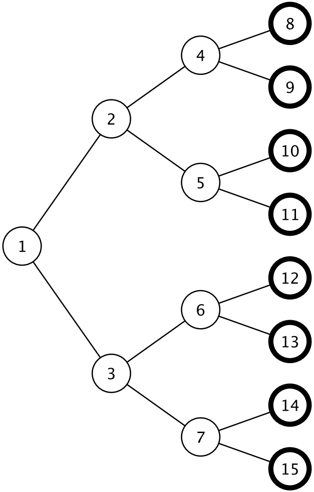
Generations



Fully observed generation 2:

$$\mathbb{G}_2 = \{4, 5, 6, 7\}$$

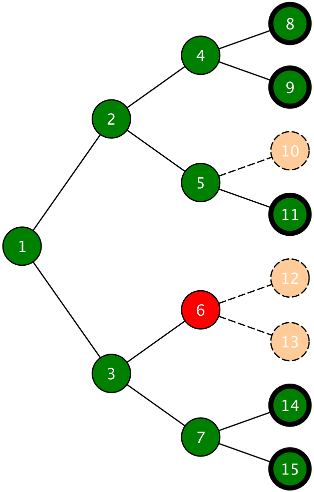
Generations



Fully observed generation n:

$$\mathbb{G}_n = \{2^n, 2^n + 1, \ldots, 2^{n+1} - 1\}$$

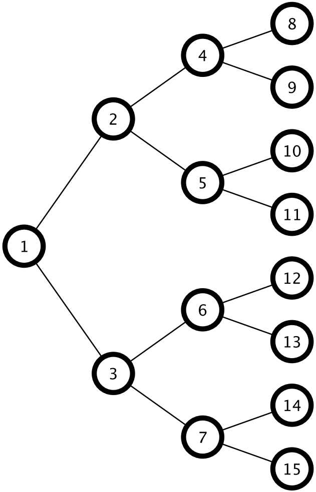
Generations



Partially observed generation n:

$$\mathbb{G}_n^* = \{k \in \mathbb{G}_n \ ; \ \delta_k = 1\}$$

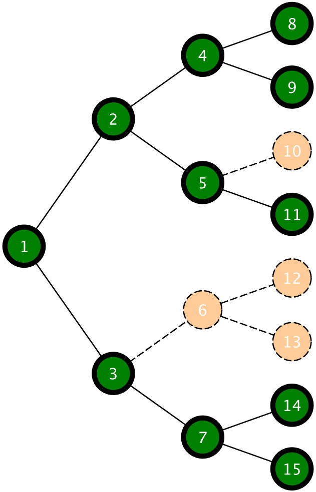
Generations



Fully observed tree up to generation n:

$$\mathbb{T}_n = \cup_{\ell=0}^{n} \mathbb{G}_\ell$$

Generations



Partially observed tree up to generation n:

$$\mathbb{T}_n^* = \{k \in \mathbb{T}_n \ ; \ \delta_k = 1\} = \cup_{\ell=0}^n \mathbb{G}_\ell^*$$

Assumptions

BAR model

$$\begin{cases} X_{2n} = a + b\,X_n + \varepsilon_{2n} \\ X_{2n+1} = c + d\,X_n + \varepsilon_{2n+1} \end{cases}$$

Assumption

- independence between (δ_k) and (X_k) and $(\varepsilon_{2k}, \varepsilon_{2k+1})$

Estimation of $\theta = (a, b, c, d)^t$

Least squares estimator minimizes

$$\Delta_n(\theta) = \frac{1}{2} \sum_{k \in \mathbb{T}_{n-1}} \delta_{2k}(X_{2k} - a - bX_k)^2 + \delta_{2k+1}(X_{2k+1} - c - dX_k)^2.$$

Least squares estimator

LS Estimator for θ

$$\widehat{\boldsymbol{\theta}}_n = \begin{pmatrix} \widehat{a}_n \\ \widehat{b}_n \\ \widehat{c}_n \\ \widehat{d}_n \end{pmatrix} = \boldsymbol{\Sigma}_{n-1}^{-1} \sum_{k \in \mathbb{T}_{n-1}} \begin{pmatrix} \delta_{2k} X_{2k} \\ \delta_{2k} X_k X_{2k} \\ \delta_{2k+1} X_{2k+1} \\ \delta_{2k+1} X_k X_{2k+1} \end{pmatrix}$$

where

$$\boldsymbol{\Sigma}_n = \begin{pmatrix} \boldsymbol{S}_n^0 & 0 \\ 0 & \boldsymbol{S}_n^1 \end{pmatrix}$$

$$\boldsymbol{S}_n^0 = \sum_{k \in \mathbb{T}_n} \delta_{2k} \begin{pmatrix} 1 & X_k \\ X_k & X_k^2 \end{pmatrix} \qquad \boldsymbol{S}_n^1 = \sum_{k \in \mathbb{T}_n} \delta_{2k+1} \begin{pmatrix} 1 & X_k \\ X_k & X_k^2 \end{pmatrix}$$

Strong consistency with convergence rate

Theorem

Under moment assumptions on the noise sequence

$$\mathbb{1}_{\{|\mathbb{G}_n^*|>0\}} \parallel \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} \parallel^2 = \mathbb{1}_{\{|\mathbb{G}_n^*|>0\}} \mathcal{O}\left(\frac{\log|\mathbb{T}_{n-1}^*|}{|\mathbb{T}_{n-1}^*|}\right)$$

Proof: martingales

Main martingale

$\widehat{\theta}_n - \theta = \Sigma_{n-1}^{-1} M_n$, where (M_n) is a martingale for the filtration of generations and the whole observation process

$$M_n = \sum_{k \in \mathbb{T}_{n-1}} \begin{pmatrix} \delta_{2k}\varepsilon_{2k} \\ \delta_{2k}X_k\varepsilon_{2k} \\ \delta_{2k+1}\varepsilon_{2k+1} \\ \delta_{2k+1}X_k\varepsilon_{2k+1} \end{pmatrix}$$

$(M_n)_{n\geq 1}$ square integrable with increasing process $< M >_n = \Gamma_{n-1}$

$$\Gamma_n = \begin{pmatrix} \sigma^2 S_n^0 & \rho S_n^{0,1} \\ \rho S_n^{0,1} & \sigma^2 S_n^1 \end{pmatrix} \quad \text{and} \quad S_n^{0,1} = \sum_{k \in \mathbb{T}_n} \delta_{2k}\delta_{2k+1} \begin{pmatrix} 1 & X_k \\ X_k & X_k^2 \end{pmatrix}$$

Martingale convergence results

(M_n) scalar $\mathcal{F}$-martingale bounded in L^2
$\Delta M_{n+1} = M_{n+1} - M_n$
Increasing process $< M >_n = \sum_{k=0}^{n} \mathbb{E}[(\Delta M_{n+1})^2 \mid \mathcal{F}_n]$

Convergence of scalar L^2 martingales

If $\lim_{n\to\infty} < M >_n = +\infty$, then $\frac{M_n}{<M>_n} \to 0$ a.s.
+ moment conditions then $\left(\frac{M_n}{<M>_n}\right)^2 = \mathcal{O}\left(\frac{log(<M>_n)}{<M>_n}\right)$ a.s.

Similar results for vector-valued martingales.
Here $< M >_n$ is a 4×4-matrix

Convergence of the increasing process

Theorem

$\mathbb{1}_{\{|\mathbb{G}_n^*|>0\}} \frac{<M>_n}{|\mathbb{T}_n^*|}$ converges a.s. to a fixed definite positive matrix.

Sketch of the proof

Laws of large numbers for

- the observation process (δ_k)
- the observed noise $(\delta_k \varepsilon_k)$
- the observed BAR $(\delta_{2k} X_k)$, $\delta_{2k+1} X_k$, $\delta_{2k} X_k^2, \ldots$

via martingale methods for various filtrations and using the auto-regressive structure

Central limit theorem

$\overline{\mathcal{E}}$ non-extinction set, complementary set of

$$\mathcal{E} = \bigcup_{n\geq 1}\{|\mathbb{G}_n^*| = 0\}.$$

New probability

$$\mathbb{P}_{\overline{\mathcal{E}}}(A) = \frac{\mathbb{P}(A \cap \overline{\mathcal{E}})}{\mathbb{P}(\overline{\mathcal{E}})}$$

Theorem

$$\sqrt{|\mathbb{T}_{n-1}^*|}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}) \quad \text{on} \quad (\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}}).$$

Test for the coefficients (a, b) vs (c, d)

H0: $(a, b) = (c, d)$ vs **H1**: $(a, b) \neq (c, d)$
Test statistic :

$$\boldsymbol{Y}_n^2 = |\mathbb{T}_{n-1}^*|(\widehat{a}_n - \widehat{c}_n, \widehat{b}_n - \widehat{d}_n)\Delta^{-1}(\widehat{a}_n - \widehat{c}_n, \widehat{b}_n - \widehat{d}_n)^t$$

where

$$\Delta = \boldsymbol{dg}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}\boldsymbol{dg} \quad \text{and} \quad \boldsymbol{dg} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}^t$$

Theorem

Under the null hypothesis ***(H0)***

$$\boldsymbol{Y}_n^2 \xrightarrow{\mathcal{L}} \chi^2(2) \qquad \text{on } (\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$$

and under the alternative hypothesis ***(H1)***

$$\lim_{n\to\infty} \|\boldsymbol{Y}_n^2\| = +\infty \qquad \text{a.s. on } (\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$$

Test for the fixed points $a/(1-b)$ vs $c/(1-d)$

H0: $a/(1-b) = c/(1-d)$ vs **H1**: $a/(1-b) \neq c/(1-d)$
Test statistic

$$Y_n^2 = |\mathbb{T}_{n-1}^*| \Delta^{-1} \big(\widehat{a}_n/(1-\widehat{b}_n) - \widehat{c}_n/(1-\widehat{d}_n) \big)^2$$

where

$$\Delta = \boldsymbol{dg}^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma} \boldsymbol{\Sigma}^{-1} \boldsymbol{dg} \quad \text{and} \quad \boldsymbol{dg} = \big(1/(1-b), a/(1-b)^2, -1/(1-d), -c/(1-d)^2 \big)^t$$

Theorem

Under the null hypothesis ***(H0)***

$$Y_n^2 \xrightarrow{\mathcal{L}} \chi^2(1) \qquad \text{on } (\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$$

and under the alternative hypothesis ***(H1)***, one has

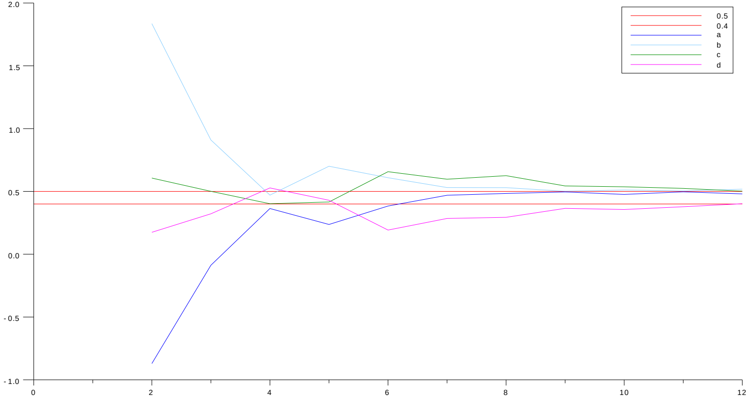
$$\lim_{n\to\infty} Y_n^2 = +\infty \qquad \text{a.s. on } (\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$$

Simulated data - parameters

Galton-Watson	$p^0(0,0)$	$p^0(0,1)$	$p^0(1,0)$	$p^0(1,1)$
symmetric case	0.02	0.04	0.04	0.90
non symmetric case	0.02	0.04	0.04	0.90
	$p^1(0,0)$	$p^1(0,1)$	$p^1(1,0)$	$p^1(1,1)$
symmetric case	0.02	0.04	0.04	0.90
non symmetric case	0.015	0.075	0.075	0.0835

BAR	a	b	c	d
symmetric case	0.5	0.5	0.5	0.5
non symmetric case	0.5	0.5	0.5	0.4

Simulated data - estimation (non symmetric case)



Simulated data - Test (a, b) vs (c, d) -1000 simulations

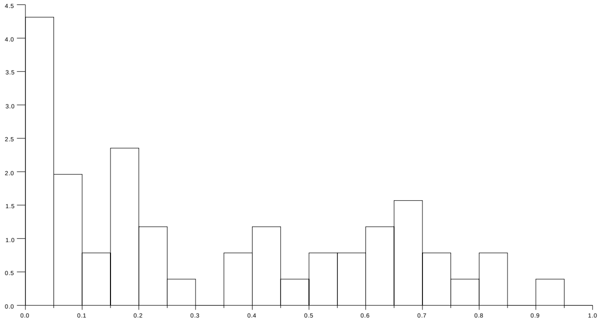
number n of generations	H0 % p-value<0.05	H1 % pvalue<0.05
$n = 7$	37.4	6.6
$n = 8$	53.6	5.5
$n = 9$	71.1	5.5
$n = 10$	86.8	6.3
$n = 11$	95.7	5.9

Simulated data - Test $a/(1-b)$ vs $c/(1-d)$ - 1000 simulations

number n of generations	H0 % p-value<0.05	H1 % p-value<0.05
$n = 7$	23.1	2.2
$n = 8$	41.3	3.3
$n = 9$	64.6	3.8
$n = 10$	82.9	4.7
$n = 11$	94.5	5.5

Real data - growth rate of Escherichia coli

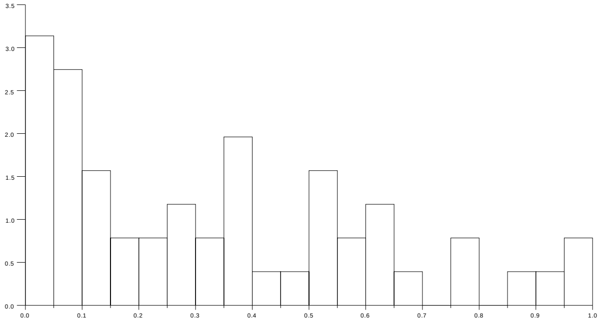
51 genealogies of cells dividing between 8 and 9 times



p-values First test (a, b) vs (c, d)

Real data - growth rate of Escherichia coli

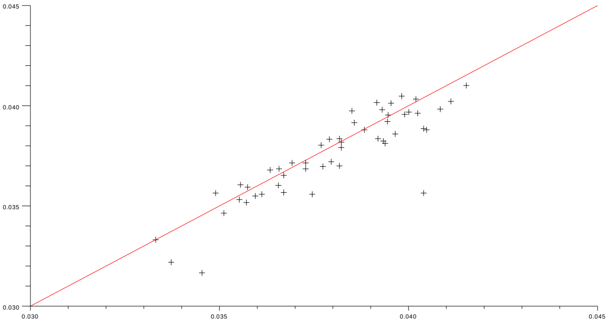
51 genealogies of cells dividing between 8 and 9 times



p-values Second test $a/(1-b)$ vs $c/(1-d)$

Real data - growth rate of Escherichia coli

51 genealogies of cells dividing between 8 and 9 times



Fixed points

References

- B. Bercu, B. de Saporta, and A. Gégout-Petit, *Electron. J. Probab.*, 2009
- R. Cowan and R. G. Staudte, *Biometrics*, 1986
- J.-F. Delmas and L. Marsalle, *Stoch. Process. and Appl.*, 2010.
- B. de Saporta, A. Gégout-Petit, and L. Marsalle, *Arxiv 1012.2012*
- J. Guyon, *Ann. Appl. Probab.*, 2007.
- T. E. Harris, *The theory of branching processes*, 1963.
- E.J. Stewart, R. Madden, and F. Taddei, *PLoS Biol.*, 2005.