Asymmetry tests for Bifurcating Auto-Regressive Processes with missing data

Benoîte de Saporta[a], Anne Gégout-Petit[b], Laurence Marsalle[c]

[a] *Univ. Bordeaux, GREThA CNRS UMR 5113, IMB CNRS UMR 5251, INRIA Bordeaux Sud Ouest team CQFD, 351 cours de la Libération, 33405 Talence Cedex, France, saporta@math.u-bordeaux1.fr, tel: +33 5 24 57 41 69, fax: +33 5 24 57 40 24*

[b] *Univ. Bordeaux, IMB CNRS UMR 5251, INRIA Bordeaux Sud Ouest team CQFD, 351 cours de la Libération, 33405 Talence Cedex, France, anne.petit@u-bordeaux2.fr*

[c] *Univ. Lille 1, Laboratoire Paul Painlevé, CNRS UMR 8524, 59655 Villeneuve d'Ascq Cedex, France, laurence.marsalle@univ-lille1.fr*

Abstract

We present symmetry tests for bifurcating autoregressive processes (BAR) when some data are missing. BAR processes typically model cell division data. Each cell can be of one of two types *odd* or *even*. The goal of this paper is to study the possible asymmetry between odd and even cells in a single observed lineage. We first derive asymmetry tests for the lineage itself, modeled by a two-type Galton-Watson process, and then derive tests for the observed BAR process. We present applications on simulated and real data.

Key words: Bifurcating autoregressive processes, Cell division data, Missing data, Two-type Galton-Watson model, Wald's test.

2010 MSC: 62M07, 62F05, 60J80, 62P10

1. Introduction

Bifurcating autoregressive processes (BAR) were first introduced by Cowan and Staudte (1986). They generalize autoregressive processes when data are

structured as a binary tree, see also (Hwang and Basawa, 2009, 2011) for processes indexed by general trees. Typically, BAR processes are involved in statistical studies of cell lineages. Cell lineage data consist of observations of some quantitative characteristic of the cells over several generations descended from an initial cell. One may need to distinguish the two offspring of a given cell according to some biological property, leading to the notion of *type*. The initial cell is labelled 1, and the two offspring of cell k are labelled $2k$ and $2k+1$, where $2k$ is of type *even*, and $2k+1$ is of type *odd*. If X_k denotes the quantitative characteristic of cell k, the first-order asymmetric BAR process is given by

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

for all $k \geq 1$. The noise sequence $(\varepsilon_{2k}, \varepsilon_{2k+1})$ represents environmental effects, while a, b, c, d are unknown real parameters related to the inherited effects. Various estimators are studied in the literature for a, b, c, d, see (Guyon, 2007; Delmas and Marsalle, 2010; Bercu et al., 2009; de Saporta et al., 2011). Here the genealogy is modeled by a two-type Galton Watson process (GW), allowing the reproduction laws to depend on both the mother's and daughter's types. The aim of this paper is to propose asymmetry tests for both the GW process defining the genealogy of the cells, and the BAR process with missing data. In particular, we propose original taylor-made estimators for the reproduction probabilities of the GW process, give their asymptotic behavior and derive Wald test for the equality of the means of the reproduction laws for even and odd mother cells. We also investigate asymmetry of the parameters a, b, c, d. A detailed study on simulated data, as well as a new investigation

of the *Escherichia coli* data of Stewart et al. (2005) are provided.

2. Notation

For all $n \geq 1$, denote the n-th generation by $\mathbb{G}_n = \{k \in \mathbb{N} : 2^n \leq k \leq 2^{n+1} - 1\}$, and the sub-tree of all cells up to the n-th generation by $\mathbb{T}_n = \bigcup_{\ell=0}^{n} \mathbb{G}_\ell$. The cardinality $|\mathbb{G}_n|$ of $\mathbb{G}_n$ is 2^n, while that of $\mathbb{T}_n$ is $|\mathbb{T}_n| = 2^{n+1} - 1$. We encode the presence or absence of cells by the process (δ_k): if cell k is observed, $\delta_k = 1$, if cell k is missing, $\delta_k = 0$. We define the sets of observed cells as $\mathbb{G}_n^* = \{k \in \mathbb{G}_n : \delta_k = 1\}$ and $\mathbb{T}_n^* = \{k \in \mathbb{T}_n : \delta_k = 1\}$. Finally, let $\mathcal{E}$ be the event corresponding to the case when there are no cell left to observe: $\mathcal{E} = \bigcup_{n\geq 1}\{|\mathbb{G}_n^*| = 0\}$ and $\overline{\mathcal{E}}$ its complementary set. For $n \geq 1$, we define the number of observed cells in the n-th generation distinguishing their type: $Z_n^0 = |\mathbb{G}_n^* \cap 2\mathbb{N}|$, $Z_n^1 = |\mathbb{G}_n^* \cap (2\mathbb{N}+1)|$, and we set, for all $n \geq 1$, $\mathbf{Z}_n = (Z_n^0, Z_n^1)$. Note that for $i \in \{0, 1\}$ and $n \geq 1$ one has $Z_n^i = \sum_{k\in\mathbb{G}_{n-1}} \delta_{2k+i}$.

3. Asymmetry in the lineage

We now describe the mechanism generating the observation process (δ_k) and recall some classical assumptions taken from (Harris, 1963). We define the cells genealogy by a two-type GW process $(\mathbf{Z}_n)$. All cells reproduce independently and with a reproduction law depending only on their type. For a mother cell of type $i \in \{0, 1\}$, we denote by $p^{(i)}(j_0, j_1)$ the probability that it has j_0 daughter of type 0 and j_1 daughter of type 1. For the cell division process we are interested in, one clearly has $p^{(i)}(0, 0) + p^{(i)}(1, 0) + p^{(i)}(0, 1) + p^{(i)}(1, 1) = 1$. The reproduction laws have moments of all order, and we can thus define the descendants matrix as the 2×2 matrix $\mathbf{P} = (p_{ij})_{0\leq i,j\leq 1}$, where

$p_{i0} = p^{(i)}(1,0) + p^{(i)}(1,1)$ and $p_{i1} = p^{(i)}(0,1) + p^{(i)}(1,1)$, for $i \in \{0,1\}$ i.e. p_{ij} is the expected number of descendants of type j of a cell of type i. We make the following main assumption.

(AO) All entries of the matrix $\mathbf{P}$ are positive and its dominant eigenvalue π satisfies $\pi > 1$.

In this case, there exist component-wise positive left eigenvectors for π. Let $\mathbf{z} = (z^0, z^1)$ be the one satisfying $z^0 + z^1 = 1$. Assumption (**AO**) means that $(\mathbf{Z}_n)$ is super-critical and ensures that extinction is not almost sure: $\mathbb{P}(\mathcal{E}) < 1$. On $\overline{\mathcal{E}}$, $|\mathbb{T}_n^*|^{-1} \sum_{l=1}^n Z_l^i$ converges to z^i, meaning that z^i is the asymptotic proportion of cells of type i. Our context is very specific because the information given by (δ_k) is more precise than the one given by $(\mathbf{Z}_n)$ used the literature, see (Guttorp, 1991). Empiric estimators of the reproduction probabilities using data up to the n-th generation are, for i, j_0, j_1 in $\{0,1\}$

$$\widehat{p}_n^{(i)}(j_0, j_1) = \frac{\sum_{k \in \mathbb{T}_{n-2}} \delta_{2k+i} \phi_{j_0}(\delta_{2(2k+i)}) \phi_{j_1}(\delta_{2(2k+i)+1})}{\sum_{k \in \mathbb{T}_{n-2}} \delta_{2k+i}},$$

where $\phi_0(x) = 1 - x$, $\phi_1(x) = x$, if the denominators are non zero (zero otherwise). The strong consistency is readily obtained on the non-extinction set $\overline{\mathcal{E}}$ by martingales methods.

Lemma 3.1. *Under* (**AO**) *and for all* i, j_0 *and* j_1 *in* $\{0,1\}$, *one has* $\lim_{n\to\infty} \mathbb{1}_{\{|\mathbb{G}_n^*|>0\}} \widehat{p}_n^{(i)}(j_0, j_1) = p^{(i)}(j_0, j_1) \mathbb{1}_{\overline{\mathcal{E}}}$ *a.s.*

Set $\mathbf{p}^{(i)} = (p^{(i)}(0,0), p^{(i)}(1,0), p^{(i)}(0,1), p^{(i)}(1,1))^t$ the vector of the 4 reproduction probabilities for a mother of type i, $\mathbf{p} = ((\mathbf{p}^{(0)})^t, (\mathbf{p}^{(1)})^t)^t$ the vector of all 8 reproductions probabilities and $\widehat{\mathbf{p}}_n = (\widehat{p}_n^{(0)}(0,0), \ldots, \widehat{p}_n^{(1)}(1,1))^t$ its estimator. As $\mathbb{P}(\overline{\mathcal{E}}) \neq 0$, we define the conditional probability $\mathbb{P}_{\overline{\mathcal{E}}}$ by $\mathbb{P}_{\overline{\mathcal{E}}}(A) = \mathbb{P}(A \cap \overline{\mathcal{E}})/\mathbb{P}(\overline{\mathcal{E}})$ for all event A.

Theorem 3.2. *Under assumption* (**AO**)*, we have the convergence* $\sqrt{|\mathbb{T}^*_{n-1}|}(\widehat{\mathbf{p}}_n - \mathbf{p}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{V})$ *on* $(\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$*, with* $\mathbf{V} = \begin{pmatrix} \mathbf{V}^0/z^0 & 0 \\ 0 & \mathbf{V}^1/z^1 \end{pmatrix}$ *and for all* i *in* $\{0,1\}$*,* $\mathbf{V}^i = \mathbf{W}^i - \mathbf{p}^{(i)}(\mathbf{p}^{(i)})^t$*,* $\mathbf{W}^i$ *is a* 4×4 *matrix with the entries of* $\mathbf{p}^{(i)}$ *on the diagonal and* 0 *elsewhere.*

Proof : For all $n \geq 2$, and $q \geq 1$, set $\mathbf{M}_q^{(n)} = \big((\mathbf{M}_q^{0(n)})^t, (\mathbf{M}_q^{1(n)})^t\big)^t$, with

$$\mathbf{M}_q^{i(n)} = \frac{1}{\sqrt{|\mathbb{T}^*_{n-1}|}} \sum_{k=1}^{q} \begin{pmatrix} \delta_{2k+i}\big((1-\delta_{2(2k+i)})(1-\delta_{2(2k+i)+1}) - p^{(i)}(0,0)\big) \\ \delta_{2k+i}\big(\delta_{2(2k+i)}(1-\delta_{2(2k+i)+1}) - p^{(i)}(1,0)\big) \\ \delta_{2k+i}\big((1-\delta_{2(2k+i)})\delta_{2(2k+i)+1} - p^{(i)}(0,1)\big) \\ \delta_{2k+i}\big(\delta_{2(2k+i)}\delta_{2(2k+i)+1} - p^{(i)}(1,1)\big) \end{pmatrix}.$$

Let $(\mathcal{G}_q)$ be the filtration of cousin cells: $\mathcal{G}_0 = \sigma\{\delta_1, \delta_2, \delta_3\}$ and for all $q \geq 1$, $\mathcal{G}_q = \mathcal{G}_{q-1} \vee \sigma\{\delta_{4q}, \delta_{4q+1}, \delta_{4q+2}, \delta_{4q+3}\}$. For all $n \geq 2$, $(\mathbf{M}_q^{(n)})$ is a $(\mathcal{G}_q)$-martingale . We apply Theorem 3.II.10 of (Duflo, 1997) with the stopping times $\nu_n = |\mathbb{T}_{n-2}|$. The $\mathbb{P}_{\overline{\mathcal{E}}}$ a.s. limit of the increasing process is

$$< \mathbf{M}^{(n)} >_{\nu_n} = \frac{1}{|\mathbb{T}^*_{n-1}|} \begin{pmatrix} \sum_{k \in \mathbb{T}_{n-2}} \delta_{2k} \mathbf{V}^0 & 0 \\ 0 & \sum_{k \in \mathbb{T}_{n-2}} \delta_{2k+1} \mathbf{V}^1 \end{pmatrix} \xrightarrow[n\to\infty]{} \mathbf{G} = \begin{pmatrix} z^0 \mathbf{V}^0 & 0 \\ 0 & z^1 \mathbf{V}^1 \end{pmatrix}.$$

In addition, the Lindeberg condition holds as the δ_k have finite moments of all order. Thus, we obtain the convergence $M^{(n)}_{|\mathbb{T}_{n-2}|} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{G})$ on $(\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$. On the other hand, $\mathbf{\Delta}_{n-1}^{-1} |\mathbb{T}^*_{n-1}| M^{(n)}_{|\mathbb{T}_{n-2}|} = \sqrt{|\mathbb{T}^*_{n-1}|}(\widehat{\mathbf{p}}_n - \mathbf{p})$, with

$$\mathbf{\Delta}_n = \begin{pmatrix} \sum_{\ell=1}^n Z_\ell^0 \mathbf{I}_4 & 0 \\ 0 & \sum_{\ell=1}^n Z_\ell^1 \mathbf{I}_4 \end{pmatrix},$$

and $\mathbf{I}_4$ is the identity matrix of size 4. As $|\mathbb{T}^*_n|^{-1} \sum_{\ell=1}^n Z_\ell^i$ converges a.s. to z^i on $(\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$, we have the asymptotic normality, using Slutsky's lemma. □

We now derive Wald's test for the asymmetry of the means of the reproduction laws. Set $m = \big(p^{(0)}(1,0) + p^{(0)}(0,1) + 2p^{(0)}(1,1)\big) - \big(p^{(1)}(1,0) + p^{(1)}(0,1) + 2p^{(1)}(1,1)\big)$ the difference of the means of the types 0 and 1 reproduction laws and $\widehat{m}_n$ its empirical estimator. Set $\mathbf{H_0^{GW}}$: $m = 0$ the symmetry hypothesis and $\mathbf{H_1^{GW}}$: $m \neq 0$. Let $(Y_n^{GW})^2$ be the test statistic defined by $Y_n^{GW} = |\mathbb{T}_{n-1}^*|^{1/2}(\widehat{\Delta}_n^{GW})^{-1/2}\widehat{m}_n$, where $\widehat{\Delta}_n^{GW} = \mathbf{dgw}^t\widehat{\mathbf{V}}_n\mathbf{dgw}$, $\mathbf{dgw} = (0,1,1,2,0-1,-1,-2)^t$, and $\widehat{\mathbf{V}}_n$ is the empirical version of $\mathbf{V}$, where z^i is replaced by $|\mathbb{T}_n^*|^{-1}\sum_{l=1}^n Z_l^i$ and the $p^{(i)}(j_0,j_1)$ are replaced by $\widehat{p}_n^{(i)}(j_0,j_1)$. Thanks to Lemma 3.1, $\widehat{\mathbf{V}}_n$ converges a.s. to $\mathbf{V}$ and the test statistic has the following asymptotic properties.

Theorem 3.3. *Under assumption* (**AO**) *and the null hypothesis* $\mathbf{H_0^{GW}}$*, one has* $(Y_n^{GW})^2 \xrightarrow{\mathcal{L}} \chi^2(1)$ *on* $(\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$*; and under the alternative hypothesis* $\mathbf{H_1^{GW}}$*, one has* $\lim_{n\to\infty}(Y_n^{GW})^2 = +\infty$ *a.s. on* $(\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$.

Proof : Let g be the function defined from $\mathbb{R}^8$ onto $\mathbb{R}$ by $g(x_1,\ldots,x_8) = \big(x_2+x_3+2x_4\big) - \big(x_6+x_7+2x_8\big)$, so that $\widehat{m}_n - m = g(\widehat{\mathbf{p}}_n) - g(\mathbf{p})$, and $\mathbf{dgw}$ is the gradient of g. Theorem 3.2 yields $\sqrt{|\mathbb{T}_{n-1}^*|}\big(g(\widehat{\mathbf{p}}_n) - g(\mathbf{p})\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Delta^{GW})$ on $(\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$, with $\Delta^{GW} = \mathbf{dgw}^t\mathbf{V}\mathbf{dgw}$. Under $\mathbf{H_0^{GW}}$, $g(\mathbf{p}) = m = 0$, so that $|\mathbb{T}_{n-1}^*|(\Delta^{GW})^{-1}g(\widehat{\mathbf{p}}_n)^2 \xrightarrow{\mathcal{L}} \chi^2(1)$ on $(\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$. One then uses Slutsky's lemma to replace Δ^{GW} by $\widehat{\Delta}_n^{GW}$. Under $\mathbf{H_1^{GW}}$, since $Y_n^{GW} = |\mathbb{T}_{n-1}^*|^{1/2}(\widehat{\Delta}_n^{GW})^{-1/2}\widehat{m}_n$ and $\widehat{m}_n$ converges to $m \neq 0$, Y_n^{GW} tends to infinity a.s. on $(\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$ □

4. Asymmetry in cells characteristic

We turn to the asymmetry of the BAR model with missing data. Assume that $\mathbb{E}[X_1^8] < \infty$ and $0 < \max(|b|,|d|) < 1$. Denote by $\mathbb{F} = (\mathcal{F}_n)$ the natural filtration associated with the BAR process: $\mathcal{F}_n = \sigma\{X_k, k \in \mathbb{T}_n\}$.

(AN.1) One has $\sup_{n\geq 0}\sup_{k\in\mathbb{G}_{n+1}}\mathbb{E}[\varepsilon_k^8|\mathcal{F}_n]<\infty$ a.s., $\forall n\geq 0$ and $k\in\mathbb{G}_{n+1}$,
$\mathbb{E}[\varepsilon_k|\mathcal{F}_n]=0$, $\mathbb{E}[\varepsilon_k^2|\mathcal{F}_n]=\sigma^2$ a.s., $\forall k\in\mathbb{G}_n$, $\mathbb{E}[\varepsilon_{2k}\varepsilon_{2k+1}|\mathcal{F}_n]=\rho$ a.s.

(AN.2) $\forall n\geq 0$, $\{(\varepsilon_{2k},\varepsilon_{2k+1})\}_{k\in\mathbb{G}_n}$ are conditionally independent given $\mathcal{F}_n$.

(AI) The sequence (δ_k) is independent from the sequences (X_k) and (ε_k).

The least-squares estimator of $\theta=(a,b,c,d)^t$ is given for all $n\geq 1$ by $\widehat{\theta}_n=(\widehat{a}_n,\widehat{b}_n,\widehat{c}_n,\widehat{d}_n)^t$ with

$$\widehat{\theta}_n=\mathbf{\Sigma}_{n-1}^{-1}\sum_{k\in\mathbb{T}_{n-1}}\begin{pmatrix}\delta_{2k}X_{2k}\\ \delta_{2k}X_kX_{2k}\\ \delta_{2k+1}X_{2k+1}\\ \delta_{2k+1}X_kX_{2k+1}\end{pmatrix},\ \mathbf{\Sigma}_n=\begin{pmatrix}\mathbf{S}_n^0 & 0\\ 0 & \mathbf{S}_n^1\end{pmatrix},$$

$$\mathbf{S}_n^i=\sum_{k\in\mathbb{T}_n}\delta_{2k+i}\begin{pmatrix}1 & X_k\\ X_k & X_k^2\end{pmatrix},\ \mathbf{S}_n^{0,1}=\sum_{k\in\mathbb{T}_n}\delta_{2k}\delta_{2k+1}\begin{pmatrix}1 & X_k\\ X_k & X_k^2\end{pmatrix}.$$

Denote also $\mathbf{L}^0$, $\mathbf{L}^1$, $\mathbf{L}^{0,1}$ their a.s. limits: $\lim_{n\to\infty}\mathbb{1}_{\{|\mathbb{G}_n^*|>0\}}\mathbf{S}_n^i/|\mathbb{T}_n^*|=\mathbb{1}_{\overline{\mathcal{E}}}\mathbf{L}^i$, $\lim_{n\to\infty}\mathbb{1}_{\{|\mathbb{G}_n^*|>0\}}\mathbf{S}_n^{0,1}/|\mathbb{T}_n^*|=\mathbb{1}_{\overline{\mathcal{E}}}\mathbf{L}^{0,1}$, see Proposition 4.2 of (de Saporta et al., 2011). We now recall Theorems 3.2 and 3.4 of (de Saporta et al., 2011).

Theorem 4.1. *Under* (**AN.1-2**), (**AO**) *and* (**AI**), *the estimator* $\widehat{\theta}_n$ *is strongly consistent* $\lim_{n\to\infty}\mathbb{1}_{\{|\mathbb{G}_n^*|>0\}}\widehat{\theta}_n=\theta\mathbb{1}_{\overline{\mathcal{E}}}$ *a.s. In addition, we have the asymptotic normality* $\sqrt{|\mathbb{T}_{n-1}^*|}(\widehat{\theta}_n-\theta)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\mathbf{\Sigma}^{-1}\mathbf{\Gamma}\mathbf{\Sigma}^{-1})$ *on* $(\overline{\mathcal{E}},\mathbb{P}_{\overline{\mathcal{E}}})$, *where*

$$\mathbf{\Sigma}=\begin{pmatrix}\mathbf{L}^0 & 0\\ 0 & \mathbf{L}^1\end{pmatrix},\quad \textit{and}\quad \mathbf{\Gamma}=\begin{pmatrix}\sigma^2\mathbf{L}^0 & \rho\mathbf{L}^{0,1}\\ \rho\mathbf{L}^{0,1} & \sigma^2\mathbf{L}^1\end{pmatrix}.$$

We now propose two different asymmetry tests. The first one compares the couples (a,b) and (c,d). Set $\mathbf{H_0^c}$: $(a,b)=(c,d)$ the symmetry hypothesis

and $\mathbf{H_1^c}$: $(a,b) \neq (c,d)$. Let $(\mathbf{Y}_n^c)^t \mathbf{Y}_n^c$ be the test statistic defined by $\mathbf{Y}_n^c = |\mathbb{T}_{n-1}^*|^{1/2}(\widehat{\mathbf{\Delta}}_n^c)^{-1/2}(\widehat{a}_n - \widehat{c}_n, \widehat{b}_n - \widehat{d}_n)^t$, with $\widehat{\mathbf{\Delta}}_n^c = |\mathbb{T}_{n-1}^*|^2 \mathbf{dgc}^t \mathbf{\Sigma}_{n-1}^{-1} \widehat{\mathbf{\Gamma}}_{n-1} \mathbf{\Sigma}_{n-1}^{-1} \mathbf{dgc}$,

$$\mathbf{dgc} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}^t, \qquad \widehat{\mathbf{\Gamma}}_n = \frac{1}{|\mathbb{T}_n^*|} \begin{pmatrix} \widehat{\sigma}_{n+1}^2 \mathbf{S}_n^0 & \widehat{\rho}_{n+1} \mathbf{S}_n^{0,1} \\ \widehat{\rho}_{n+1} \mathbf{S}_n^{0,1} & \widehat{\sigma}_{n+1}^2 \mathbf{S}_n^1 \end{pmatrix},$$

$\widehat{\sigma}_n^2 = |\mathbb{T}_n^*|^{-1} \sum_{k \in \mathbb{T}_{n-1}^*} (\widehat{\varepsilon}_{2k}^2 + \widehat{\varepsilon}_{2k+1}^2)$, $\widehat{\rho}_n = |\mathbb{T}_{n-1}^{*01}|^{-1} \sum_{k \in \mathbb{T}_{n-1}} \widehat{\varepsilon}_{2k} \widehat{\varepsilon}_{2k+1}$, $\mathbb{T}_n^{*01} = \{k \in \mathbb{T}_n : \delta_{2k}\delta_{2k+1} = 1\}$ and for all $k \in \mathbb{G}_n$, $\widehat{\varepsilon}_{2k} = \delta_{2k}(X_{2k} - \widehat{a}_n - \widehat{b}_n X_k)$, $\widehat{\varepsilon}_{2k+1} = \delta_{2k+1}(X_{2k+1} - \widehat{c}_n - \widehat{d}_n X_k)$.

Theorem 4.2. *Under assumptions* (**AN.1-2**), (**AO**), (**AI**) *and the null hypothesis* $\mathbf{H_0^c}$, *one has* $(\mathbf{Y}_n^c)^t \mathbf{Y}_n^c \xrightarrow{\mathcal{L}} \chi^2(2)$ *on* $(\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$; *and under the alternative hypothesis* $\mathbf{H_1^c}$, *one has* $\lim_{n\to\infty} \|\mathbf{Y}_n^c\|^2 = +\infty$ *a.s. on* $(\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$.

Proof : We mimic the proof of Theorem 3.3 with g the function defined from $\mathbb{R}^4$ onto $\mathbb{R}^2$ by $g(x_1, x_2, x_3, x_4) = \big(x_1 - x_3, x_2 - x_4\big)^t$. □

Our second test compares the fixed points $a/(1-b)$ and $c/(1-d)$, which are the asymptotic means of X_{2k} and X_{2k+1} respectively. Set $\mathbf{H_0^f}$: $a/(1-b) = c/(1-d)$ the symmetry hypothesis and $\mathbf{H_1^f}$: $a/(1-b) \neq c/(1-d)$. Let $(Y_n^f)^2$ be the test statistic defined by $Y_n^f = |\mathbb{T}_{n-1}^*|^{1/2}(\widehat{\mathbf{\Delta}}_n^f)^{-1/2}\big(\widehat{a}_n/(1-\widehat{b}_n) - \widehat{c}_n/(1-\widehat{d}_n)\big)$, where $\widehat{\mathbf{\Delta}}_n^f = |\mathbb{T}_{n-1}^*|^2 \mathbf{dgf}^t \mathbf{\Sigma}_{n-1}^{-1} \widehat{\mathbf{\Gamma}}_{n-1} \mathbf{\Sigma}_{n-1}^{-1} \mathbf{dgf}$, $\mathbf{dgf} = \big(1/(1-b), a/(1-b)^2, -1/(1-d), -c/(1-d)^2\big)^t$.

Theorem 4.3. *Under assumptions* (**AN.1-2**), (**AO**), (**AI**) *and the null hypothesis* $\mathbf{H_0^f}$, *one has* $(Y_n^f)^2 \xrightarrow{\mathcal{L}} \chi^2(1)$ *on* $(\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$; *and under the alternative hypothesis* $\mathbf{H_1^f}$, *one has* $\lim_{n\to\infty} (Y_n^f)^2 = +\infty$ *a.s. on* $(\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}})$.

Proof : We mimic again the proof of Theorem 3.3 with g the function defined from $\mathbb{R}^4$ onto $\mathbb{R}$ by $g(x_1, x_2, x_3, x_4) = \big(x_1/(1-x_2) - x_3/(1-x_4)\big)^t$. □

5. Application to simulated data

We now study the behavior of our tests on simulated data. For each test, we compute, in function of the generation n and for different thresholds, the proportion of rejections under hypotheses $\mathbf{H_0}$ and $\mathbf{H_1}$, the latter being an indicator of the power of the test. Proportions are computed on a sample of 1000 replicated trees. In Table 1, the observed proportions of p-values under the given thresholds are close to the expected proportions of rejection under

Generation	Under $\mathbf{H_0^{GW}}$			Under $\mathbf{H_1^{GW}}$		
	$p < 0.05$	$p < 0.01$	$p < 0.001$	$p < 0.05$	$p < 0.01$	$p < 0.001$
7	6.4	1.9	0.3	27.8	11.8	03.6
8	5.6	1.4	0.3	44.2	22.2	07.6
9	5.5	1.1	0.3	58.6	38.5	17.0
10	5.7	1.5	0.2	79.4	60.8	35.9
11	4.8	1.0	0.1	93.1	82.0	64.2

Table 1: Proportions of p-values under the 0.05, 0.01 and 0.001 thresholds of the asymmetry tests for the means of the GW process (1000 replicas) $\mathbf{p}^{(0)} = (0.04, 0.08, 0.08, 0.8)$ (under ($\mathbf{H1}$), $\mathbf{p}^{(1)} = (0.15, 0.08, 0.08, 0.69)$)

$\mathbf{H_0^{GW}}$ suggesting that the asymptotic law of the statistic $(\mathbf{Y}_n^{GW})^2$ is available by generation 8. Under $\mathbf{H_1^{GW}}$, the power of the test increases from 27.8 % for generation 7 to 93.1 % for generation 11 for a risk of type 1 fixed at 0.05. In Table 2, the observed proportions of p-values under the given thresholds, are close to the expected proportions of rejection under $\mathbf{H_0^c}$ suggesting that the asymptotic law of the statistic $\|\mathbf{Y}_n^c\|^2$ is also available at generation 8. Under

$\mathbf{H_1^c}$, the power of the test increases from 37.4 % for generation 7 to 95.7 % for generation 11 for a risk of type 1 fixed at 0.05. In Table 3, the observed

Gen	Under $\mathbf{H_0^c}$			Under $\mathbf{H_1^c}$		
	$p < 0.05$	$p < 0.01$	$p < 0.001$	$p < 0.05$	$p < 0.01$	$p < 0.001$
7	6.6	2.2	0.6	37.4	19.7	08.0
8	5.5	1.5	0.3	53.6	31.0	14.6
9	5.5	1.3	0.3	71.1	52.3	30.3
10	6.3	1.2	0.1	86.8	75.5	56.1
11	5.9	0.6	0.1	95.7	90.8	81.4

Table 2: Proportions of p-values under the 0.05, 0.01 and 0.001 thresholds of the asymmetry test for the parameters of the BAR process (1000 replicas) $a = b = 0.5$ (under (**H1**), $c = 0.5; d = 0.4$)

proportions go away from the expected ones under $\mathbf{H_0^f}$, suggesting that the asymptotic law of the statistic is not reached before the 10th generation. The power is also weak until the 10th generation.

6. Application to real data: aging detection of Escherichia coli

To study aging in the single cell organism E. coli, Stewart et al. (2005) filmed 94 colonies of dividing cells, determining the complete lineage and the growth rate of each cell. E. coli is a rod-shaped bacterium that reproduces by dividing in the middle. Each cell inherits an old end or *pole* from its mother, and creates a new pole. Therefore, each cell has a *type*: old pole or new pole inducing asymmetry in the cell division. Stewart et al. (2005) propose

Gen	Under $\mathbf{H_0^f}$			Under $\mathbf{H_1^f}$		
	$p < 0.05$	$p < 0.01$	$p < 0.001$	$p < 0.05$	$p < 0.01$	$p < 0.001$
7	2.2	0.7	0	23.1	07.4	01.4
8	3.3	0.5	0.1	41.3	20.5	06.1
9	3.8	0.5	0	64.6	41.6	18.6
10	4.7	0.8	0	82.9	68.1	46.3
11	5.5	0.7	0.1	94.5	88.5	74.5

Table 3: Proportions of p-values under the 0.05, 0.01 and 0.001 thresholds of the asymmetry test for the fixed points of the BAR process (1000 replicas) $a = b = 0.5$ (under (**H1**), $c = 0.5; d = 0.4$)

a statistical study of the mean genealogy and pair-wise comparison of sister cells assuming their independence which is not verified in the lineage. We ran our tests of the null hypotheses $\mathbf{H_0^{GW}}$, $\mathbf{H_0^c}$ and $\mathbf{H_0^f}$ on the 51 data sets

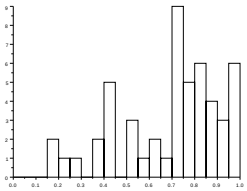
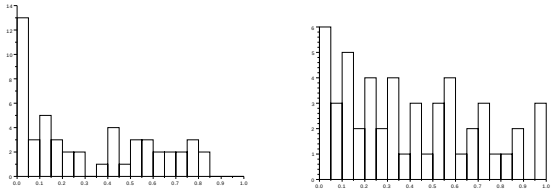


Figure 1: Histogram of the 51 p-values of the test $\mathbf{H_0^{GW}}$

issued of the 94 colonies containing at least eight generations. Figure 1 shows that the hypothesis of equality of the expected number of observed offspring between two sisters is not rejected whatever the data set. This result is not

surprising: data are missing most frequently because the cells were out of the range of the camera. The null hypotheses of the two tests are rejected for



(a) $\mathbf{H_0^c}$ (b) $\mathbf{H_0^f}$

Figure 2: Histogram of the 51 p-values of the tests $\mathbf{H_0^c}$ and $\mathbf{H_0^f}$.

one set in four for $\mathbf{H_0^c}$ and for one in eight for $\mathbf{H_0^f}$. A global conclusion on the asymmetry between the old pole and new pole cells is not easy. Regarding the simulations results in Tables 2 and 3, this lack of evidence is probably due to a low power of the tests at generations 8 and 9.

References

Bercu, B., de Saporta, B., Gégout-Petit, A., 2009. Asymptotic analysis for bifurcating autoregressive processes via a martingale approach. Electron. J. Probab. 14, no. 87, 2492–2526.

Cowan, R., Staudte, R.G., 1986. The bifurcating autoregressive model in cell lineage studies. Biometrics 42, 769–783.

Delmas, J.F., Marsalle, L., 2010. Detection of cellular aging in a Galton-Watson process. Stoch. Process. and Appl. 120, 2495–2519.

Duflo, M., 1997. Random iterative models. volume 34 of *Applications of Mathematics*. Springer-Verlag, Berlin.

Guttorp, P., 1991. Statistical inference for branching processes. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York.

Guyon, J., 2007. Limit theorems for bifurcating Markov chains. Application to the detection of cellular aging. Ann. Appl. Probab. 17, 1538–1569.

Harris, T.E., 1963. The theory of branching processes. Die Grundlehren der Mathematischen Wissenschaften, Bd. 119, Springer-Verlag, Berlin.

Hwang, S.Y., Basawa, I.V., 2009. Branching Markov processes and related asymptotics. J. Multivariate Anal. 100, 1155–1167.

Hwang, S.Y., Basawa, I.V., 2011. Asymptotic optimal inference for multivariate branching-Markov processes via martingale estimating functions and mixed normality. J. Multivariate Anal. 102, 1018–1031.

de Saporta, B., Gégout-Petit, A., Marsalle, L., 2011. Parameters estimation for asymmetric bifurcating autoregressive processes with missing data. Electron. J. Statist. 5, 1313–1353.

Stewart, E., Madden, R., Paul, G., Taddei, F., 2005. Aging and death in an organism that reproduces by morphologically symmetric division. PLoS Biol. 3, e45.