Bifurcating autoregressive processes and cell division data

Benoîte de Saporta, Anne Gégout-Petit, Laurence Marsalle

University of Bordeaux | University of Lille

University of Bordeaux
Inria CQFD
France

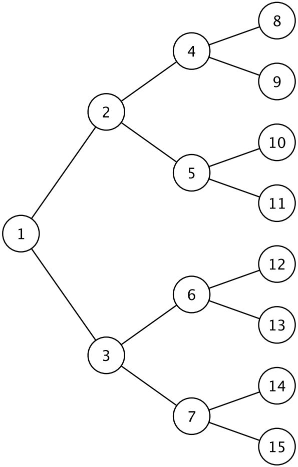
University of Hong Kong – 24 April 2013

Outline

Cell division

film

Escherichia coli

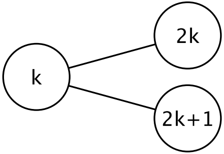


Observation genealogical tree
Originality dependence structure

First BAR model

[Cowan & Staudte 1986] Bifurcating AutoRegressive model

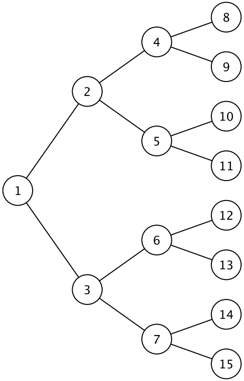
$$\begin{cases} X_{2k} &= a + bX_k + \epsilon_{2k} \\ X_{2k+1} &= a + bX_k + \epsilon_{2k+1} \end{cases}$$



$(\epsilon_{2k}, \epsilon_{2k+1})$ gaussian iid
$\mathbb{E}[\epsilon_{2k+i}] = \sigma^2$, $\mathbb{E}[\epsilon_{2k}\epsilon_{2k+1}] = \rho$
stationary regime if $X_1 \sim \mathcal{N}(\frac{a}{1-b}, \frac{\sigma^2}{1-b^2})$

First BAR model

[Cowan & Staudte 1986] Bifurcating AutoRegressive model

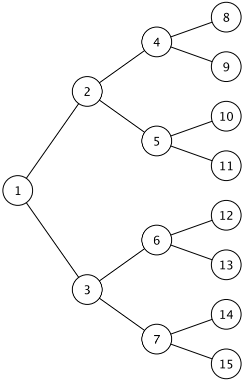


$$\begin{cases} X_{2k} & = & a + bX_k + \epsilon_{2k} \\ X_{2k+1} & = & a + bX_k + \epsilon_{2k+1} \end{cases}$$

$(\epsilon_{2k}, \epsilon_{2k+1})$ gaussian iid
$\mathbb{E}[\epsilon_{2k+i}] = \sigma^2$, $\mathbb{E}[\epsilon_{2k}\epsilon_{2k+1}] = \rho$
stationary regime if $X_1 \sim \mathcal{N}(\frac{a}{1-b}, \frac{\sigma^2}{1-b^2})$

First BAR model

[Cowan & Staudte 1986] Bifurcating AutoRegressive model



$$\begin{cases} X_{2k} &= a + bX_k + \epsilon_{2k} \\ X_{2k+1} &= a + bX_k + \epsilon_{2k+1} \end{cases}$$

$(\epsilon_{2k}, \epsilon_{2k+1})$ gaussian iid
$\mathbb{E}[\epsilon_{2k+i}] = \sigma^2$, $\mathbb{E}[\epsilon_{2k}\epsilon_{2k+1}] = \rho$
stationary regime if $X_1 \sim \mathcal{N}(\frac{a}{1-b}, \frac{\sigma^2}{1-b^2})$

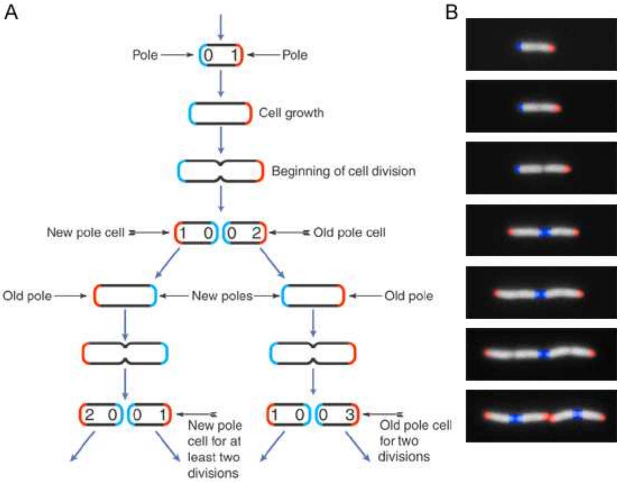
Estimate the parameters to measure correlations

- b mother-daughter correlation
- $\phi = b^2 + (1-b^2)\rho/\sigma^2$ sister-sister correlation

Asymmetry in cell division

[Stewart & al. 2005]

Do single cell organisms age ?



Asymmetric BAR process

[Guyon 2007] Asymmetric model

$$\begin{cases} X_{2k} = a + bX_k + \epsilon_{2k} \\ X_{2k+1} = c + dX_k + \epsilon_{2k+1} \end{cases}$$

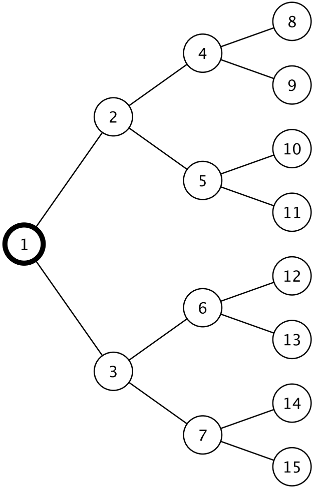
$(\epsilon_{2k}, \epsilon_{2k+1})$ gaussian iid, $\mathbb{E}[\epsilon_{2k+i}] = \sigma^2$, $\mathbb{E}[\epsilon_{2k}\epsilon_{2k+1}] = \rho$
no stationarity

Estimate the parameters to test symmetry

- $(a, b) = (c, d)$?
- $a/(1 - b) = c/(1 - d)$?

Bifurcating Markov chains approach with generation-wise tree structure

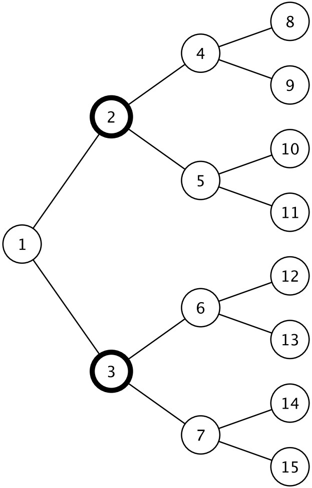
Generations



Generation 0:

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

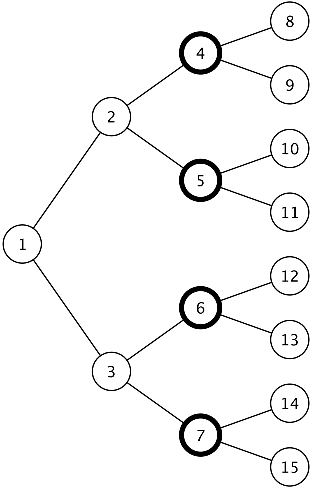
Generations



Generation 1:

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

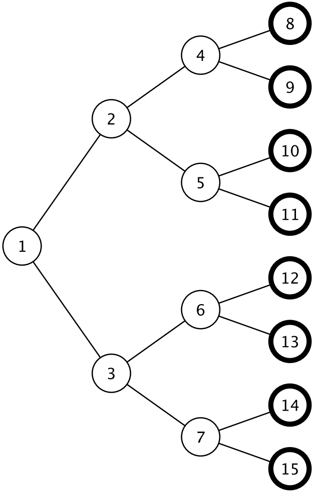
Generations



Generation 2:

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

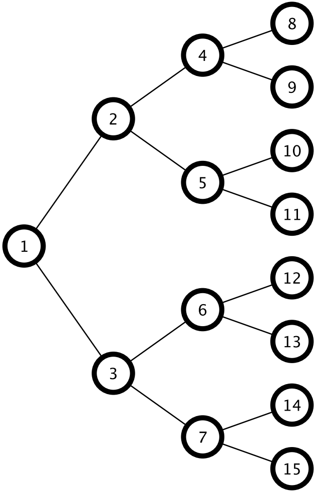
Generations



Generation n:

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

Generations



Tree up to Generation n:

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

Bifurcating Markov chains

- definition of a Markov model on a binary tree

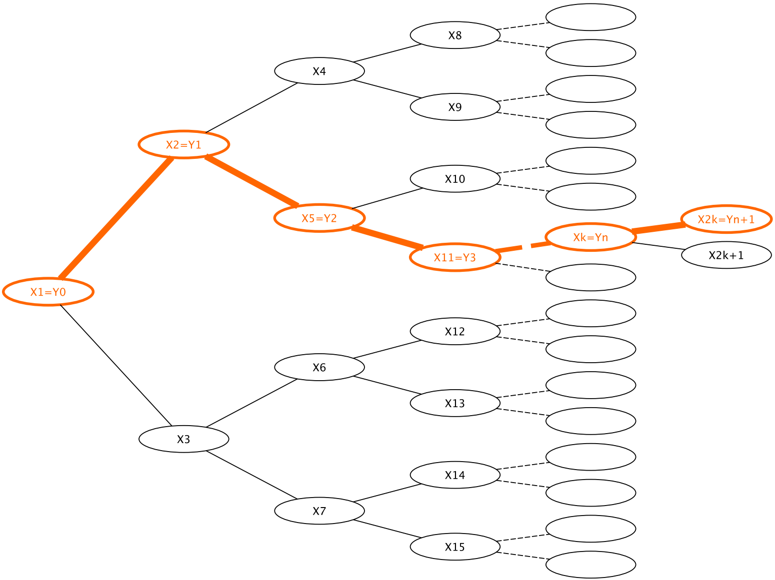
$$\mathbb{E}\left[\prod_{k\in\mathbb{G}_n} f_k(X_{2k},X_{2k+1}) \,\middle|\, \sigma(X_j, j\in\mathbb{T}_n)\right] = \prod_{k\in\mathbb{G}_n} Pf_k(X_k)$$

- asymptotic behavior of (X_k) given by an induced Markov chain

$$\left\{\begin{array}{rcl} Y_0 & = & X_1, \\ Y_{n+1} & = & A_{n+1} + B_{n+1} Y_n \end{array}\right.$$

random lineage (A_n, B_n) iid with distribution
$(a+\epsilon_2, b)\mathbb{1}_{\{\zeta=1\}} + (c+\epsilon_3, d)\mathbb{1}_{\{\zeta=0\}}$, $\zeta \sim$ Bernoulli$(1/2)$

Induced Markov chain



First contribution

[Bercu, dS, Gégout-Petit 2009] Asymmetric model

$$\begin{cases} X_{2k} &= a + bX_k + \epsilon_{2k} \\ X_{2k+1} &= c + dX_k + \epsilon_{2k+1} \end{cases}$$

Assumptions

$\mathcal{F}_n = \sigma\{X_k, k \in \mathbb{T}_n\}$ generation-wise filtration

- moments of order 8 for the noise
- martingale difference sequence
  $\mathbb{E}[\epsilon_{2k+i}|\mathcal{F}_n] = 0$ for all $k \in \mathbb{G}_n$, ϵ_{2k+i} independent of $\epsilon_{2k'+j}$ conditionnally to $\mathcal{F}_n$ for all $k \neq k' \in \mathbb{G}_n$
- $\mathbb{E}[\epsilon_{2k+i}^2|\mathcal{F}_n] = \sigma^2$, $\mathbb{E}[\epsilon_{2k}\epsilon_{2k+1}|\mathcal{F}_n] = \rho$ for all $k \in \mathbb{G}_n$

- convergence rate for the estimators
- martingale approach

Martingale approach

Convergence of martingales in L^2

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

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

- identify a (vector) martingale for the generation-wise filtration
- compute the limit of the quadratic variation $< M >_n \sim |\mathbb{T}_n|$
- apply the theorem of convergence with rate ?

Martingale approach

Convergence of martingales in L^2

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

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

- identify a (vector) martingale for the generation-wise filtration
- compute the limit of the quadratic variation $< M >_n \sim |\mathbb{T}_n|$
- prove the theorem of convergence with rate for martingales on a binary tree

Real data

Escherichia coli data of [Stewart & al. 2005]

- 94 films = 94 genealogies
- 4 to 9 generations of cells in each genealogy
- average growth rate 0.037
- no complete genealogy: cells out of scope, overlapping, ...

Real data

Escherichia coli data of [Stewart & al. 2005]

- 94 films = 94 genealogies
- 4 to 9 generations of cells in each genealogy
- average growth rate 0.037
- no complete genealogy: cells out of scope, overlapping, ...

Our test procedure does not apply to these data

Real data

Escherichia coli data of [Stewart & al. 2005]

- 94 films = 94 genealogies
- 4 to 9 generations of cells in each genealogy
- average growth rate 0.037
- no complete genealogy: cells out of scope, overlapping, ...

Our test procedure does not apply to these data

$\Longrightarrow$ New procedure taking missing data into account

Outline

Galton-Watson model

[Delmas & Marsalle 2010]

- each cell has a type 0 (even – new pole) or 1 (odd – old pole)
- probability $p(j_0, j_1)$ for a cell to have j_0 daughter of type 0 and j_1 daughters of type 1, drawn independently for each cell
- Z_n number of observed cells in generation n Galton-Watson process
- if a cell is not observed, its offspring are not observed either
- inference for partially observed BAR process through the bifurcating Markov chain framework

Galton-Watson model

[Delmas & Marsalle 2010]

- each cell has a type 0 (even – new pole) or 1 (odd – old pole)
- probability $p(j_0, j_1)$ for a cell to have j_0 daughter of type 0 and j_1 daughters of type 1, drawn independently for each cell
- Z_n number of observed cells in generation n Galton-Watson process
- if a cell is not observed, its offspring are not observed either
- inference for partially observed BAR process through the bifurcating Markov chain framework

The number of daughters of each type should also depend on the type of the mother

Two-type Galton-Watson model

- $\delta_k = 1$ if cell k is observed, 0 otherwise
- probability $p^{(i)}(j_0, j_1)$ for a mother cell of type i to have j_0 daughter of type 0 et j_1 daughter of type 1, drawn independently for each cell
- Z_n^i number of cells of type i ingeneration n, (Z_n^0, Z_n^1) two-type Galton-Watson process
- if a cell is not observed, its offspring are not observed either

Extinction

Descendants matrix

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

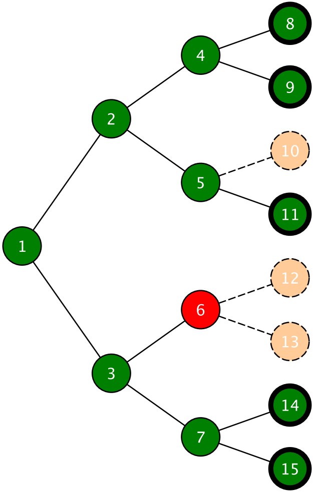
$p_{i0} = p^{(i)}(1,0) + p^{(i)}(1,1)$: mean number of daughters of type 0
$p_{i1} = p^{(i)}(0,1) + p^{(i)}(1,1)$: mean number of daughters of type 1
for a mother of type i

Probability of extinction

π spectral radius of P

- if $\pi \leq 1$, almost sure extinction
- if $\pi > 1$, extinction with probability < 1

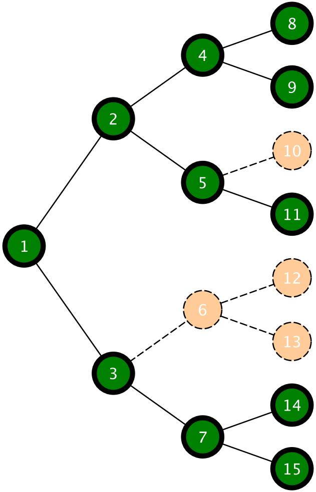
Observed generations



Observed generation n

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

Observed generations



Observed tree up to generation n

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

Partially observed BAR process

$$\begin{cases} X_{2k} = a + b\, X_k + \epsilon_{2k} \\ X_{2k+1} = c + d\, X_k + \epsilon_{2k+1} \end{cases}$$

Assumptions

- independence between (δ_k) and X_1, $(\epsilon_{2k}, \epsilon_{2k+1})$
- noise martingale difference sequence with moments up to order 8

Least squares estimation of $\boldsymbol{\theta} = (a, b, c, d)^t$: minimize

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

Empirical estimators for the moments of the noise

Estimator of $\boldsymbol{\theta}$

Least squares estimator for $\boldsymbol{\theta}$

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

with

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

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

Convergence rate

Theorem

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

Proof: martingale approach

- identify a (vector) martingale for the generation-wise filtration with observations
- compute the limit of the quadratic variation
- theorem on the convergence rate of martingales on a Galton-Watson binary tree

Main martingale

$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} = \boldsymbol{S}_{n-1}^{-1}\boldsymbol{M}_n$, with $(\boldsymbol{M}_n)$ martingale for the generation-wise filtration of the process and observations

$$\boldsymbol{M}_n = \sum_{k\in\mathbb{T}_{n-1}} \begin{pmatrix} \delta_{2k}\epsilon_{2k} \\ \delta_{2k}X_k\epsilon_{2k} \\ \delta_{2k+1}\epsilon_{2k+1} \\ \delta_{2k+1}X_k\epsilon_{2k+1} \end{pmatrix}$$

$(\boldsymbol{M}_n)_{n\geq 1}$ square integrable with quadratic variation $<\boldsymbol{M}>_n = \boldsymbol{\Gamma}_{n-1}$

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

Convergence of the quadratic variation

Laws of large numbers for the observations (δ_k), the noise $(\delta_k \epsilon_k)$ processes

- scalar martingales for various filtrations

Laws of large numbers for the BAR $(\delta_{2k+i} X_k^q)$ processes

- specific form of the autoregression
- assumption $\max\{|b|, |d|\} < 1$

Central limit theorem

Theorem

Conditionally to non extinction

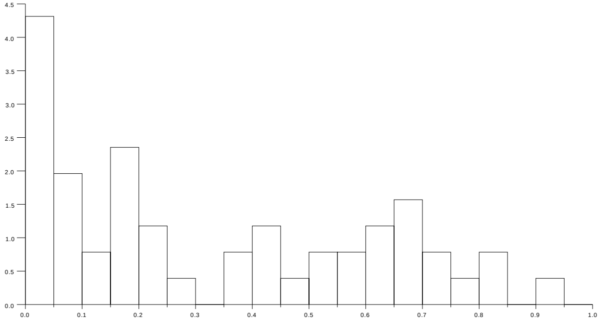
$$\sqrt{|\mathbb{T}^*_{n-1}|}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \boldsymbol{S}^{-1}\boldsymbol{\Gamma}\boldsymbol{S}^{-1})$$

Two main difficulties

- random $|\mathbb{T}^*_{n-1}|$ normalization
- result only valid conditionally to non extinction: on the non extinction set $\overline{\mathcal{E}} = \cap\{|\mathbb{G}^*_n| > 0\}$ endowed with the probability $\mathbb{P}_{\overline{\mathcal{E}}}(\cdot) = \mathbb{P}(\cdot \cap \overline{\mathcal{E}})/\mathbb{P}(\overline{\mathcal{E}})$

Symmetry tests: Escherichia coli data

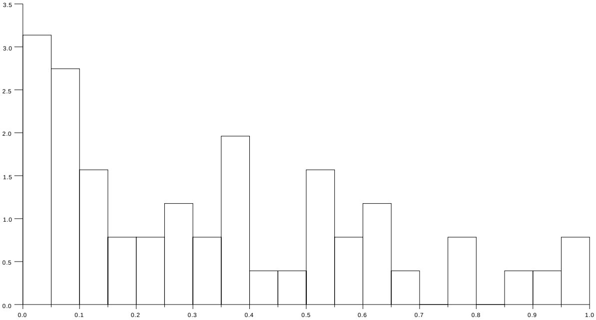
p-values for the 51 genealogies with 8 or 9 generations



Test $(a, b) = (c, d)$

Symmetry tests: Escherichia coli data

p-values for the 51 genealogies with 8 or 9 generations



Test $a/(1-b) = c/(1-d)$

New model

Simulations $\Longrightarrow$ low power of the tests for 8 or 9 generations

Multiple-tree estimation

- use several genealogies (in fixed number) for inference
- genealogies are iid samples of the partially observed BAR process with the same parameters
- new estimator ($\neq$ average of single-tree estimators)
- union of non-extinction sets
- new proofs of convergence with the same ideas
- inference and symmetry test for the Galton Watson process

Multiple-tree estimator

Least squares estimator for $\boldsymbol{\theta}$

$$\widehat{\boldsymbol{\theta}}_n = \Big(\sum_{j=1}^{m} \boldsymbol{S}_{n-1}(j)\Big)^{-1} \sum_{j=1}^{m} \sum_{k\in\mathbb{T}_{n-1}} \begin{pmatrix} \delta_{j,2k} X_{j,2k} \\ \delta_{j,2k} X_{j,k} X_{j,2k} \\ \delta_{j,2k+1} X_{j,2k+1} \\ \delta_{j,2k+1} X_{j,k} X_{j,2k+1} \end{pmatrix}$$

with

$$\boldsymbol{S}_n(j) = \begin{pmatrix} \boldsymbol{S}_n^0(j) & 0 \\ 0 & \boldsymbol{S}_n^1(j) \end{pmatrix}$$

$$\boldsymbol{S}_n^i(j) = \sum_{k\in\mathbb{T}_n} \delta_{j,2k+i} \begin{pmatrix} 1 & X_{j,k} \\ X_{j,k} & X_{j,k}^2 \end{pmatrix}$$

Multiple-tree analysis of E. coli data: BAR

Estimation of $\theta \Longrightarrow$ assumption $\max\{|b|,|d|\} < 1$ holds true

a	0.0203 [0.0197; 0.0210]	c	0.0195 [0.0188; 0.0201]
b	0.4615 [0.4437; 0.4792]	d	0.4782 [0.4605; 0.4959]

Estimation of the moments of the noise

σ^2	$1.81 \cdot 10^{-5}$ $[1.12 \cdot 10^{-5}; 2.50 \cdot 10^{-5}]$
ρ	$0.48 \cdot 10^{-5}$ $[0.44 \cdot 10^{-5}; 0.52 \cdot 10^{-5}]$

Tests

hypothesis $(a,b) = (c,d)$ rejected (p-value $= 10^{-5}$),
hypothesis $a/(1-b) = c/(1-d)$ rejected (p-value $= 2 \cdot 10^{-3}$)

Multiple-tree analysis of E. coli data: Galton-Watson

Estimation of the reproduction laws

$p^{(0)}(0,0)$	0.35579 [0.35574; 0.35583]	$p^{(1)}(0,0)$	0.35611 [0.35606; 0.35616]
$p^{(0)}(1,0)$	0.03621 [0.03620; 0.03622]	$p^{(1)}(1,0)$	0.04707 [0.04706; 0.04708]
$p^{(0)}(0,1)$	0.04740 [0.04739; 0.04741]	$p^{(1)}(0,1)$	0.03755 [0.03754; 0.03756]
$p^{(0)}(1,1)$	0.56060 [0.56055; 0.56065]	$p^{(1)}(1,1)$	0.55928 [0.55923; 0.55933]

Estimation of π: 1.204 [1.191; 1.217]
$\Longrightarrow$ assumption $\pi > 1$ holds true

Tests
hypothesis of equality of the means of the reproduction laws not rejected (p-value $= 0.9$),
assumption of equality between the vectors rejected (p-value $= 2 \cdot 10^{-5}$)

Outline

Random coefficient model

$$\begin{cases} X_{2k} & = & (a + \varepsilon_{2k}) & + & (b + \eta_{2k})\, X_k \\ X_{2k+1} & = & (c + \varepsilon_{2k+1}) & + & (d + \eta_{2k+1})\, X_k \end{cases}$$

Assumptions

- $(\varepsilon_{2k}, \eta_{2k}, \varepsilon_{2k+1}, \eta_{2k+1})$ iid
- moments up to order 32
- missing data modeled by a simple supercritical Galton Watson process

Estimators

- Least squares estimator of $\boldsymbol{\theta}$: same formula
- modified least squares estimators for the moments of the noise: minimize

$$\frac{1}{2}\sum_{\ell=1}^{n-1}\sum_{k\in\mathbb{G}_\ell}(\widehat{\epsilon}_{2k}^2-\mathbb{E}[\epsilon_{2k}^2|\mathcal{F}_\ell^{\mathcal{O}}])^2+(\widehat{\epsilon}_{2k+1}^2-\mathbb{E}[\epsilon_{2k+1}^2|\mathcal{F}_\ell^{\mathcal{O}}])^2$$

$$\frac{1}{2}\sum_{\ell=1}^{n-1}\sum_{k\in\mathbb{G}_\ell}(\widehat{\epsilon}_{2k}\widehat{\epsilon}_{2k+1}-\mathbb{E}[\epsilon_{2k}\epsilon_{2k+1}|\mathcal{F}_\ell^{\mathcal{O}}])^2$$

where $(\mathcal{F}_n^{\mathcal{O}})$ generation-wise filtration with observations and

$$\left\{\begin{array}{lcl}\epsilon_{2k} & = & \delta_{2k}(\varepsilon_{2k}+\eta_{2k}X_k),\\ \epsilon_{2k+1} & = & \delta_{2k+1}(\varepsilon_{2k+1}+\eta_{2k+1}X_k),\end{array}\right. \quad \left\{\begin{array}{lcl}\widehat{\epsilon}_{2k} & = & \delta_{2k}(X_{2k}-\widehat{a}_n-\widehat{b}_nX_k),\\ \widehat{\epsilon}_{2k+1} & = & \delta_{2k}(X_{2k+1}-\widehat{c}_n-\widehat{d}_nX_k).\end{array}\right.$$

Convergence

Convergence rate

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

Central limit theorem

Conditionally to non extinction

$$\sqrt{|\mathbb{T}_{n-1}^*|}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \boldsymbol{S}^{-1}\boldsymbol{\Gamma}\boldsymbol{S}^{-1})$$

- identify a (vector) martingale for the generation-wise filtration with observations
- compute the limit of the quadratic variation
- theorem on the convergence rate of martingales on a Galton-Watson binary tree

Main martingale

$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} = \boldsymbol{S}_{n-1}^{-1}\boldsymbol{M}_n$, with $(\boldsymbol{M}_n)$ martingale for the generation -wise filtration with observations

$$\boldsymbol{M}_n = \sum_{k\in\mathbb{T}_{n-1}} \begin{pmatrix} \delta_{2k}\epsilon_{2k} \\ \delta_{2k}X_k\epsilon_{2k} \\ \delta_{2k+1}\epsilon_{2k+1} \\ \delta_{2k+1}X_k\epsilon_{2k+1} \end{pmatrix}$$

$$\begin{cases} \epsilon_{2k} & = \quad \delta_{2k}(\varepsilon_{2k} + \eta_{2k}X_k), \\ \epsilon_{2k+1} & = \quad \delta_{2k+1}(\varepsilon_{2k+1} + \eta_{2k+1}X_k), \end{cases}$$

quadratic variation $< \boldsymbol{M} >_n = \boldsymbol{\Gamma}_{n-1}$, 4×4 matrix with terms of the form $\sum_{k\in\mathbb{T}_{n-1}} \delta_{2k+i}X_k^q$, $0 \leq q \leq 4$

Convergence of the quadratic variation

We do not want to suppose

$$\max\{|b+\eta_2|, |d+\eta_3|\} < 1$$

$\Longrightarrow$ no majoration to make asymmetry vanish
impossible to use the martingale approach martingale directly

Convergence of the quadratic variation

We do not want to suppose

$$\max\{|b+\eta_2|, |d+\eta_3|\} < 1$$

$\Longrightarrow$ no majoration to make asymmetry vanish
impossible to use the martingale approach martingale directly

$\Longrightarrow$ laws of large numbers by bifurcating Markov chain approach

Bifurcating Markov chain on a Galton-Watson tree

Bifurcating Markov chain on $\mathbb{R} \cup \partial$

$$X_k^* = X_k \mathbb{1}_{\{\delta_k = 1\}} + \partial \mathbb{1}_{\{\delta_k = 0\}}$$

bifurcating Markov kernel on $(\mathbb{R} \cup \partial)$ $Pf(\partial) = f(\partial, \partial, \partial)$ and

$$\begin{aligned} Pf(x) &= p(1,1)\mathbb{E}\left[f\left(x, (b+\eta_2)x + a + \varepsilon_2, (d+\eta_3)x + c + \varepsilon_3\right)\right] \\ &\quad + p(1,0)\mathbb{E}\left[f\left(x, (b+\eta_2)x + a + \varepsilon_2, \partial\right)\right] \\ &\quad + p(0,1)\mathbb{E}\left[f\left(x, \partial, (d+\eta_3)x + c + \varepsilon_3\right)\right] \\ &\quad + p(0,0)f(x, \partial, \partial) \end{aligned}$$

Sub-Markovian kernels on $\mathbb{R}$

$$P_0(x, A) = (p(1,1) + p(1,0))\mathbb{E}\left[\mathbb{1}_A\big((a+\varepsilon_2) + (b+\eta_2)x\big)\right]$$

Bifurcating Markov chain on a Galton-Watson tree

Bifurcating Markov chain on $\mathbb{R} \cup \partial$

$$X_k^* = X_k \mathbb{1}_{\{\delta_k = 1\}} + \partial \mathbb{1}_{\{\delta_k = 0\}}$$

bifurcating Markov kernel on $(\mathbb{R} \cup \partial)$ $Pf(\partial) = f(\partial, \partial, \partial)$ and

$$\begin{aligned} Pf(x) &= p(1,1)\mathbb{E}\left[f\left(x, (b+\eta_2)x + a + \varepsilon_2, (d+\eta_3)x + c + \varepsilon_3\right)\right] \\ &\quad + p(1,0)\mathbb{E}\left[f\left(x, (b+\eta_2)x + a + \varepsilon_2, \partial\right)\right] \\ &\quad + p(0,1)\mathbb{E}\left[f\left(x, \partial, (d+\eta_3)x + c + \varepsilon_3\right)\right] \\ &\quad + p(0,0) f(x, \partial, \partial) \end{aligned}$$

Sub-Markovian kernels on $\mathbb{R}$

$$P_1(x, A) = (p(1,1) + p(0,1))\mathbb{E}\left[\mathbb{1}_A\big((c + \varepsilon_3) + (d + \eta_3)x\big)\right]$$

Induced Markov chain

(A_n, B_n) iid $\sim (a+\epsilon_2, b+\eta_2)\mathbb{1}_{\{\zeta=1\}} + (c+\epsilon_3, d+\eta_3)\mathbb{1}_{\{\zeta=0\}}$,
$\zeta \sim$ Bernoulli$\big((p(1,1)+p(1,0))/\pi\big)$ where π mean of the reproduction law

$$\begin{cases} Y_0 &= X_1, \\ Y_{n+1} &= A_{n+1} + B_{n+1} Y_n \end{cases}$$

- Markov kernel $Q = (P_0 + P_1)/\pi$
- Many to one formula

$$\frac{1}{\pi^n} \sum_{k \in \mathbb{G}_n} \mathbb{E}[f(X_k)\mathbb{1}_{\{k \in \mathbb{T}_n^*\}}] = \mathbb{E}[f(Y_n)]$$

- Law of large numbers: ν distribution of X_1

$$\Big\| \frac{1}{\pi^n} \sum_{k \in \mathbb{G}_n^*} f(X_k) \Big\|_{L^2}^2 = \frac{\nu Q^n f^2}{\pi^n} + \frac{2}{\pi^2} \sum_{\ell=0}^{n-1} \frac{1}{\pi^\ell} \nu Q^\ell P(Q^{n-\ell-1} f \otimes Q^{n-\ell-1} f)$$

Ergodicity of the induced chain

- invariant distribution $\mu \sim \sum B_1 \cdots B_{n-1} A_n$
- geometric ergodicity for polynomials up to degree q if

$$\mathbb{E}[|B_1|^q] = \frac{p(1,0)+p(1,1)}{\pi}\mathbb{E}[|b+\eta_2|^q] + \frac{p(0,1)+p(1,1)}{\pi}\mathbb{E}[|d+\eta_3|^q] < 1$$

  replace assumption $\max\{|b|,|d|\} < 1$
- law of large numbers for X_k^q requires moments of order $4q$
- convergence of the quadratic variation
- rate of convergence of the estimators via martingale approach

Outline

Bifrucating Markov chain vs martingale approach

	Martingale	Markov chain
noise	martingale difference sequence moments of order q	iid moments of order $4q$
b and d	max < 1	weighted mean < 1
observations	two-type Galton-Watson process	simple Galton-Watson process two-type ?

References

[Cowan & Staudte 1986] COWAN & STAUDTE The bifurcating autoregressive model in cell lineage studies. *Biometrics* (1986).

[Guyon 2007] GUYON Limit theorems for bifurcating Markov chains. Application to the detection of cellular aging. *Ann. Appl. Probab.* (2007)

[Delmas & Marsalle 2010] DELMAS & MARSALLE Detection of cellular aging in a Galton-Watson process. *Stoch. Process. and Appl.* (2010)

[Stewart & al. 2005] STEWART, MADDEN, PAUL, TADDEI Aging and death in an organism that reproduces by morphologically symmetric division. *PLoS Biol.* (2005)

[Bercu, dS, Gégout-Petit 2009] BERCU, DE SAPORTA, GÉGOUT-PETIT Asymptotic analysis for bifurcating autoregressive processes via a martingale approach. *Electron. J. Probab.* (2009)

[dS, Gégout-Petit, Marsalle 2011] DE SAPORTA, GÉGOUT-PETIT, MARSALLE Parameters estimation for asymmetric bifurcating autoregressive processes with missing data. *Electron. J. Statist.* (2011)

[dS, Gégout-Petit, Marsalle 2012] DE SAPORTA, GÉGOUT-PETIT, MARSALLE Symmetry tests for bifurcating autoregressive processes with missing data. *Statistics & Probability Letters* (2012)

random coefficents DE SAPORTA, GÉGOUT-PETIT, MARSALLE Random coefficients bifurcating autoregressive processes. Arxiv 1205.4840

multiple trees DE SAPORTA, GÉGOUT-PETIT, MARSALLE Statistical study of asymmetry in cell lineage data. Arxiv 1205.3658