Analyse asymptotique des processus autoregressifs de bifurcation avec données manquantes

Benoîte de Saporta, Anne Gégout-Petit, Laurence Marsalle

▶ **To cite this version:**

Benoîte de Saporta, Anne Gégout-Petit, Laurence Marsalle. Analyse asymptotique des processus autoregressifs de bifurcation avec données manquantes. 42èmes Journées de Statistique, 2010, Marseille, France. inria-00494793

HAL Id: inria-00494793
https://hal.inria.fr/inria-00494793

Submitted on 24 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analyse asymptotique des processus autoregressifs de bifurcation avec données manquantes

Benoîte de Saporta[1] & Anne Gégout-Petit[2] & Laurence Marsalle[3]

1. Université de Bordeaux, GREThA, CNRS, UMR 5113, IMB CNRS, UMR 5251, INRIA Bordeaux Sud-Ouest CQFD, 351, Avenue de la libération F33405 Talence, France
2. Université de Bordeaux, IMB CNRS, UMR 5251, INRIA Bordeaux Sud-Ouest CQFD, 351, Avenue de la libération F33405 Talence, France
3. Université de Lille 1, U.M.R. CNRS 8524, U.F.R. de Mathématiques 59 655 Villeneuve d'Ascq Cedex

Abstract We study the asymptotic behavior of the least squares estimators of the unknown parameters of bifurcating autoregressive processes with missing data. The process of missing data is modelled by a two-type Galton-Watson process. Under very weak assumptions on the driven noise of the process, namely conditional pair-wise independence and suitable moment conditions, we establish the almost sure convergence of our estimators. All our analysis relies on non-standard asymptotic results for martingales.

Résumé Nous étudions le comportement asymptotique de l'estimateur des paramètres d'un processus autorégressif de bifurcation dans le contexte de données manquantes. Les données manquantes sont modélisées par un processus de Galton-Watson multi-type à deux catégories. Sous des hypothèses faibles sur le bruit de l'autorégression, (indépendance conditionnelle paire par paire et conditions de moments), nous établissons la convergence presque sûre de notre estimateur. Notre travail repose sur des résultats asymptotiques non-standard pour les martingales.

Mots clés processus autorégressif de bifurcation ; processus de Galton-Watson multi-type ; données manquantes ; séries temporelles indexées par un arbre ; martingales ; estimateur des moindres carrés ; convergence presque sûre.

1 Introduction, processus BAR.

Les processus autorégressifs de bifurcation (BAR) sont une adaptation des processus autorégressifs à des données structurées en arbre binaire. Ils ont été introduits par Cowan and Staudte [2] pour des données de division cellulaire, chaque individu d'une génération donnant naissance à deux filles à la génération suivante. Les données correspondent à une caractéristique quantitative associée à chaque cellule sur plusieurs générations (diamètre, taux de croisssance,...). L'évolution de cette caractéristique dépend de la cellule mère via

l'autorégression et des conditions environnementales (bruit).

La cellule initiale est numérotée 1, et les deux descendants de la cellule n sont numérotés $2n$ et $2n+1$. X_n est la caractéristique quantitative de l'individu n. L'équation du processus BAR d'ordre 1 pour $n \geq 1$ est donnée par :

$$\left\{ \begin{array}{rcccccl} X_{2n} & = & a_{2n} & + & b_{2n}X_n & + & \varepsilon_{2n}, \\ X_{2n+1} & = & a_{2n+1} & + & b_{2n+1}X_n & + & \varepsilon_{2n+1}. \end{array} \right. \tag{1}$$

Les $(\varepsilon_{2n}, \varepsilon_{2n+1})$ correspondent au bruit environnemental, les hypothèses sur ce bruit sont données à la section 4. Nous supposons que

$$\left\{ \begin{array}{l} a_{2n} = a, \\ b_{2n} = b, \end{array} \right. \quad \text{et} \quad \left\{ \begin{array}{l} a_{2n+1} = c, \\ b_{2n+1} = d. \end{array} \right.$$

avec $(a, b, c, d) \in \mathbb{R}^4$ et

$$0 < \max(|b|, |d|) < 1 \qquad \text{and} \qquad |a| + |c| \neq 0.$$

Comme expliqué dans [1], un processus BAR est un processus AR indexé par un arbre binaire où chaque noeud représente un individu. Pour tout $n \geq 1$, la n-ieme génération est notée $\mathbb{G}_n = \{2^n, 2^n + 1, \ldots, 2^{n+1} - 1\}$ et $\mathbb{G}_0 = \{1\}$, $\mathbb{G}_1 = \{2, 3\}$ sont les premières générations. On note $\mathbb{T}_n = \bigcup_{k=0}^{n} \mathbb{G}_k$ le sous-arbre des individus jusqu'à la n-ieme génération et $\mathbb{T}$ l'arbre dans sa totalité : $\mathbb{T} = \bigcup_{n=0}^{\infty} \mathbb{T}_n$.

Par exemple, à la figure 1, nous donnons un arbre à 5 générations, les cellules en pointillés sont celles qui ne sont pas observées. Le processus générant ces données manquantes est introduit à la section suivante.

2 Processus générant les données manquantes

Le processus générant les données manquantes est noté $(\delta_k)_{k \in \mathbb{T}} \in \{0, 1\}^{\mathbb{T}}$. La caractéristique de la cellule k est observée si $\delta_k = 1$ ou encore $\delta_k = \mathbb{I}_{\{X_k \text{ est observé }\}}$. Comme le processus BAR sous jacent est dissymétrique, nous distinguons deux type de cellules : les cellules paires (type 0) et impaires (type 1). On note :

$p^i(1,1)$	la probabilité qu'une cellule de type i donne naissance à	deux filles
$p^i(1,0)$	"	une fille de type 0
$p^i(0,1)$	"	une fille de type 1
$p^i(0,0)$	"	aucune fille

On en déduit la matrice P avec $(p_{ij})_{0 \leq i,j \leq 1}$ le nombre moyen de descendants de type j d'une cellule de type i :

$$\left(\begin{array}{cc} p^0(1,1) + p^0(1,0) & p^0(1,1) + p^0(0,1) \\ p^1(1,1) + p^1(1,0) & p^1(1,1) + p^1(0,1) \end{array} \right) \tag{2}$$

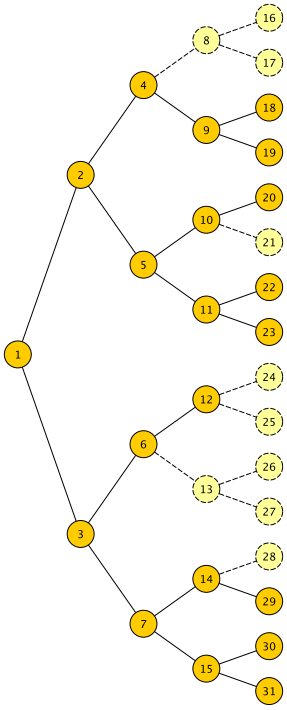

FIG. 1 – Arbre associé à un processus de bifurcation. Les cellules en pointillés ne sont pas observées.

On a alors $E[\delta_{4k}|\delta_{2k} \neq 0] = p_{00}$, $E[\delta_{4k+1}|\delta_{2k} \neq 0] = p_{01}$, etc.

Nous faisons les hypothèses suivantes sur le processus (δ_k).

(HP.1) Les procesus $(\delta_n)_{n\in\mathbb{T}}$ et $(X_n)_{n\in\mathbb{T}}$ sont indépendants

(HP.2) $\delta_2 = \delta_3 = 1$ et si une cellule est manquante alors ses descendants le sont aussi.

(HP.3) Quel que soit $(i,j) \in \{0,1\}^2$, $p_{ij} > 0$.

(HP.4) Le rayon spectral π de la matrice P vérifie $\pi > 1$.

Les hypothèses énoncées ci-dessus entraînent que pour un $k > 1$ donné, la loi du vecteur $(\delta_{2^j.k}, \ldots, \delta_{2^j.k+2^j-1})$ sachant $\delta_k \neq 0$ est identique à celle du vecteur $(\delta_{2^j.2}, \ldots, \delta_{2^j.2+2^j-1})$ si $k \equiv 0[2]$ ou celle de $(\delta_{2^j.3}, \ldots, \delta_{2^j.3+2^j-1})$ si $k \equiv 1[2]$.

Si on note

$$Z_n = (Z_n^0, Z_n^1) = \left(\sum_{k\in\mathbb{G}_{n-1}} \delta_{2k}, \sum_{k\in\mathbb{G}_{n-1}} \delta_{2k+1} \right) \qquad n \geq 1, \tag{3}$$

Z_n^0 (resp. Z_n^1) est le nombre de cellules paires (resp. impaires) à la génération n. $(Z_n)_{n\geq 1}$ est un processus de Galton-Watson multi-type qui vérifie $E[Z_{n+1}] = {}^tPE[Z_n]$. On note $|\mathbb{G}_n^*| = Z_n^0 + Z_n^1$ le nombre d'observations à la génération n et $|\mathbb{T}_n^*| = \sum_{k=0}^n |\mathbb{G}_k^*|$, le nombre

total d'observations jusqu'à la génération n. Les propriétés détaillées du processus Z_n sont données dans [5] et utilisées tout au long de ce travail. Nous rappelons ici la principale :

Proposition 2.1 *Sous les hypothèses* (**HP.2**)–(**HP.4**)*, il existe un événement A, tel que* $P(A) > 0$ *et*

$$\lim_{n\to\infty} \mathbb{I}_A \frac{Z_n^0}{|\mathbb{G}_n^*|} = z^0 \quad \text{et} \quad \lim_{n\to\infty} \mathbb{I}_A \frac{Z_n^1}{|\mathbb{G}_n^*|} = z^1 \qquad p.s.$$

On en déduit que le nombre d'observations tend vers l'infini sur A. On va démontrer des résultats de convergence p.s. sur cet ensemble.

3 Estimateur

Notre but est d'estimer $\theta = (a, b, c, d)^t$ d'après l'observation des individus non manquants jusqu'à la génération n. Nous proposons l'estimateur des moindres carrés $\widehat{\theta}_n$ qui minimise

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

Celui-ci est donné par

$$(\widehat{\theta}_n) = \begin{pmatrix} \widehat{a}_n \\ \widehat{b}_n \\ \widehat{c}_n \\ \widehat{d}_n \end{pmatrix} = \begin{pmatrix} \left(S_{n-1}^0\right)^{-1} & 0 \\ 0 & \left(S_{n-1}^1\right)^{-1} \end{pmatrix} \sum_{k\in\mathbb{T}_{n-1}} \begin{pmatrix} \delta_{2k}X_{2k} \\ \delta_{2k}X_kX_{2k} \\ \delta_{2k+1}X_{2k+1} \\ \delta_{2k+1}X_kX_{2k+1} \end{pmatrix}, \tag{4}$$

avec $S_{n-1}^0 = \sum_{k\in\mathbb{T}_{n-1}} \delta_{2k} \begin{pmatrix} 1 & X_k \\ X_k & X_k^2 \end{pmatrix}$, $S_{n-1}^1 = \sum_{k\in\mathbb{T}_{n-1}} \delta_{2k+1} \begin{pmatrix} 1 & X_k \\ X_k & X_k^2 \end{pmatrix}$. De même on utilisera plus bas la notation $S_{n-1}^{0,1} = \sum_{k\in\mathbb{T}_{n-1}} \delta_{2k}\delta_{2k+1} \begin{pmatrix} 1 & X_k \\ X_k & X_k^2 \end{pmatrix}$.

On déduit facilement des équation (1) et (4) que

$$\widehat{\theta}_n - \theta = \begin{pmatrix} \left(S_{n-1}^0\right)^{-1} & 0 \\ 0 & \left(S_{n-1}^1\right)^{-1} \end{pmatrix} \sum_{k\in\mathbb{T}_{n-1}} \begin{pmatrix} \delta_{2k}\varepsilon_{2k} \\ \delta_{2k}X_k\varepsilon_{2k} \\ \delta_{2k+1}\varepsilon_{2k+1} \\ \delta_{2k+1}X_k\varepsilon_{2k+1} \end{pmatrix} = \begin{pmatrix} \left(S_{n-1}^0\right)^{-1} & 0 \\ 0 & \left(S_{n-1}^1\right)^{-1} \end{pmatrix} M_n \tag{5}$$

4 Propriétés de Martingales

Nos résultats de convergence reposent sur le fait que la suite (M_n) introduite ci-dessus est une martingale. Nous introduisons les différentes filtrations : $\mathcal{F}_n$ est la σ-algèbre des individus de l'arbre $\mathbb{T}_n$: $\mathcal{F}_n = \sigma\{X_k, k \in \mathbb{T}_n\}$. Pour ce qui de la tribu observée $\mathcal{O}_n$, nous supposons que toute l'histoire du processus δ est connue à l'instant 0 : $\mathcal{D} = \sigma\{\delta_k, k \in \mathbb{T}\}$ et $\mathcal{O}_n = \mathcal{D} \vee \sigma\{\delta_k X_k, k \in \mathbb{T}_n\}$. Nous faisons les hypothèses suivantes sur le processus (ε_n) :

(H.1) Pour tout $n \geq 0$ et $\forall k \in \mathbb{G}_{n+1}$, $\mathbb{E}[\varepsilon_k|\mathcal{F}_n] = 0$ et $\mathbb{E}[\varepsilon_k^2|\mathcal{F}_n] = \sigma^2 > 0$ p.s.

(H.2) Pour tout $n \geq 0$ et pour $k \neq l \in \mathbb{G}_{n+1}$, si $[k/2] \neq [l/2]$, ε_k et ε_l sont conditionnellement indépendants sachant $\mathcal{F}_n$, et sinon $\mathbb{E}[\varepsilon_k\varepsilon_l|\mathcal{F}_n] = \rho$ p.s. avec $\rho < \sigma^2$

(H.3)

$$\sup_{n\geq 0} \sup_{k\in\mathbb{G}_{n+1}} \mathbb{E}[\varepsilon_k^4|\mathcal{F}_n] < \infty \qquad \text{p.s.}$$

Proposition 4.1 *Sous les hypothèses* (**HP.1**), (**H.1**)–(**H.3**), *le processus* M_n *défini par (5) est une* $\mathcal{O}_n$*-martingale de carré intégrable dont le crochet est donné par :*

$$< M >_n = \begin{pmatrix} \sigma^2 S^0_{n-1} & \rho S^{0,1}_{n-1} \\ \\ \rho S^{0,1}_{n-1} & \sigma^2 S^1_{n-1} \end{pmatrix}$$

Comme dans [1], le résultat clé est la convergence des matrices S^0_{n-1}, S^1_{n-1} et $S^{0,1}_{n-1}$ normalisées par $|\mathbb{T}^*_n|$ vers des limites finies.

5 Résultats

Proposition 5.1. *Supposons que* (ε_n) *vérifie* (**H.1**) – (**H.3**) *et que le processus de Galton Watson vérifie* (**HP.1**) – (**HP.4**). *Alors il existe des matrices finies* L^0, L^1 *et* $L^{0,1}$ *telles que*

$$\lim_{n\to\infty} \frac{S^1_n}{|\mathbb{T}^*_n|} = \mathbb{I}_A L^1 \qquad \lim_{n\to\infty} \frac{S^0_n}{|\mathbb{T}^*_n|} = \mathbb{I}_A L^0 \qquad \lim_{n\to\infty} \frac{S^{0,1}_n}{|\mathbb{T}^*_n|} = \mathbb{I}_A L^{0,1} \qquad a.s. \quad (6)$$

En utilisant des technniques de martingales comme dans [1], ce résultat implique la convergence p.s. de $\widehat{\theta}_n$ sur l'ensemble des A.

Théorème 5.1. *Sous les hypothèses* (**H.1**) – (**H.3**) *et* (**HP.1**) – (**HP.4**), $\mathbb{I}_A\widehat{\theta}_n$ *converge p.s. vers* $\mathbb{I}_A\theta$.

Références

[1] Bercu, B., de Saporta, B., Gégout-Petit, A. Asymptotic analysis for bifurcating autoregressive processes via a martingale approach. *Electronic Journal of Probability*, 14, 2492–2526, 2009.

[2] Cowan, R., Staudte, R. G. The bifurcating autoregressive model in cell lineage studies. *Biometrics 42* (1986), 769–783.

[3] Delmas, J.-F., Marsalle, L. Detection of cellular aging in a Galton-Watson process. *arXiv, 0807.0749* (2008).

[4] Duflo, M. *Random iterative models*, vol. 34 of *Applications of Mathematics*. Springer-Verlag, Berlin, 1997.

[5] Harris, T. E. *The Theory of Branching Processes*. Cambridge University Press, Cambridge, 1985.