Estimation et identification de modèles autorégressifs non-linéaires multidimensionnels

Joseph Rynkiewicz
SAMOS/MATISSE, UMR 8595, Université de Paris1
rynkiewi@univ-paris1.fr

29 septembre 2000

Résumé

Ce travail concerne l'estimation paramétrique de modèle autorégressifs non-linéaires multidimensionnels et ses propriétés asymptotiques. Pour une série scalaire l'estimateur du maximum de vraisemblance d'un modèle dont l'innovation est gaussienne, coïncide avec l'estimateur des moindres carrés. Néanmoins, ce n'est, en général, plus vraie lorsque les observations sont multidimensionnels. Dans ce cas l'estimateur du maximum de vraisemblance pour un modèle gaussien est le paramètre qui minimise le déterminant de la covariance empirique du bruit. Sans hypothèses de normalité on montrera la consistance forte et la normalité asymptotique de cet estimateur. De plus, on donne une loi du logarithme itéré pour cet estimateur. Ces propriétés fourniront un critère d'identification presque sûre du modèle dans le cadre de la sélection de modèles par un contraste pénalisé comme le BIC. Finalement nous appliquerons ces résultats au perceptrons multicouches.

This work concerns parametric estimation for nonlinear multidimensional autoregressive models and its asymptotic properties. For a scalar series, the least square estimator match with the gaussian maximum likelihood estimator. But, in general, it is not true when the obervations are multidimensionnal. In this case the gaussian maximum likelihood estimator is the parameter minimizing the determinant of the error covariance matrix. Without normality asumption, we establish the strong consistency and the asymptotic normality of this estimator. Furthermore, we give a law of the iterated logarithm for this estimator. These properties yield a result about almost sure identification of the true model within the framework of model selection using penalized contrast, like BIC. Finally, we apply these results to multilayer perceptron.

Classification AMS : 62 F 12, 62 M 10

Keywords Nonlinear AR process, Maximum likelihood estimator, law of iterated logarithm, almost sur model identification, multilayer perceptron.

1 Le modèle

Pour une série (Y_t), $t \in \mathbb{N}^*$ on notera Y_t^{t+l} le vecteur $(Y_t, \cdots, Y_{t+l})^T$ avec $l \in \mathbb{N}^*$.
On considère le modèle suivant :

$$Y_{t+1} = F_W(Y_{t-p+1}^t) + \varepsilon_{t+1} \tag{1}$$

où

- p est l'ordre de régression du modèle.
- (Y_t), $t \in \mathbb{Z}$, $t \geq -p+1$ est une suite de variables aléatoires de $\mathbb{R}^d$.
- F_W est une fonction paramétrique, continûment dérivable par rapport à ses paramètres, avec pour vecteur paramètre $W \in \mathbb{R}^D$,
- (ε_t), $t \in \mathbb{N}^*$, est une suite de variables aléatoires vectorielles indépendantes identiquement distribuées de matrice de covariance $\Gamma \in \mathbb{R}^{(d+1)d/2}$ inversible et inconnue.
- On supposera les observations initiales y^0_{-p+1} connues et fixées

Notation 1 *On note le vecteur paramètre* $\theta = (W, \Gamma^{-1}) \in \Theta = \Theta^W \times \Theta^{\Gamma^{-1}}$ *, où* $W \in \Theta^W \subset \mathbb{R}^D$, $\Gamma^{-1} \in \Theta^{\Gamma^{-1}} \subset \mathbb{R}^{\frac{d(d+1)}{2}}$ *donc en posant* $B = D + (d+1)\,d/2$, B *est le nombre total de paramètres et* $\theta \in \Theta \subset \mathbb{R}^B$. *Pour simplifier les calculs on estime* Γ^{-1} *à la place de* Γ, *puisque cette matrice est supposée inversible, les deux paramétrisations sont équivalentes.*

Notation 2 *Dans toute la suite, si* X *est un vecteur multidimensionnel,* $X(i)$ *désignera sa i ème coordonnée.*

Notation 3 *Si* X *est une matrice inversible on notera* x^{-1}_{ij} *les coefficients de* X^{-1}.

Notation 4 *On note* $\mathbb{F}_t$ *la tribu engendrée par* $Y_{-p+1}, \cdots, Y_t$, $t > 0$

Remarque 1 *Toute cette étude est valable si on considère que* Θ *est inclus dans un espace polonais, c'est-à-dire un espace métrique complet et séparable, mais par souci de clarté on exposera les résultats pour des paramètres réels.*

Nous supposons, dans un premiers temps, que l'innovation ε est gaussienne, on peut alors calculer la log-vraisemblance du modèle. En effet, la vraisemblance s'écrit en fonction des observations $(y_1, \cdots, y_n)$ et du paramètre θ.

$$L_\theta(y_1, ..., y_n) = \frac{1}{(2\pi)^d \det(\Gamma)}^{\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{t=1}^{n}(y_t - F_W(y_{t-p}^{t-1}))^T\Gamma^{-1}(y_t - F_W(y_{t-p}^{t-1}))\right)$$

d'où la log-vraisemblance :

$$l_\theta(y_1, \cdots, y_n) := \ln(L_\theta(y_1, ..., y_n))$$

$$= -\frac{n}{2}\ln(\det(\Gamma)) - \frac{1}{2}\sum_{t=1}^{n}(y_t - F_W(y_{t-p}^{t-1}))^T\Gamma^{-1}(y_t - F_W(y_{t-p}^{t-1})) + Cte. \tag{2}$$

2 Maximisation de la log-vraisemblance

On va d'abord montrer que pour maximiser cette fonction, on peut simplement trouver les paramètres $\hat{W}_n$ qui minimisent le déterminant de la covariance empirique du bruit, et poser $\hat{\Gamma}_n^{-1}$ égale à l'inverse de la covariance empirique calculée grâce aux paramètres $\hat{W}_n$.

2.1 Expression de $\hat{\Gamma}_n^{-1}$ en fonction de $\hat{W}_n$

On rappelle trois formules classiques que l'on utilisera par la suite :
- Si A de coefficient a_{ij}, est une matrice constante et X une matrice de coefficients x_{ij} :

$$\frac{\partial}{\partial x_{ij}} Tr(AX) = a_{ji}. \tag{3}$$

– En supposant maintenant X inversible on a :

$$\frac{\partial}{\partial x_{ij}} \ln(\det(X)) = x_{ji}^{-1}. \tag{4}$$

– Si A, B, C sont trois matrices de taille convenable, la trace de leur produit est invariante par permutation circulaire :

$$Tr(ABC) = Tr(BCA) = Tr(CAB). \tag{5}$$

On a le résultat suivant :

Proposition 1 *Notons $\Gamma_n(W)$ la covariance empirique :*

$$\Gamma_n(W) = \frac{1}{n}\sum_{t=1}^{n}(y_t - F_W(y_{t-p}^{t-1}))(y_t - F_W(y_{t-p}^{t-1}))^T$$

et

$$\Gamma_n^{-1}(W) = \left(\frac{1}{n}\sum_{t=1}^{n}(y_t - F_W(y_{t-p}^{t-1}))(y_t - F_W(y_{t-p}^{t-1}))^T\right)^{-1}.$$

Soit $\hat{\theta}_n = (\hat{W}_n, \hat{\Gamma}_n^{-1})$, l'estimateur du maximum de vraisemblance, on a :

$$\hat{W}_n = \arg\min_{W\in\Theta^W}\left(\frac{1}{2}\ln\det\left(\Gamma_n(W)\right)\right) \tag{6}$$

et

$$\hat{\Gamma}_n^{-1} = \left(\Gamma_n(\hat{W}_n)\right)^{-1}$$

en supposant que $\Gamma_n(\hat{W}_n)$ est bien inversible.

Preuve Fixons les paramètres W, la dérivée de la log-vraisemblance (multipliée par $\frac{2}{n}$) par rapport au coefficient Γ_{ij}^{-1} s'écrit :

$$\frac{\partial}{\partial\Gamma_{ij}^{-1}}(\ln(\det(\Gamma_n^{-1}(W))) - \frac{1}{n}\sum_{t=1}^{n}\frac{\partial}{\partial\Gamma_{ij}^{-1}}Tr\left((y_t - F_W(y_{t-p}^{t-1}))^T\Gamma^{-1}(y_t - F_W(y_{t-p}^{t-1}))\right)$$

grâce aux formules (4) et (5) cela s'écrit :

$$\Gamma_{ji} - \frac{1}{n}\sum_{t=1}^{n}\frac{\partial}{\partial\Gamma_{ij}^{-1}}Tr\left((y_t - F_W(y_{t-p}^{t-1}))(y_t - F_W(y_{t-p}^{t-1})^T\Gamma^{-1}\right).$$

En utilisant la formule (3) on obtient :

$$\frac{\partial\frac{2}{n}l_\theta(y_1,\cdots,y_n)}{\partial\Gamma_{ij}^{-1}} = \Gamma_{ji} - \frac{1}{n}\sum_{t=1}^{n}\left(y_t - F_W(y_{t-p}^{t-1})\right)(j)\times\left(y_t - F_W(y_{t-p}^{t-1})\right)(i)$$

Les fonctions

$$\Gamma^{-1} \longmapsto (\ln(\det(\Gamma^{-1})) - \frac{1}{n}\sum_{t=1}^{n}Tr\left((y_t - F_W(y_{t-p}^{t-1}))^T\Gamma^{-1}(y_t - F_W(y_{t-p}^{t-1}))\right)$$

et

$$\Gamma^{-1} \longmapsto (\ln(\det(\Gamma^{-1}))$$

ont la même matrice hessienne (cf formules (3) et (5)). De plus la fonction $\Gamma^{-1} \longmapsto \ln\det(\Gamma^{-1})$ est une fonction strictement concave sur l'ensemble convexe des matrices symétriques définies positives (cf [7] théorème 7.6.7) donc

$$\Gamma^{-1} \longmapsto (\ln(\det(\Gamma^{-1})) - \frac{1}{n}\sum_{t=1}^{n} Tr\left((y_t - F_W(y_{t-p}^{t-1}))^T\Gamma^{-1}(y_t - F_W(y_{t-p}^{t-1}))\right)$$

aussi.

Elle atteint son maximum en l'unique point où sa dérivée s'annule. Ainsi, pour W fixé, le maximum de la log-vraisemblance en fonction de la matrice Γ^{-1} s'exprime en fonction de W :

$$\hat{\Gamma}_n^{-1}(W) = \left(\frac{1}{n}\sum_{t=1}^{n}(y_t - F_W(y_{t-p}^{t-1}))(y_t - F_W(y_{t-p}^{t-1}))^T\right)^{-1}.$$

Le vecteur paramètre $\hat{\theta}_n = (\hat{W}_n, \hat{\Gamma}_n^{-1})$ qui maximise la log-vraisemblance vérifie donc :

$$\hat{W}_n = \arg\min_{W\in\Theta^W}\left(\frac{n}{2}\ln(\det(\Gamma_n(W))) + \frac{1}{2}\sum_{t=1}^{n} Tr(y_t - F_W(y_{t-p}^{t-1})^T\Gamma_n^{-1}(W)(y_t - F_W(y_{t-p}^{t-1}))\right) \tag{7}$$

et

$$\hat{\Gamma}_n^{-1} = \left(\Gamma_n(\hat{W}_n)\right)^{-1}.$$

Ce qui est équivalent, en remplaçant $\Gamma_n^{-1}(W)$ par $\Gamma_n^{-1}(\hat{W}_n)$ dans (7) à :

$$\hat{W}_n = \arg\min_{W\in\Theta^W}\left(\frac{1}{2}\ln\det\left(\Gamma_n(W)\right)\right)$$

et

$$\hat{\Gamma}_n^{-1} = \left(\Gamma_n(\hat{W}_n)\right)^{-1}$$

■

En général, on n'a pas de forme explicite de $\hat{W}_n = \arg\min_{W\in\Theta^W}(\ln\det(\Gamma_n(W)))$, néanmoins une solution acceptable est de savoir calculer la dérivée de cette fonction et d'approcher ce minimum par optimisation différentielle. Ainsi, pour maximiser la log-vraisemblance, il suffit d'optimiser cette fonction le long de la sous-variété de $(W,\Gamma) := \mathbb{R}^{(D+(d+1)*d/2)}$ définie par $\Gamma = \Gamma_n(W)$.

2.2 Dérivée de $\frac{1}{2}\ln\det(\Gamma_n(W))$

On suppose ici que $\forall y_1^p \in \left(\mathbb{R}^d\right)^p$, la fonction $W \longmapsto F_W(y_1^p)$ est continûment dérivable.

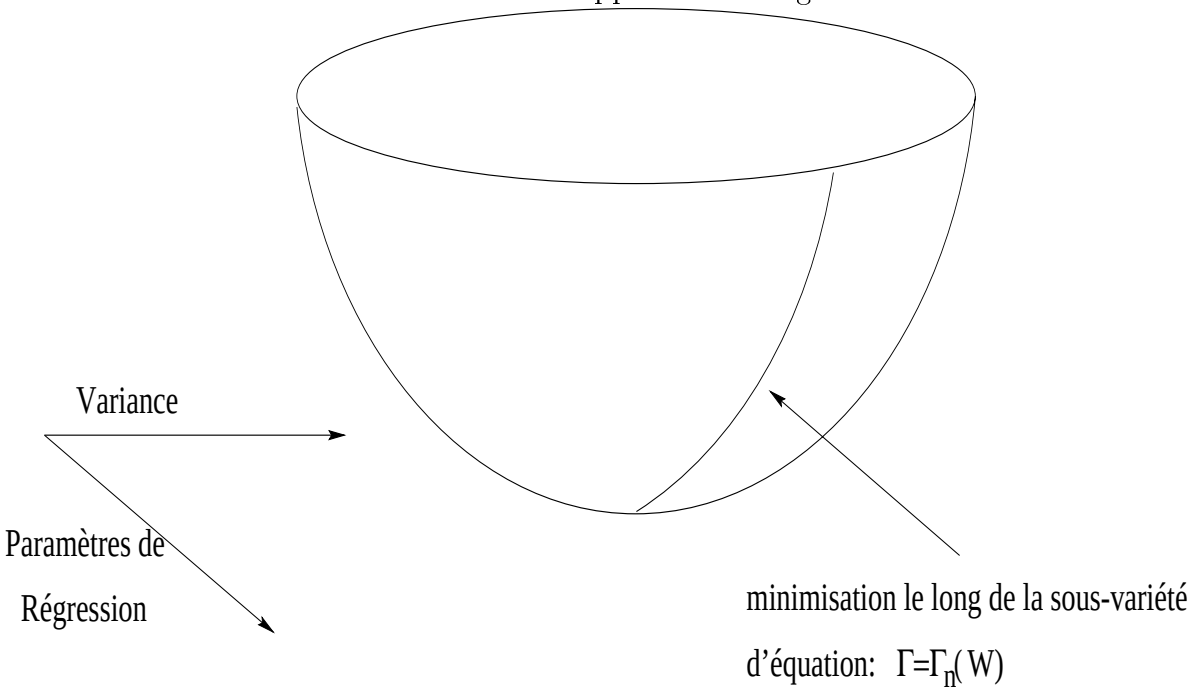
Notation 5 *Dans la suite, si X est une matrice symétrique, la notation :*

$$(X_{ij})_{ind} := (X_{ij})_{1\le i\le j\le d}$$

indiquera le vecteur :

$$(X_{11}, X_{12}, \cdots, X_{1d}, X_{22}, X_{23}, \cdots X_{2d}, \cdots, X_{dd})^T.$$

FIG. 1 – Minimisation de l'opposée de la log-vraisemblance



La fonction $\ln(\det(\Gamma_n(W)))$ s'exprime comme la fonction composée $f(g(W))$ avec :

- $g : \mathbb{R}^D \longrightarrow \mathbb{R}^{d(d+1)/2}$ telle que :

$$\Gamma_{ij} = \Gamma_{ji} = g_{ij}(W) = g_{ji}(W) := \frac{1}{n}\sum_{t=1}^{n}\left(y_t - F_W(y_{t-p}^{t-1})\right)(i) \times \left(y_t - F_W(y_{t-p}^{t-1})\right)(j).$$

- $f : \mathbb{R}^{d(d+1)/2} \longrightarrow \mathbb{R}$ telle que :

$$f(\Gamma) = \frac{1}{2}\ln(\det(\Gamma)).$$

Grâce à la formule de dérivée de fonction composée (Cartan [1] théorème 2.2.1), on aura pour tout $k \in \{1, ..., D\}$[1]

$$\frac{\partial}{\partial W_k}(\frac{1}{2}\ln(\det(\Gamma_n(W)))) = \left(\frac{\partial}{\partial \Gamma_{ij}}(\ln(\det(\Gamma_n(W)))\right)_{ind}^{T}\left(\frac{\partial \Gamma_{ij}}{\partial W_k}\right)_{ind}$$

avec

$$\frac{\partial \Gamma_{ij}}{\partial W_k} = \frac{1}{n}\sum_{t=1}^{n}\frac{\partial}{\partial W_k}\left[\left(y_t - F_W(y_{t-p}^{t-1})\right)(i) \times \left(y_t - F_W(y_{t-p}^{t-1})\right)(j)\right]$$

soit

$$\frac{\partial \Gamma_{ij}}{\partial W_k} = \frac{1}{n}\sum_{t=1}^{n}\left[-\frac{\partial F_W(y_{t-p}^{t-1})(i)}{\partial W_k} \times \left(y_t - F_W(y_{t-p}^{t-1})\right)(j) - \frac{\partial F_W(y_{t-p}^{t-1})(j)}{\partial W_k}\left(y_t - F_W(y_{t-p}^{t-1})\right)(i)\right]$$

et grâce à la formule (4) :

$$\frac{\partial}{\partial \Gamma_{ij}}\ln(\det(\Gamma_n(W)) = \Gamma_{ij}^{-1} = \Gamma_{ji}^{-1},$$

car la matrice $\Gamma_n^{-1}(W)$ est symétrique.

[1] La notation $\frac{\partial f(X)}{\partial X_k}$ signifie la k-ème coordonnée de la dérivée de $f(X)$ au point X

On en déduit la dérivée de la log-vraisemblance par rapport à l'élément W_k du vecteur paramètre W :

$$\frac{\partial}{\partial W_k}(\frac{1}{2}\ln(\det(\Gamma_n(W)))) = \left(\Gamma_{ij}^{-1}\right)_{ind}^T \left(\frac{\Gamma_{ij}}{\partial W_k}\right)_{ind}$$

et

$$\frac{\partial \Gamma_{ij}}{\partial W_k} = \frac{1}{n}\sum_{t=1}^{n}\left[-\frac{\partial F_W(y_{t-p}^{t-1})(i)}{\partial W_k} \times \left(y_t - F_W(y_{t-p}^{t-1})\right)(j) - \frac{\partial F_W(y_{t-p}^{t-1})(j)}{\partial W_k}\left(y_t - F_W(y_{t-p}^{t-1})\right)(i)\right]$$

d'où la formule de la dérivée :

$$\frac{\partial}{\partial W_k}(\frac{1}{2}\ln(\det(\Gamma_n(W)))) =$$

$$\left(\Gamma_{ij}^{-1}\right)_{ind}^T \left(\frac{1}{n}\sum_{t=1}^{n} -\frac{\partial F_W(y_{t-p}^{t-1})(i)}{\partial W_k} \times \left(y_t - F_W(y_{t-p}^{t-1})\right)(j) - \frac{\partial F_W(y_{t-p}^{t-1})(j)}{\partial W_k}\left(y_t - F_W(y_{t-p}^{t-1})\right)(i)\right)_{ind}. \tag{8}$$

Connaissant la dérivée de $\frac{1}{2}\ln\det(\Gamma_n(W))$, il est maintenant facile de la minimiser par une technique d'optimisation différentielle (cf Press et al. [9]).

3 Propriétés statistiques de l'estimateur

On appellera le contraste $U_n(\theta) = -\frac{1}{n}l_\theta(y_1,\cdots,y_n)$ (cf 3.2.1) : contraste associé à la vraisemblance. Nous allons montrer que l'estimateur du minimum de contraste converge presque sûrement vers le bon paramètre (consistance) et qu'il converge suffisamment vite (théorème de la limite centrale, loi du logarithme itéré) pour pouvoir aussi identifier le modèle par un contraste pénalisé. Cela, même dans le cas où l'innovation n'est plus gaussienne, mais garde un moment fini d'ordre suffisamment grand.

3.1 Hypothèses de base

La chaîne vectorisée $\left(Y_{t-p+1}^t\right)_{t>0}$vérifie l'équation :

$$Y_{t-p+1}^t = \begin{pmatrix} Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-p+1} \end{pmatrix} = \begin{pmatrix} F_W(Y_{t-1},\cdots,Y_{t-p}) \\ Y_{t-1} \\ \vdots \\ Yt-p+1 \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{9}$$

La loi forte des grands nombres est assurée par le théorème suivant : (cf Duflo [3])

Théorème 1 *Soit le modèle vérifiant (9). supposons que le bruit (ε_t) a une densité positive par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ avec un moment d'ordre $a \geq 1$. Si il existe des nombres positifs $\nu_1,\cdots,\nu_p$ tels que $\nu_1+\cdots+\nu_p<1$, une constante $\kappa \geq 0$ et une norme $\|.\|$de $\mathbb{R}^d$ satisfaisant pour tout $y \in \left(\mathbb{R}^d\right)^p$:*

$$\|F_{W_0}(y_1^p)\| \leq \nu_1\|y_1\| + \cdots + \nu_p\|y_p\| + \kappa,$$

alors la chaîne vectorisée $\left(Y_{t-p+1}^t\right)_{t>0}$ est stable, géométriquement ergodique et sa mesure invariante μ_0 a une densité par rapport à la mesure de Lebesgue qui admet un moment d'ordre a.

Remarque 2 *Soit $(Y_t)_{t\in\mathbb{N}^*}$ un processus ergodique à valeurs dans $\mathbb{R}^d$, alors $\forall q>0, \forall g$ fonction intégrable, le processus $\left(g(Y_{t-q}^t)\right)_{t\in\mathbb{N}^*}$ est un processus ergodique.*

Dans la suite, on supposera que le modèle vérifie les hypothèses (**H**) :

1. Le processus vérifie les hypothèses du théorème 1, avec un moment d'ordre $a \geq 2$
2. Θ est un compact de $\mathbb{R}^B$ et le vrai paramètre $\theta_0 = \left(W_0, \Gamma_0^{-1}\right)$ appartient à l'intérieur de Θ .
3. Pour tout $Y_1^p \in \left(\mathbb{R}^d\right)^p$, $F_W(Y_1^p)$ est continue sur Θ^W par rapport au paramètre W.
4. Pour toute matrice de covariance $\Gamma^{-1} \in \Theta^{\Gamma^{-1}}$, $\rho(\Gamma) \geq \lambda_{min} > 0$, où $\rho(\Gamma)$ est le rayon spectral de $\Gamma = \left(\Gamma^{-1}\right)^{-1}$.
5. Le modèle est supposé identifiable, c'est-à-dire : $F_{W'} = F_W \Leftrightarrow W' = W$.

3.2 Consistance de l'estimateur.

3.2.1 Vérification des propriétés de contraste :

La fonction $U_n(\theta)$ associée à la vraisemblance est définie par :

$$U_n(\theta) = -\frac{1}{n} l_\theta(y_1, \cdots, y_n)$$

$U_n(\theta)$ est un processus contraste relatif à une fonction $\theta \longmapsto K(\theta_0, \theta)$ si $U_n(\theta)$ est $\mathcal{F}_n$-adapté et si :

$$\lim_{n\to\infty} U_n(\theta) - U_n(\theta_0) \xrightarrow{p.s.} K(\theta_0, \theta) \geq 0$$

avec

$$K(\theta_0, \theta) = 0 \Leftrightarrow \theta = \theta_0$$

On a, en notant $\varepsilon_t^W = Y_t - F_W(Y_{t-p}^{t-1})$:

$$U_n(\theta) - U_n(\theta_0) = \frac{1}{n}(l_{\theta_0}(y_1, \cdots, y_n) - l_\theta(y_1, \cdots, y_n))$$

$$= \frac{1}{2}\ln((\frac{\det\Gamma}{\det\Gamma_0})) + \frac{1}{2n}\left(\sum_{t=1}^n \left(\varepsilon_t^W\right)^T \Gamma^{-1}\left(\varepsilon_t^W\right) - \sum_{t=1}^n \left(\varepsilon_t^{W_0}\right)^T \Gamma_0^{-1}\left(\varepsilon_t^{W_0}\right)\right).$$

On sait qu'on peut toujours trouver une base, qui diagonalise simultanément Γ_0 et Γ. Soit Z la matrice ayant pour vecteurs colonnes les vecteurs de cette base. $Z^T\Gamma_0 Z$ est la matrice identité et $Z^T\Gamma Z$ est une matrice diagonale dont les termes diagonaux sont notés $\sigma_i^2 > 0$, $i = 1, \cdots, d$. On aura :

$$\frac{\det\Gamma}{\det\Gamma_0} = \det\left(Z^{-1}\Gamma_0^{-1}\left(Z^T\right)^{-1} Z^T\Gamma Z\right) = \prod_{i=1}^d \sigma_i^2.$$

Posons, pour $W^* \in \{W, W_0\}$, $\widetilde{\varepsilon}_t^{W^*} = Z^{-1}\varepsilon_t^{W^*}$, on aura :

$$\sum_{t=1}^n \left(\varepsilon_t^{W_0}\right)^T \Gamma_0^{-1}\left(\varepsilon_t^{W_0}\right) = \sum_{t=1}^n \sum_{i=1}^d \left(\widetilde{\varepsilon}_t^{W_0}(i)\right)^2$$

et

$$\sum_{t=1}^{n} \left(\varepsilon_t^W\right)^T \Gamma^{-1} \left(\varepsilon_t^W\right) = \sum_{t=1}^{n} \sum_{i=1}^{d} \frac{1}{\sigma_i^2} \left(\widetilde{\varepsilon}_t^W(i)\right)^2,$$

donc

$$U_n(\theta) - U_n(\theta_0) = \frac{1}{2} \sum_{i=1}^{d} \ln(\sigma_i^2) + \frac{1}{2n} \sum_{t=1}^{n} \sum_{i=1}^{d} \frac{1}{\sigma_i^2} \left(\widetilde{\varepsilon}_t^W(i)\right)^2 - \frac{1}{2n} \sum_{t=1}^{n} \sum_{i=1}^{d} \left(\widetilde{\varepsilon}_t^{W_0}(i)\right)^2.$$

La condition **H**-1 assure que pour tout θ et toute norme $\|.\|$, $\left\|\widetilde{\varepsilon}_t^W\right\|^2$est intégrable par rapport à la mesure invariante μ_0. Par la loi forte des grands nombres on aura p.s. :

$$\lim_{n \longrightarrow \infty} U_n(\theta) - U_n(\theta_0) := K(\theta_0, \theta)$$

avec

$$K(\theta_0, \theta) = \frac{1}{2} \sum_{i=1}^{d} \left(\ln(\sigma_i^2) + \frac{1}{\sigma_i^2} E_{\theta_0} \left[\left(\widetilde{\varepsilon}^W(i)\right)^2 \right] - E_{\theta_0} \left[\left(\widetilde{\varepsilon}^{W_0}(i)\right)^2 \right] \right).$$

Maintenant Yao [12] montre que le processus associé aux moindres carrés est un contraste, grâce au théorème suivant :

Théorème 2 *Sous les hypothèses (**H**), en définissant :*

$$V_n(W) = \frac{1}{n} \sum_{t=1}^{n} \left\| Y_t - F_W(Y_{t-p}^{t-1}) \right\|^2$$

on aura :

$$\lim_{n \longrightarrow \infty} \left[V_n(W) - V_n(W_0) \right] \stackrel{p.s.}{=} \int_{\left(\mathbb{R}^d\right)^p} \left\| F_W\left(y_1^p\right) - F_{W_0}\left(y_1^p\right) \right\|^2 \mu_0\left(dy_1^p\right) \geq 0.$$

Il est facile de voir que ce théorème est vrai pour toute norme $\|.\|_Q$ associée à une matrice Q définie positive. Il est donc vrai pour la norme :

$$\|X\| = \|ZX\|_Q$$

où Q est une matrice définie positive de $\mathbb{R}^{d \times d}$ et $X \in \mathbb{R}^d$.

Cela implique que :

Lemme 1 *Pour tout $i \in \{1, \cdots, d\}$, et tout $W \in \Theta^W$*

$$E_{\theta_0} \left[\left(\widetilde{\varepsilon}^W(i)\right)^2 \right] \geq E_{\theta_0} \left[\left(\widetilde{\varepsilon}^{W_0}(i)\right)^2 \right].$$

Preuve Supposons qu'il existe $i \in \{1, \cdots, d\}$ tel que :

$$E_{\theta_0} \left[\left(\widetilde{\varepsilon}^W(i)\right)^2 \right] < E_{\theta_0} \left[\left(\widetilde{\varepsilon}^{W_0}(i)\right)^2 \right].$$

Sans perte de généralité, on peut supposer que c'est l'indice $i = 1$. Il existe alors $\delta_2, \cdots, \delta_d$ strictement positifs tels que :

$$E_{\theta_0} \left[\left(\widetilde{\varepsilon}^W(1)\right)^2 \right] + \sum_{i=2}^{d} \delta_i E_{\theta_0} \left[\left(\widetilde{\varepsilon}^W(i)\right)^2 \right] < E_{\theta_0} \left[\left(\widetilde{\varepsilon}^{W_0}(1)\right)^2 \right],$$

car pour tout $i \in \{1, \cdots, d\}$, $E_{\theta_0}\left[\left(\widetilde{\varepsilon}^W(i)\right)^2\right] < \infty$. En posant :

$$Q = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \delta_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \delta_d \end{pmatrix}$$

et pour tout $X \in \mathbb{R}^d$, $\|X\| = \|ZX\|_Q$ on aura sous les hypothèses de la section 3.1 :

$$\lim_{n \longrightarrow \infty} [V_n(W) - V_n(W_0)] < 0$$

ce qui contredit le théorème 2. ■

On peut alors montrer que le processus associé à la vraisemblance est un contraste et cela même si le bruit n'est pas gaussien puisque :

$$K(\theta_0, \theta) = \frac{1}{2} \sum_{i=1}^{d} \left(\ln(\sigma_i^2) + \frac{1}{\sigma_i^2} E_{\theta_0}\left[\left(\widetilde{\varepsilon}^W(i)\right)^2\right] - E_{\theta_0}\left[\left(\widetilde{\varepsilon}^{W_0}(i)\right)^2\right] \right)$$

donc

$$K(\theta_0, \theta) \geq \frac{1}{2} \sum_{i=1}^{d} \left(\ln(\sigma_i^2) + \frac{1}{\sigma_i^2} E_{\theta_0}\left[\left(\widetilde{\varepsilon}^{W_0}(i)\right)^2\right] - E_{\theta_0}\left[\left(\widetilde{\varepsilon}^{W_0}(i)\right)^2\right] \right).$$

Mais, par construction du changement de base : $E_{\theta_0}\left[\left(\widetilde{\varepsilon}^{W_0}(i)\right)^2\right] = 1$, par conséquent :

$$K(\theta_0, \theta) \geq \frac{1}{2} \sum_{i=1}^{d} \frac{1}{\sigma_i^2} \left(\sigma_i^2 \ln(\sigma_i^2) + 1 - \sigma_i^2 \right) \geq 0,$$

car la fonction $x \longmapsto x \ln x + 1 - x$ pour $x \in \mathbb{R}^{+*}$ est positive, nulle seulement pour $x = 1$. L'hypothèse d'identifiabilité **H**-5, assure que $K(\theta_0, \theta) = 0$, seulement pour $\theta = \theta_0$.

3.2.2 Consistance forte

Dans le cadre des hypothèses (**H**), une condition suffisante assurant la consistance forte de $(\hat{\theta}_n)$ est (cf Guyon [6] section 3.4) :

Lemme 2 *Pour $\eta > 0$, posons*

$$\omega_n(\eta) = \sup \left\{ |U_n(\theta_\alpha) - U_n(\theta_\beta)| \, ; \|\theta_\alpha - \theta_\beta\| \leq \eta \right\}.$$

Si $\theta \longmapsto K(\theta_0, \theta)$ est continue et si il existe une suite (ϵ_k) réelle et décroissante vers 0 telle que, pour tout entier $k > 0$,

$$\lim_{n \longrightarrow \infty} P_{\theta_0}\left[\limsup_{n \to \infty} \left(\omega_n(\frac{1}{k}) > \epsilon_k \right) \right] = 0,$$

alors l'estimateur du minimum de contraste est fortement consistant.

On a :

$$l_\theta(y_1^{p+1}) = \ln\det\Gamma + \left(y_{p+1} - F_W(y_1^p)\right)^T \Gamma^{-1} \left(y_{p+1} - F_W(y_1^p)\right).$$

Posons $\Theta^* = \Theta \cap \mathbb{Q}^B$, où $\mathbb{Q}$ est l'ensemble des nombres rationnels. Cet ensemble est dense dans Θ et dénombrable. Soit la famille de fonctions (g_η) , $\eta > 0$.

$$\begin{array}{l} g_\eta(y_1^{p+1}) := \\ \sup\left\{\left|l_{\theta_\alpha}(y_1^{p+1}) - l_{\theta_\beta}(y_1^{p+1})\right| ; \|(\theta_\alpha) - (\theta_\beta)\| \leq \eta,\ (\theta_\alpha, \theta_\beta) \in \Theta^* \times \Theta^*\right\}. \end{array}$$

La fonction g_η est une variable aléatoire.

Pour tout $y_1^{p+1} \in \left(\mathbb{R}^d\right)^{p+1}$, par continuité de la fonction

$$(\theta_\alpha, \theta_\beta) \longmapsto \left|l_{\theta_\alpha}(y_1^{p+1}) - l_{\theta_\beta}(y_1^{p+1})\right|$$

et la densité de Θ^*dans Θ, on aura :

$$\begin{array}{l} \sup\left\{\left|l_{\theta_\alpha}(y_1^{p+1}) - l_{\theta_\beta}(y_1^{p+1})\right| ; \|(\theta_\alpha) - (\theta_\beta)\| \leq \eta,\ (\theta_\alpha, \theta_\beta) \in \Theta^* \times \Theta^*\right\} \\ = \sup\left\{\left|l_{\theta_\alpha}(y_1^{p+1}) - l_{\theta_\beta}(y_1^{p+1})\right| ; \|(\theta_\alpha) - (\theta_\beta)\| \leq \eta,\ (\theta_\alpha, \theta_\beta) \in \Theta \times \Theta\right\}. \end{array}$$

En notant $\lambda_{max} = \sup\left\{\rho(\Gamma);\ \Gamma^{-1} \in \Theta^{\Gamma^{-1}}\right\}$ (qui existe car Θ est compact), on a la majoration :

$$\sup_{\theta\in\Theta}\left|l_\theta(y_1^{p+1})\right| \leq h(y_1^{p+1}) \tag{10}$$

avec

$$h(y_1^{p+1}) := d \times \sup\left(\left|\ln(\lambda_{max})\right|, \left|\ln(\lambda_{min})\right|\right) + \frac{1}{\lambda_{min}}\left(\|y_{p+1}\|^a + \xi_1\|y_1\|^a + \cdots + \xi_p\|y_p\|^a + \kappa\right),$$

$\xi_1, \cdots, \xi_p$ et κ étant des constantes positives finies dont l'existence est assurée grâce à la condition **H**-1, alors $g_\eta \leq 2h$ avec h intégrable par rapport à la mesure invariante.

La continuité uniforme (car Θ est compact) de

$$(\theta_\alpha, \theta_\beta) \longmapsto \left|l_{\theta_\alpha}(y_1^{p+1}) - l_{\theta_\beta}(y_1^{p+1})\right|$$

implique, par convergence dominée, que

$$\theta \longmapsto K(\theta_0, \theta)$$

est continue.

De même on aura

$$\forall\,(y_1, \cdots, y_{p+1}) \in \left(\mathbb{R}^d\right)^{p+1},\ \lim_{\eta\longrightarrow 0} g_\eta(y_1, \cdots, y_{p+1}) = 0$$

donc par convergence dominée :

$$\lim_{\eta\longrightarrow 0} E_{\theta_0}\left[g_\eta(Y_1, \cdots, Y_{p+1})\right] = 0.$$

Finalement on aura P_{θ_0} p.s. :

$$\omega_n(\eta) \leq \frac{1}{n}\sum_{t=1}^{n} g_\eta(Y_t, \cdots, Y_{t-p}).$$

Il suffit donc de choisir pour $k \in \mathbb{N}^*$: $\epsilon_k = 2 \times E_{\theta_0}\left[g_{\frac{1}{k}}(Y_1, \cdots, Y_{p+1})\right]$ (ϵ_k décroît bien vers 0) pour que (en notant *i.s.* pour infiniment souvent)

$$\limsup_{n\to\infty}\left\{\omega_n\left(\frac{1}{k}\right) \geq \epsilon_k\right\} = \left\{\omega_n\left(\frac{1}{k}\right) \geq \epsilon_k \; i.s.\right\}$$

$$\subseteq \frac{1}{n}\sum_{t=1}^{n} g_{\frac{1}{k}}(Y_t, \cdots, Y_{t-p}) \geq \epsilon_k \; i.s.$$

sur $A := \left\{\frac{1}{n}\sum_{t=1}^{n} g_{\frac{1}{k}}(Y_t, \cdots, Y_{t-p}) \geq 2 \times E_{\theta_0}\left[g_{\frac{1}{k}}(Y_1, \cdots, Y_{p+1})\right] \; i.s.\right\}$, $\frac{1}{n}\sum_{t=1}^{n} g_{\frac{1}{k}}(Y_t, \cdots, Y_{t-p})$, ne peut converger vers $E_{\theta_0}\left[g_{\frac{1}{k}}(Y_1, \cdots, Y_{p+1})\right]$, A est donc un ensemble de mesure nulle. Cela montre le théorème suivant :

Théorème 3 *Dans le cadre des hypothèses (**H**), l'estimateur du minimum de contraste $U_n(\theta)$ est fortement consistant.*

3.3 Normalité asymptotique

Le Théorème de la Limite Centrale pour $\hat{\theta}_n$ nécessite des hypothèses supplémentaires sur les dérivées de $\theta \longmapsto U_n(\theta)$. Il faut s'assurer que les dérivées secondes, les carrés des dérivées premières sont bien intégrables, et avoir un contrôle sur la croissance des dérivées secondes.

3.3.1 Les dérivées d'ordre 1 et 2

On a

$$\frac{\partial U_n(\theta)}{\partial W_k} =$$

$$(\Gamma_{ij}^{-1})_{ind}^{T}\left(\frac{1}{n}\sum_{t=1}^{n}\frac{\partial F_W(y_{t-p}^{t-1})(i)}{\partial W_k} \times \left(y_t - F_W(y_{t-p}^{t-1})\right)(j) + \frac{\partial F_W(y_{t-p}^{t-1})(j)}{\partial W_k}\left(y_t - F_W(y_{t-p}^{t-1})\right)(i)\right)_{ind}$$

et

$$\frac{\partial U_n(\theta)}{\partial \Gamma_{ij}^{-1}} = (\Gamma ij) - \left(\frac{1}{n}\sum_{t=1}^{n}\left(y_t - F_W(y_{t-p}^{t-1})\right)(i)\left(y_t - F_W(y_{t-p}^{t-1})\right)(j)\right).$$

On en déduit les dérivées d'ordre 2 :

$$\frac{\partial^2 U_n(\theta)}{\partial W_k \partial W_l} =$$

$$(\Gamma_{ij}^{-1})_{ind}^{T}\left(\frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2 F_W(y_{t-p}^{t-1})(i)}{\partial W_k \partial W_l} \times \left(y_t - F_W(y_{t-p}^{t-1})\right)(j) - \frac{\partial F_W(y_{t-p}^{t-1})(i)}{\partial W_k}\frac{\partial F_W(y_{t-p}^{t-1})(j)}{\partial W_l}\right)_{ind}$$

$$+(\Gamma_{ij}^{-1})_{ind}^{T}\left(\frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2 F_W(y_{t-p}^{t-1})(j)}{\partial W_k \partial W_l} \times \left(y_t - F_W(y_{t-p}^{t-1})\right)(i) - \frac{\partial F_W(y_{t-p}^{t-1})(j)}{\partial W_k}\frac{\partial F_W(y_{t-p}^{t-1})(i)}{\partial W_l}\right)_{ind}$$

ainsi que

$$\frac{\partial^2 U_n(\theta)}{\partial \Gamma_{ij}^{-1}\partial \Gamma_{kl}^{-1}} = \frac{\partial(\Gamma ij)}{\partial \Gamma_{kl}^{-1}}$$

et

$$\frac{\partial^2 U_n(\theta)}{\partial W_k \partial \Gamma_{ij}^{-1}} = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial F_W(y_{t-p}^{t-1})(i)}{\partial W_k} \times \left(y_t - F_W(y_{t-p}^{t-1})\right)(j) + \frac{\partial F_W(y_{t-p}^{t-1})(j)}{\partial W_k}\left(y_t - F_W(y_{t-p}^{t-1})\right)(i).$$

La dérivée de $\theta \longmapsto U_n(\theta)$ (resp. $W \longmapsto F_W$) sera noté $\nabla U_n(\theta)$ (resp. ∇F_W). De même la dérivée seconde de $\theta \longmapsto U_n(\theta)$ (resp. $W \longmapsto F_W$) sera noté $HU_n(\theta)$ (resp. HF_W).

3.3.2 Hypothèses supplémentaires

On suppose qu'il existe un voisinage V de θ_0 tel que les hypothèses (**N**) suivantes soient vérifiées.

1. Le bruit a un moment d'ordre $2a$ avec $a \geq 2$ et par le théorème 1, $(Y_t)_{t\in\mathbb{N}^*}$ a alors un moment d'ordre $2a$.
2. Pour tout $y_1^p \in \left(\mathbb{R}^d\right)^p$, les dérivées d'ordre 3 de $W \longmapsto F_W(y_1^p)$, $W \in \Theta^W \cap V$ sont continues.
3. Pour tout $W \in \Theta^W \cap V$, les dérivées d'ordres 1, 2 et 3 de $W \longmapsto F_W$ sont μ_0-p.s. continues par rapport à y_1^p .
4. Pour tout $y_1^p \in \left(\mathbb{R}^d\right)^p$, $\forall k, j \,:\, 1 \leq k, j \leq D$

$$\left\| \frac{\partial F_{W_0}(y^{(p)})}{\partial W_k} \right\| \leq Cte \times \left(1 + \left\|y^{(p)}\right\|^{a/2}\right)$$

et

$$\left\| \frac{\partial^2 F_{W_0}(y^{(p)})}{\partial W_k \partial W_j} \right\| \leq Cte \times \left(1 + \left\|y^{(p)}\right\|^{a/2}\right).$$

5. Pour tout $y_1^p \in \left(\mathbb{R}^d\right)^p$, pour tout $W \in \Theta^W \cap V$,$\forall k, j, l \,:\, 1 \leq k, j, l \leq D$

$$\left\| \frac{\partial^3 F_W(y^{(p)})}{\partial W_k \partial W_j \partial W_l} \right\| \leq Cte \times \left(1 + \left\|y^{(p)}\right\|^{a/2}\right).$$

Remarque 3 *La condition **N**-5 implique qu'il existe un module de continuité ζ tel que $\forall W \in \Theta^W \cap V$:*

$$\left\|HF_W(y^{(p)}) - HF_{W_0}(y^{(p)})\right\| \leq \zeta(\|W - W_0\|)\left(1 + \left\|y^{(p)}\right\|^{a/2}\right).$$

Remarque 4 *Notons que la compacité de Θ, la remarque 3 et la condition **N**-4 impliquent qu'il existe une constante $\gamma > 0$ telle que $\forall W \in \Theta^W \cap V$:*

$$\left\|HF_W(y^{(p)})\right\| \leq \gamma(1 + \left\|y^{(p)}\right\|^{a/2}).$$

Remarque 5 *La remarque précédente nous permet de déduire un contrôle sur les accroissements de la dérivée première $\forall W \in \Theta^W \cap V$:*

$$\forall k, i \left\| \frac{\partial F_W(y^{(p)})}{\partial W_k} - \frac{\partial F_{W_0}(y^{(p)})}{\partial W_k} \right\| \leq \gamma \|\theta - \theta_0\| \times \left(1 + \left\|y^{(p)}\right\|^{a/2}\right).$$

Donc on aura aussi l'existence d'une constante finie γ telle que $\forall W \in \Theta^W \cap V$:

$$\forall k \,:\, \left\| \frac{\partial F_W(y^{(p)})}{\partial W_k} \right\| \leq Cte \times \left(1 + \left\|y^{(p)}\right\|^{a/2}\right).$$

Remarque 6 *De la même façon, le contrôle sur cette dérivée nous donne un contrôle sur les accroissements de la fonction de régression elle-même* $\forall W \in \Theta^W \cap V$ *:*

$$\forall k,i,\ \left\|F_W(y^{(p)}) - F_{W_0}(y^{(p)})\right\| \leq \gamma \left\|\theta - \theta_0\right\| \times \left(1 + \left\|y^{(p)}\right\|^{a/2}\right).$$

On aura alors

Proposition 2 *Sous les hypothèses (**H**) et (**N**), on a pour toute loi initiale de la chaîne* $\left(Y_{t-p+1}^t\right)_{t\in\mathbb{N}}$ *:*

$$HU_n(\theta_0) \xrightarrow{p.s.} I_0 \tag{11}$$

où I_0 *est une matrice symétrique.*

Preuve Il suffit de montrer que chaque terme de la Hessienne $\left(HU_n(\theta_0)\right)_{ij}$, $1 \leq i \leq j \leq B$ converge presque sûrement vers $(I_0)_{ij}$ et pour cela montrer qu'ils sont tous dominés par une fonction intégrable :

Terme de la forme $\frac{\partial^2 U_n(\theta_0)}{\partial W_k \partial W_l}$ **:** On a

$$\left\|\frac{\partial^2 U_n(\theta_0)}{\partial W_k \partial W_l}\right\| \leq$$

$$\left\|(\Gamma_{0_{ij}}^{-1})_{ind}^T \left(\frac{1}{n}\sum_{t=1}^n \frac{\partial^2 F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_k \partial W_l} \times \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(j) - \frac{\partial F_{W_0}(y_{t-p}^{t-1})(j)}{\partial W_k}\frac{\partial F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_l}\right)_{ind}\right\|$$

$$+ \left\|(\Gamma_{0_{ij}}^{-1})_{ind}^T \left(\frac{1}{n}\sum_{t=1}^n \frac{\partial^2 F_{W_0}(y_{t-p}^{t-1})(j)}{\partial W_k \partial W_l} \times \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(i) - \frac{\partial F_{W_0}(y_{t-p}^{t-1})(j)}{\partial W_k}\frac{\partial F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_l}\right)_{ind}\right\|$$

qui est majoré grâce à la condition **H**-4 par :

$$\frac{2}{\lambda_{min}}\sum_{ind}\left\|\frac{1}{n}\sum_{t=1}^n \frac{\partial^2 F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_k \partial W_l} \times \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(j) - \frac{\partial F_{W_0}(y_{t-p}^{t-1})(j)}{\partial W_k}\frac{\partial F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_l}\right\|$$

ce qui est majoré par

$$\frac{2}{\lambda_{min}}\sum_{ind}\frac{1}{n}\sum_{t=1}^n \left\|\frac{\partial^2 F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_k \partial W_l}\right\| \left\|\left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(j)\right\| + \left\|\left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(i)\right\| \left\|\frac{\partial F_{W_0}(y_{t-p}^{t-1})(j)}{\partial W_k}\right\|.$$

Mais grâce à la remarque 4 et à la condition (**H**)-1, sachant que $a \geq 2$, il existe une constante $\gamma > 0$ finie telle que pour tout $t \in \mathbb{N}^*$:

$$\left\|\frac{\partial^2 F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_k \partial W_l}\right\| \left\|\left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(j)\right\| < \gamma(1 + \left\|y_{t-p}^t\right\|^a).$$

Donc par la loi forte des grands nombres, $\frac{\partial^2 U_n(\theta_0)}{\partial W_k \partial W_l}$ converge presque sûrement vers le nombre fini :

$$E_{\theta_0}\left[(\Gamma_{0_{ij}}^{-1})_{ind}^T \left(\frac{\partial^2 F_W(Y_1^p)(i)}{\partial W_k \partial W_l} \times \left(Y_{p+1} - F_W(Y_1^P)\right)(j) - \frac{\partial^2 F_W(Y_1^p)(j)}{\partial W_k \partial W_l} \times \left(Y_{p+1} - F_W(Y_1^P)\right)(i)\right)_{ind}\right.$$

$$\left. + (\Gamma_{0_{ij}}^{-1})_{ind}^T \left(\frac{\partial^2 F_W(Y_1^p)(j)}{\partial W_k \partial W_l} \times \left(Y_{p+1} - F_W(Y_1^P)\right)(i) - \frac{\partial^2 F_W(Y_1^p)(i)}{\partial W_k \partial W_l} \times \left(Y_{p+1} - F_W(Y_1^P)\right)(j)\right)_{ind}\right].$$

Terme de la forme $\frac{\partial^2 U_n(\theta_0)}{\partial W_k \partial \Gamma_{ij}^{-1}}$ **:** On a

$$\left\| \frac{\partial^2 U_n(\theta_0)}{\partial W_k \partial \Gamma_{ij}^{-1}} \right\| =$$

$$\left\| \frac{1}{n} \sum_{t=1}^{n} \frac{\partial F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_k} \times \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(j) + \frac{\partial F_{W_0}(y_{t-p}^{t-1})(j)}{\partial W_k} \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(i) \right\|,$$

ce qui est majoré par

$$\frac{1}{n} \sum_{t=1}^{n} \left\| \frac{\partial F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_k} \right\| \left\| \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(j) \right\| + \left\| \frac{\partial F_{W_0}(y_{t-p}^{t-1})(j)}{\partial W_k} \right\| \left\| \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(i) \right\|$$

et grâce aux conditions (**H**)-1 et (**N**)-4, il existe une constante $\gamma > 0$ finie telle que pour tout $t \in \mathbb{N}^*$:

$$\left\| \frac{\partial F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_k} \right\| \left\| \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(j) \right\| + \left\| \frac{\partial F_{W_0}(y_{t-p}^{t-1})(j)}{\partial W_k} \right\| \left\| \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(i) \right\|$$

$$\leq \gamma(1 + \left\| y_{t-p}^{t} \right\|^a).$$

On peut encore appliquer la loi des grands nombres et $\frac{\partial^2 U_n(\theta_0)}{\partial W_k \partial \Gamma_{ij}^{-1}}$ converge presque sûrement vers :

$$E_{\theta_0} \left[\frac{\partial F_{W_0}(Y_1^p)(i)}{\partial W_k} \times \left(Y_{p+1} - F_{W_0}(Y_1^P)\right)(j) - \frac{\partial F_{W_0}(Y_1^p)(j)}{\partial W_k} \times \left(Y_{p+1} - F_W(Y_1^P)\right)(i) \right] < \infty.$$

Terme de la forme $\frac{\partial^2 U_n(W,\Gamma^{-1})}{\partial \Gamma_{ij}^{-1} \partial \Gamma_{kl}^{-1}} = \frac{\partial(\Gamma ij)}{\partial \Gamma_{kl}^{-1}}$ **:** Ce terme est constant en y, donc il est intégrable. ■

Démontrons maintenant la proposition qui établit la normalité asymptotique du processus $\nabla U_n(\theta_0)$:

Proposition 3 *Sous les hypothèses (**H**) et (**N**), on a pour toute loi initiale de la chaîne* $\left(Y_{t-p+1}^t\right)_{t\in\mathbb{N}}$

$$\sqrt{n} \nabla U_n(\theta_0) \xrightarrow{Loi} N(0, J_0) \tag{12}$$

où J_0 *est une matrice symétrique.*

Pour montrer cette proposition, posons

$$M_n = -n \nabla U_n(\theta_0) \tag{13}$$

M_n est une martingale, montrons qu'elle est de carré intégrable. Pour cela on va montrer que chaque terme de son crochet est intégrable.

Notation 6 *Notons* $\widetilde{U}_\theta(y_{t-p}^t)$ *la fonction :*

$$\widetilde{U}_\theta(y_{t-p}^t) = \ln \det \Gamma + \left(y_t - F_W(y_{t-p}^{t-1})\right)^T \Gamma^{-1} \left(y_t - F_W(y_{t-p}^{t-1})\right)$$

et $\nabla \widetilde{U}_\theta(y_{t-p}^t)$ *la dérivée de* $\widetilde{U}_\theta(y_{t-p}^t)$ *par rapport aux paramètres.*

on a : $M_t - M_{t-1} = \nabla \widetilde{U}_{\theta_0}(y_{t-p}^t)$,

Lemme 3 *Il existe une constante strictement positive* γ *telle que pour tout* $t \in \mathbb{N}^*$

$$\left\| \nabla \widetilde{U}_{\theta_0}(y_{t-p}^t)^T \nabla \widetilde{U}_{\theta_0}(y_{t-p}^t) \right\| < \gamma \left(1 + \left\| y_{t-p}^t \right\|^{2a}\right) \tag{14}$$

Preuve On va examiner chaque terme.

Terme de la forme $\frac{\partial \widetilde{U}_{\theta_0}(y_{t-p}^t)}{\partial W_k} \times \frac{\partial \widetilde{U}_{\theta_0}(y_{t-p}^t)}{\partial W_l}$ **:** Il vaut

$$(\Gamma_{0_{ij}}^{-1})_{ind}^T \left(\frac{\partial F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_k} \times \left(y_t - F_W(y_{t-p}^{t-1})\right)(j) + \frac{\partial F_{W_0}(y_{t-p}^{t-1})(j)}{\partial W_k} \left(y_t - F_W(y_{t-p}^{t-1})\right)(i) \right)_{ind} \times$$
$$(\Gamma_{0_{ij}}^{-1})_{ind}^T \left(\frac{\partial F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_l} \times \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(j) + \frac{\partial F_{W_0}(y_{t-p}^{t-1})(j)}{\partial W_l} \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(i) \right)_{ind}$$

La norme de ce terme est majorée par :

$$\left(\frac{1}{\lambda_{min}} \sum_{ind} \left\| \frac{\partial^2 F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_k \partial W_l} \right\| \left\| \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(j) \right\| + \left\| \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(i) \right\| \left\| \frac{\partial F_{W_0}(y_{t-p}^{t-1})(j)}{\partial W_k} \right\| \right)^2 .$$

Mais par la remarque 4 :

$$\left\| \frac{\partial^2 F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_k \partial W_l} \right\| \left\| \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(j) \right\| < \gamma(1 + \left\| y_{t-p}^t \right\|^a)$$

donc on aura

$$\left(\frac{1}{\lambda_{min}} \sum_{ind} \left\| \frac{\partial^2 F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_k \partial W_l} \right\| \left\| \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(j) \right\| + \left\| \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(i) \right\| \left\| \frac{\partial F_{W_0}(y_{t-p}^{t-1})(j)}{\partial W_k} \right\| \right)^2$$
$$\leq \left[\frac{2}{\lambda_{min}} \sum_{ind} \gamma(1 + \left\| y_{t-p}^t \right\|^a) \right]^2$$

ce qui assure qu'il existe une constante γ_1 positive finie telle que :

$$\left\| \frac{\partial \widetilde{U}_{\theta_0}(y_{t-p}^t)}{\partial W_k} \times \frac{\partial \widetilde{U}_{\theta_0}(y_{t-p}^t)}{\partial W_l} \right\| \leq \gamma_1 \left(1 + \left\| y_{t-p}^t \right\|^{2a}\right).$$

Terme de la forme $\frac{\partial \widetilde{U}_{\theta_0}(y_{t-p}^t)}{\partial W_k} \times \frac{\partial \widetilde{U}_{\theta_0}(y_{t-p}^t)}{\partial \Gamma_{ij}^{-1}}$ **:** Il vaut

$$(\Gamma_{0_{ij}}^{-1})_{ind}^T \left(\frac{\partial F_{W_0}(y_{t-p}^{t-1})(i)}{\partial W_k} \times \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(j) + \frac{\partial F_{W_0}(y_{t-p}^{t-1})(j)}{\partial W_k} \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(i) \right)_{ind} \times$$
$$\left((\Gamma_{0_{ij}}) - \left(\left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(i) \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(j) \right) \right).$$

Le module de ce terme sera alors majoré par

$$\frac{1}{\lambda_{min}} \sum_{ind} \left\| \frac{\partial F_{W_0}(y_{t-p}^{t-1})}{\partial W_k} \right\| \left\| \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right) \right\|^3 + \frac{2\lambda_{max}}{\lambda_{min}} \sum_{ind} \left\| \frac{\partial F_{W_0}(y_{t-p}^{t-1})}{\partial W_k} \right\| \left\| \left(y_t - F_{W_0}(y_{t-p}^{t-1})\right) \right\|$$

et grâce aux conditions (**N**)-4, (**N**)-1 et (**H**)-1, on en déduit l'existence d'un constante $\gamma_2 > 0$ telle que

$$\left\| \frac{\partial \widetilde{U}_{\theta_0}(y_{t-p}^t)}{\partial W_k} \times \frac{\partial \widetilde{U}_{\theta_0}(y_{t-p}^t)}{\partial \Gamma_{ij}^{-1}} \right\| \leq \gamma_2 \left(1 + \left\| y_{t-p}^t \right\|^{2a}\right).$$

Terme de la forme $\frac{\partial \widetilde{U}_{\theta_0}(y_{t-p}^t)}{\partial \Gamma_{ij}^{-1}} \times \frac{\partial \widetilde{U}_{\theta_0}(y_{t-p}^t)}{\partial \Gamma_{kl}^{-1}}$ **:** Il vaut

$$\begin{array}{l}\left((\Gamma_{0_{ij}}) - \left(\left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(i)\left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(j)\right)\right) \times \\ \left((\Gamma_{0_{kl}}) - \left(\left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(k)\left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)(l)\right)\right)\end{array}$$

dont le module est majoré par

$$2\lambda_{max}\left\|\left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)\right\|^2 + \left\|\left(y_t - F_{W_0}(y_{t-p}^{t-1})\right)\right\|^4.$$

Grâce aux conditions (**H**)-1 et (**N**)-1, on en déduit l'existence d'une constante $\gamma_3 > 0$ telle que

$$\left\|\frac{\partial \widetilde{U}_{\theta_0}(y_{t-p}^t)}{\partial \Gamma_{ij}^{-1}} \times \frac{\partial \widetilde{U}_{\theta_0}(y_{t-p}^t)}{\partial \Gamma_{kl}^{-1}}\right\| \leq \gamma_3\left(1 + \left\|y_{t-p}^t\right\|^{2a}\right).$$

Maintenant, en prenant $\gamma_0 = \sup\{\gamma_1, \gamma_2, \gamma_3\}$, on a bien

$$\left\|\nabla \widetilde{U}_{\theta_0}(y_{t-p}^t)\nabla \widetilde{U}_{\theta_0}(y_{t-p}^t)^T\right\| < B \times \gamma_0\left(1 + \left\|y_{t-p}^t\right\|^{2a}\right)$$

■

La proposition 3 sera donc prouvée si (M_n) satisfait la condition de Lindeberg suivante (Duflo [4]) :

Proposition 4 *En notant $\mathcal{F}_t$ la tribu engendrée par Y_{-p+1}^t, pour tout $\epsilon > 0$ on a :*

$$L_n := \frac{1}{n}\sum_{t=1}^{n} E\left[\left\|\nabla \widetilde{U}_{\theta_0}(y_{t-p}^t)\right\|^2 \mathbb{I}_{\left\{\left\|\nabla \widetilde{U}_{\theta_0}(y_{t-p}^t)\right\| \geq \epsilon\sqrt{n}\right\}} | \mathcal{F}_{t-1}\right] \xrightarrow{P_{\theta_0}} 0.$$

Preuve Soit $A > 0$ et :

$$F_n(A) := \frac{1}{n}\sum_{t=1}^{n} E\left[\left\|\nabla \widetilde{U}_{\theta_0}(y_{t-p}^t)\right\|^2 \mathbb{I}_{\left\{\left\|\nabla \widetilde{U}_{\theta_0}(y_{t-p}^t)\right\| \geq \epsilon A\right\}} | \mathcal{F}_{t-1}\right] := \frac{1}{n}\sum_{t=1}^{n} h\left(y_{t-p}^{t-1}, A\right)$$

avec

$$h\left(y_{t-p}^{t-1}, A\right) = E\left[\nabla \widetilde{U}_{\theta_0}(y_{t-p}^t)^T \nabla \widetilde{U}_{\theta_0}(y_{t-p}^t) \mathbb{I}_{\left\{\left\|\nabla \widetilde{U}_{\theta_0}(y_{t-p}^t)\right\| \geq \epsilon A\right\}} | \mathcal{F}_{t-1}\right].$$

D'après le lemme 3, il existe une constante γ_0 telle que

$$h(y_{t-p}^{t-1}, A) \leq \gamma_0\left(1 + \left\|y_{t-p}^{t-1}\right\|^{2a}\right).$$

Par la loi forte des grands nombres, on a :

$$F_n(A) \xrightarrow{p.s.} \Phi(A) = \int_{\left(\mathbb{R}^d\right)^p} h(y_1^p, A)\mu_0(dy_1^p).$$

Φ est décroissante et positive. Le théorème de convergence dominée montre que, quand A tend vers ∞, $\Phi(A)$ tend vers 0 . Enfin, pour A fixé, on a, si n est assez grand : $\epsilon\sqrt{n} > A$, et $L_n = F_n\left(\epsilon\sqrt{n}\right) \leq F_n(A)$, donc p.s. $\limsup_n L_n \leq \Phi(A)$. Finalement, en faisant tendre $A \longrightarrow \infty$, on obtient p.s. :

$$\lim_{n \longrightarrow \infty} L_n = 0$$

■

On peut maintenant établir le théorème de normalité asymptotique :

Théorème 4 *On suppose satisfaites les hypothèses (**H**) et (**N**) et que le bruit a un moment d'ordre $2a$ avec $a \geq 2$. Alors pour toute loi initiale de la chaîne vectorisée $\left(Y_{t-p+1}^t\right)_{t>0}$:*

$$\sqrt{n} I_0\left[(\hat{\theta}_n) - (\theta_0)\right] \xrightarrow{Loi} N(0, J_0).$$

Preuve Soit V un voisinage de θ_0. Puisque $\hat{\theta}_n \longrightarrow \theta_0$ p.s., il existe $n_0(\omega)$ tel que $\hat{\theta}_n \in V$, pour tout $n \geq n_0(\omega)$. Par la formule de Taylor (avec reste intégral) :

$$0 = \nabla U_n\left(\hat{\theta}_n\right) = \nabla U_n(\theta_0) + \Delta_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \tag{15}$$

avec

$$\Delta_n(\hat{\theta}_n) = \int_0^1 HU_n\left[\hat{\theta}_n + u\left(\hat{\theta}_n - \theta_0\right)\right] du.$$

Supposons vérifiée la condition suivante (voir lemme 4) :

$$\Delta_n(\hat{\theta}_n) - HU_n(\theta_0) \xrightarrow{P_{\theta_0}} 0.$$

D'après la proposition 2 qui assure : $HU_n(\theta_0) \xrightarrow{p.s.} I_0$, le théorème est prouvé, puisque alors :

$$\sqrt{n}\Delta_n(\theta_n)\left(\hat{\theta}_n - \theta_0\right) = -\sqrt{n}\nabla U_n(\theta_0)$$

assure que

$$\lim_{n\longrightarrow\infty} \sqrt{n} I_0\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{en\ loi} N(0, J_0).$$

Lemme 4 *Dans le cadre du théorème 4, on a :*

$$\Delta_n(\hat{\theta}_n) - HU_n(\theta_0) \xrightarrow{p.s.} 0$$

Soit par la proposition 2, ($HU_n(\theta_0) \xrightarrow{p.s.} I_0$) :

$$\Delta_n(\hat{\theta}_n) \xrightarrow{p.s.} I_0$$

Nous allons d'abord d'établir le lemme :

Lemme 5 *Il existe un module de continuité β tel que pour tout $\theta \in V$,*

$$\left\| H\widetilde{U}_\theta(y_1^{p+1}) - H\widetilde{U}_{\theta_0}(y_1^{p+1})) \right\| \leq \beta(\|\theta - \theta_0\|)(1 + \left\|y_1^{p+1}\right\|^a)$$

ce qui implique l'existence d'une constante γ telle que pour tout $\theta \in V$:

$$\left\| H\widetilde{U}_\theta(y_1^{p+1}) \right\| \leq \gamma(1 + \left\|y_1^{p+1}\right\|^a).$$

Preuve Il suffit de vérifier que chaque terme de la dérivée d'ordre 3 par rapport au paramètre de $\widetilde{U}_\theta(y_1^{p+1})$ est dominé par une expression du type $\gamma(1 + \left\|y_1^{p+1}\right\|^a)$, pour $\Theta \in V$.

Terme de la forme $\frac{\partial^3 \widetilde{U}_\theta(y_1^{p+1})}{\partial W_k \partial W_l \partial W_m}$: On a

$$\left\| \frac{\partial^3 \widetilde{U}_\theta(y_1^{p+1})}{\partial W_k \partial W_l \partial W_m} \right\| =$$

$$\left\| (\Gamma_{ij}^{-1})_{ind}^T \left(\frac{\partial^3 F_W(y_1^p)(i)}{\partial W_k \partial W_l \partial W_m} \times (y_{p+1} - F_W(y_1^p))(j) - \frac{\partial^2 F_W(y_1^p)(i)}{\partial W_k \partial W_l} \frac{\partial F_W(y_1^p)(j)}{\partial W_m} \right)_{ind} \right.$$

$$\left. -(\Gamma_{ij}^{-1})_{ind}^T \left(\frac{\partial^2 F_W(y_1^p)(j)}{\partial W_k \partial W_m} \times \frac{\partial F_W(y_1^p)(i)}{\partial W_l} + \frac{\partial F_W(y_1^p)(j)}{\partial W_k} \frac{\partial^2 F_W(y_1^p)(i)}{\partial W_l \partial W_m} \right)_{ind} \right.$$

$$+(\Gamma_{ij}^{-1})_{ind}^{T}\left(\frac{\partial^3 F_W(y_1^p)(j)}{\partial W_k\partial W_l\partial W_m}\times\left(y_{p+1}-F_W(y_1^p)\right)(i)-\frac{\partial^2 F_W(y_1^p)(j)}{\partial W_k\partial W_l}\frac{\partial F_W(y_1^p)(i)}{\partial W_m}\right)_{ind}$$
$$-(\Gamma_{ij}^{-1})_{ind}^{T}\left(\frac{\partial^2 F_W(y_1^p)(i)}{\partial W_k\partial W_m}\times\frac{\partial F_W(y_1^p)(j)}{\partial W_l}+\frac{\partial F_W(y_1^p)(i)}{\partial W_k}\frac{\partial^2 F_W(y_1^p)(j)}{\partial W_l\partial W_m}\right)_{ind}\Bigg\|,$$

ce qui est majoré par

$$\tfrac{2}{\lambda_{min}}\sum_{ind}\left\|\tfrac{\partial^3 F_W(y_1^p)(i)}{\partial W_k\partial W_l\partial W_m}\times\left(y_{p+1}-F_W(y_1^p)\right)(j)+\tfrac{\partial^2 F_W(y_1^p)(i)}{\partial W_k\partial W_l}\tfrac{\partial F_W(y_1^p)(j)}{\partial W_m}\right\|$$
$$+\tfrac{2}{\lambda_{min}}\sum_{ind}\left\|\tfrac{\partial^2 F_W(y_1^p)(j)}{\partial W_k\partial W_m}\times\tfrac{\partial F_W(y_1^p)}{\partial W_l}+\tfrac{\partial F_W(y_1^p)(j)}{\partial W_k}\tfrac{\partial^2 F_W(y_1^p)(i)}{\partial W_l\partial W_m}\right\|.$$

En tenant compte de la condition (**N**)-5, des remarques 4 et 5, ainsi que de la condition (**H**)-1, il existe une constante finie γ telle que :

$$\left\|\frac{\partial^3\widetilde{U}_\theta(y_1^{p+1})}{\partial W_k\partial W_l\partial W_m}\right\|\leq\gamma(1+\left\|y_1^{p+1}\right\|^a).$$

Terme de la forme $\frac{\partial^3\widetilde{U}_\theta(y_1^{p+1})}{\partial W_k\partial W_l\partial\Gamma_{ij}^{-1}}$: On a

$$\left\|\frac{\partial^3\widetilde{U}_\theta(y_1^{p+1})}{\partial W_k\partial W_l\partial\Gamma_{ij}^{-1}}\right\|=$$

$$\Bigg\|\frac{\partial^2 F_W(y_{t-p}^{t-1})(i)}{\partial W_k\partial W_l}\times\left(y_t-F_W(y_{t-p}^{t-1})\right)(j)-\frac{\partial F_W(y_{t-p}^{t-1})(j)}{\partial W_k}\frac{\partial F_W(y_{t-p}^{t-1})(i)}{\partial W_l}$$
$$\frac{\partial^2 F_W(y_{t-p}^{t-1})(j)}{\partial W_k\partial W_l}\times\left(y_t-F_W(y_{t-p}^{t-1})\right)(i)-\frac{\partial F_W(y_{t-p}^{t-1})(i)}{\partial W_k}\frac{\partial F_W(y_{t-p}^{t-1})(j)}{\partial W_l}\Bigg\|.$$

Grâce aux remarques 4 et 5, ainsi qu'à la condition (**H**)-1, il existe une constante finie γ telle que :

$$\left\|\frac{\partial^3\widetilde{U}_\theta(y_1^{p+1})}{\partial W_k\partial W_l\partial\Gamma_{ij}^{-1}}\right\|\leq\gamma(1+\left\|y_1^{p+1}\right\|^a).$$

Termes de la forme $\frac{\partial^3\widetilde{U}_\theta(y_1^{p+1})}{\partial W_k\partial\Gamma_{cd}^{-1}\partial\Gamma_{ij}^{-1}}$ et $\frac{\partial^3\widetilde{U}_\theta(y_1^{p+1})}{\partial\Gamma_{kl}\partial\Gamma_{cd}^{-1}\partial\Gamma_{ij}^{-1}}$: On a

$$\frac{\partial^3\widetilde{U}_\theta(y_1^{p+1})}{\partial W_k\partial\Gamma_{cd}^{-1}\partial\Gamma_{ij}^{-1}}=0$$

et

$$\frac{\partial^3\widetilde{U}_\theta(y_1^{p+1})}{\partial\Gamma_{kl}^{-1}\partial\Gamma_{cd}^{-1}\partial\Gamma_{ij}^{-1}}=\frac{\partial^2(\Gamma_{kl})}{\partial\Gamma_{cd}^{-1}\partial\Gamma_{ij}^{-1}}$$

qui sont constants en y, donc majorés par une expression du type $\gamma(1+\left\|y_1^{p+1}\right\|^a)$ pour $\Theta\in V$ et γ fini.

Les majorations de la dérivée troisième du contraste implique immédiatement l'existence du module de continuité du lemme 5 ■

Preuve du lemme 4 : Pour $\theta \in V$, considérons :

$$n \left\| HU_n(\theta) - HU_n(\theta_0) \right\|$$

Grâce au lemme 5, on a l'inégalité pour tout $\theta \in V$:

$$n \left\| HU_n(\theta) - HU_n(\theta_0) \right\| \leq 2 \sum_{t=1}^{n} \beta(\|\theta - \theta_0\|)(1 + \left\| y_{t-p}^t \right\|^a)$$

donc

$$\left\| \Delta_n(\hat{\theta}_n) - HU_n(\theta_0) \right\| = \left\| \int_0^1 \left\{ HU_n \left[\hat{\theta}_n + u \left(\hat{\theta}_n - \theta_0 \right) \right] - HU_n(\theta_0) \right\} du \right\|$$

est majoré par

$$\beta(\|\theta - \theta_0\|) \frac{2}{n} \sum_{t=1}^{n} (1 + \left\| y_{t-p}^t \right\|^a).$$

Par la loi forte des grands nombres $\frac{2}{n} \sum_{t=1}^{n} (1 + \left\| y_{t-p}^t \right\|^a)$ converge p.s. et comme $\hat{\theta}_n \longrightarrow \theta_0$ p.s., on en déduit la convergence p.s. vers 0 de $\Delta_n(\hat{\theta}_n) - HU_n(\theta_0)$ ■

Remarque 7 *Dans le cas gaussien on peut préciser la forme des matrices I_0 et J_0. En effet, dans ce cas, la martingale (M_n) (cf égalité (13)) est la dérivée de l'opposée de la log-vraisemblance, elle est de carré intégrable et le processus croissant qui lui est associé est :*

$$J_n(\theta_0) := \sum_{t=1}^{n} \left(\frac{\left(\frac{\partial}{\partial \theta_k} L_{\theta_0}\left(y_{t-p}^t \right) \right)_{1 \leq k \leq B}}{L_{\theta_0}\left(y_{t-p}^t \right)} \right) \left(\frac{\left(\frac{\partial}{\partial \theta_k} L_{\theta_0}\left(y_{t-p}^t \right) \right)_{1 \leq k \leq B}}{L_{\theta_0}\left(y_{t-p}^t \right)} \right)^T$$

De plus la dérivée seconde de l'opposée de la log-vraisemblance est

$$Z_n(\theta_0) := \sum_{t=1}^{n} - \frac{\left(\frac{\partial}{\partial \theta_l} \left(\frac{\partial}{\partial \theta_k} L_{\theta_0}\left(y_{t-p}^t \right) \right)_{1 \leq k \leq B} \right)_{1 \leq l \leq B}}{L_{\theta_0}\left(y_{t-p}^t \right)}$$

$$+ \left(\frac{\left(\frac{\partial}{\partial \theta_k} L_{\theta_0}\left(y_{t-p}^t \right) \right)_{1 \leq k \leq B}}{L_{\theta_0}\left(y_{t-p}^t \right)} \right) \left(\frac{\left(\frac{\partial}{\partial \theta_k} L_{\theta_0}\left(y_{t-p}^t \right) \right)_{1 \leq k \leq B}}{L_{\theta_0}\left(y_{t-p}^t \right)} \right)^T$$

*Mais, comme sous les hypothèses (**N**), on peut échanger espérance et dérivation, on a :*

$$E_{\theta_0} \left[\frac{\left(\frac{\partial}{\partial \theta_l} \left(\frac{\partial}{\partial \theta_k} L_{\theta_0}\left(y_{t-p}^t \right) \right)_{1 \leq k \leq B} \right)_{1 \leq l \leq B}}{L_{\theta_0}\left(y_{t-p}^t \right)} \right] = \int_{\left(\mathbb{R}^d \right)^{p+1}} \frac{\partial^2}{\partial \theta_l \theta_k} L_{\theta_0}\left(y_1^{p+1} \right) dy_1^{p+1} = 0$$

donc

$$\lim_{n \to \infty} \frac{1}{n} Z_n(\theta_0) = I_0 = \lim \frac{1}{n} J_n(\theta_0) = J_0$$

où $I_0 = J_0$ est la matrice d'information de Fisher du modèle.

Donc, dans le cas gaussien, $\hat{\theta}_n$ est l'estimateur du maximum de vraisemblance, il est fortement consistant et asymptotiquement efficace.

3.4 Vitesse et loi du logarithme itéré

Dans cette section, pour une matrice réelle et symétrique A, $\lambda_{max}A$ (resp. $\lambda_{min}A$) désignera la plus grande (resp. la plus petite) valeur propre de A. Montrons le théorème suivant :

Théorème 5 *Sous les hypothèses (**H**) et (**N**), si le bruit a un moment d'ordre $> 2a$, $a \geq 2$ et si les matrices I_0 et J_0 sont inversibles, on a presque sûrement :*

$$\limsup_n \sqrt{\frac{n}{2\ln\ln n}}\,\|DU_n(\theta_0)\| \leq \sqrt{\lambda_{max}J_0} \tag{16}$$

$$\limsup_n \sqrt{\frac{n}{2\ln\ln n}}\,\left\|\hat{\theta}_n - \theta_0\right\| \leq \sqrt{\frac{\lambda_{max}J_0}{\lambda_{min}I_0}}. \tag{17}$$

Preuve C'est une adaptation de la preuve de [8].

Pour u, un vecteur de $\mathbb{R}^B$ notons

$$\widetilde{M}_n := \langle M_n, u\rangle = \sum_{t=1}^{n} \left\langle \nabla\widetilde{U}_{\theta_0}(y_{t-p}^t), u \right\rangle$$

C'est une martingale de puissance $2+2\alpha$ intégrable pour tout $\alpha \in]0, \frac{a}{2}-1]$. Notons :

$$T_t = E\left[\left|\widetilde{M}_{t+1} - \widetilde{M}_t\right|^{2+2\alpha} |F_t\right]^{1/(2+2\alpha)}$$

et

$$\tau_n = \sum_{t=1}^{n} T_t^2 = u^T \langle M\rangle_n u.$$

En vertu de (14), $\frac{\tau_n}{n} \longrightarrow u^T J_0 u$ presque sûrement, qui est strictement positif car J_0 est supposée inversible. On aura alors $\tau_n \longrightarrow \infty$ presque sûrement. La loi du logarithme itéré pour une martingale de puissance $2+2\alpha$ intégrable (Duflo[5], Corollaire 6) assure que presque sûrement :

$$\limsup_n \frac{\left|\widetilde{M}_n\right|}{\sqrt{2\tau_{n-1}\ln\ln\tau_{n-1}}} \leq 1$$

si la série $\sum \left(\frac{T_n^2}{\tau_n}\right)^{1+\alpha}$ est p.s. convergente.

Posons $s_n := T_1^{2+2\alpha} + \cdots + T_n^{2+2\alpha}$. Pour $\alpha < a/2 - 1$, grâce au lemme 3, on a la loi forte des grands nombres pour $(T_n^{2+2\alpha})$, donc $\frac{s_n}{n} \longrightarrow \gamma \geq 0$ presque sûrement. Par ailleurs $\left(\frac{T_n^2}{\tau_n}\right)^{1+\alpha} \sim Cte \times \frac{T^{2+2\alpha}}{n^{1+\alpha}}$ et par la transformation d'Abel

$$\sum_{t=1}^{n} \frac{T_t^{2+2\alpha}}{t^{1+\alpha}} = \frac{s_n}{n^{1+\alpha}} + \sum_{t=1}^{n-1}\left[\frac{1}{t^{1+\alpha}} - \frac{1}{(t+1)^{1+\alpha}}\right] - s_1.$$

Puisque $\frac{s_n}{n^{1+\alpha}} \longrightarrow 0$ p.s. et

$$\frac{1}{t^{1+\alpha}} - \frac{1}{(t+1)^{1+\alpha}} \sim \frac{1+\alpha}{t^{2+\alpha}},$$

la série $\sum \frac{T^{2+2\alpha}}{n^{1+\alpha}}$ est presque sûrement convergente et il en est de même pour $\sum \left(\frac{T_n^2}{\tau_n}\right)^{1+\alpha}$.

Comme $2\tau_{n-1}\ln\ln\tau_{n-1} \sim 2nu^T J_0 u \ln\ln n$ on obtient presque sûrement :

$$\limsup_n \sqrt{\frac{n}{2\ln\ln n}}\,|\langle DU_n(\theta_0), u\rangle| \leq \sqrt{u^T J_0 u}$$

d'où la L.L.I.

$$\limsup_n \sqrt{\frac{n}{2\ln\ln n}}\,\|DU_n(\theta_0)\| \leq \sqrt{\lambda_{max} J_0}.$$

Pour la seconde L.L.I.

$$\limsup_n \sqrt{\frac{n}{2\ln\ln n}}\,\left\|\hat{\theta}_n - \theta_0\right\| \leq \sqrt{\frac{\lambda_{max} J_0}{\lambda_{min} I_0}},$$

elle se déduit de la première grâce au développement de Taylor (15) et le lemme 4 ■

3.5 Identification presque sûre

Suivant la présentation de Guyon [6], on suppose que l'espace des paramètres $\Theta \subset \mathbb{R}^M$ correspond au modèle majorant. Soit $\mathbb{O}$ une famille finie de sous-espaces de $\mathbb{R}^M$, $\delta \in \mathbb{O}$ l'élément générique de $\mathbb{O}$, $|\delta|$ sa dimension et $\Theta_\delta := \Theta \cap \delta$ le sous-espace (sous-modèle) paramétrique associé. On suppose que la vraie valeur est $\theta_{0,\delta_0} \in \Theta_{\delta_0}$, $\delta_0 \in \mathbb{O}$ étant le sous-espace minimal associé à θ_{0,δ_0}. Soit $(c\,(n))$ une suite positive. Au vu de la réalisation $(Y_t)_{-p<t\leq n}$, on utilise comme fonction de décision le contraste pénalisé à la vitesse $c\,(n)$ par la dimension du modèle :

$$CP_{n,\delta}\,(\theta) := U_n(\theta_\delta) + \frac{c(n)}{n}\,|\delta| \tag{18}$$

pour $\delta \in \mathbb{O}$ et $\theta_\delta \in \Theta_\delta$.

Notons : $\overline{CP}_{n,\delta} = \overline{U}_{n,\delta} + \frac{c(n)}{n}\,|\delta|$ avec $\overline{U}_{n,\delta} = U_n(\hat{\theta}_{n,\delta})$ et $\hat{\theta}_{n,\delta} = \arg\min_{\theta_\delta \in \Theta_\delta} U_n(\theta_\delta)$. On choisira $\hat{\delta}_n = \arg\min_{\delta\in\mathbb{F}} \overline{CP}_{n,\delta}$, qui répond au principe de parcimonie d'Akaiké avec la vitesse $c(n)$.

Appliquant les résultats de (Senoussi [10], Guyon [6]), nous avons le résultat suivant d'identification presque sûre du vrai modèle δ_0.

Théorème 6 *On se place dans le cadre du théorème 5. Si la vitesse de pénalisation $c(n)$ est telle que :*

$$\lim_n \frac{c(n)}{n} = 0$$

et

$$\liminf_n \frac{c(n)}{2\ln\ln n} > \frac{\lambda_{max} J_0}{2\lambda_{min} I_0}$$

alors, le couple $\left(\hat{\delta}_n, \hat{\theta}_{n,\delta_n}\right)$ converge P_{θ_0}-p.s. vers la vraie valeur $(\delta_0, \theta_{0,\delta_0})$.

Preuve Il suffit d'appliquer le théorème (3.4.8) de (Guyon [6]) dont les conditions d'application se vérifient immédiatement ici grâce au théorème 5.

4 Application au perceptron multicouches (MLP)

Un MLP est une fonction paramétrique non-linéaire. Par exemple la figure 2 représente la fonction :

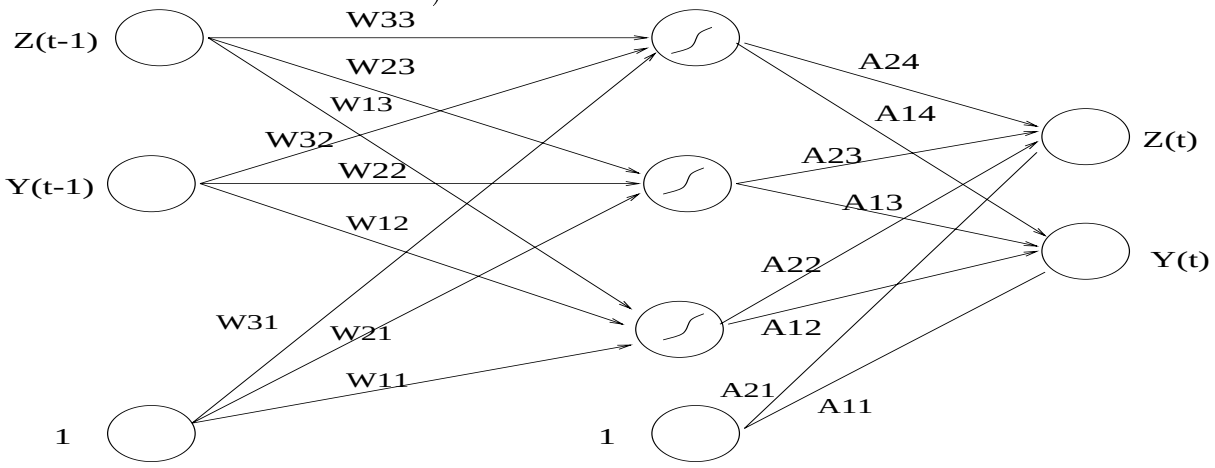
$$Z_t = A_{24} f\left(W_{33} Z_{t-1} + W_{32} Y_{t-1} + W_{31}\right) + A_{23} f(W_{23} Z_{t-1} + W_{22} Y_{t-1} + W_{21}) \\ + A_{22} f(W_{13} Z_{t-1} + W_{12} Y_{t-1} + W_{11}) + A_{21}$$

et

$$Y_t = A_{14} f\left(W_{33} Z_{t-1} + W_{32} Y_{t-1} + W_{31}\right) + A_{13} f(W_{23} Z_{t-1} + W_{22} Y_{t-1} + W_{21}) \\ + A_{12} f(W_{13} Z_{t-1} + W_{12} Y_{t-1} + W_{11}) + A_{11}$$

où "f" est la fonction d'activation (généralement "tanh" ou une fonction sigmoïde).

FIG. 2 – Fonction MLP (On note A_{ij} les poids de sorties pour les différentier des poids W_{ij} entre l'entrée et la couche cachée.)



4.1 Calcul de la dérivée du contraste

Pour obtenir les paramètres (poids du MLP) $W \in \mathbb{R}^D$ qui maximisent la vraisemblance des observations il faut minimiser la fonction $W \to \frac{1}{2}\ln(\det(\Gamma))$. On rappelle que par l'équation 8, le gradient de cette fonction s'écrit :

$$\left(\frac{\partial}{\partial W_k}(\frac{1}{2}\ln(\det(\Gamma)))\right)_{1\leq k\leq D}$$

avec

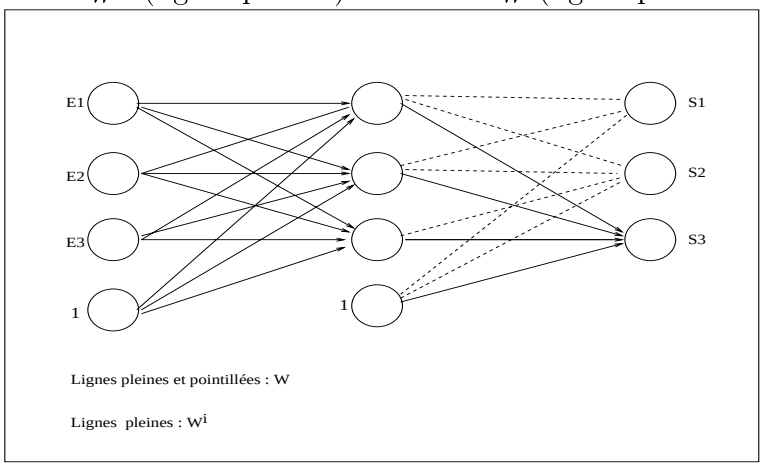
$$\frac{\partial}{\partial W_k}(\frac{1}{2}\ln(\det(\Gamma))) = \left(\Gamma_{ij}^{-1}\right)_{ind}^T \left(\frac{\Gamma_{ij}}{\partial W_k}\right)_{ind}$$

et

$$\frac{\partial \Gamma_{ij}}{\partial W_k} = \sum_{t=1}^{n} \left[-\frac{\partial F_W(y_{t-p}^{t-1})(i)}{\partial W_k} \times \left(y_t - F_W(y_{t-p}^{t-1})\right)(j) - \frac{\partial F_W(y_{t-p}^{t-1})(j)}{\partial W_k} \left(y_t - F_W(y_{t-p}^{t-1})\right)(i) \right]$$

Calcul de $\frac{\partial F_W(y_{t-p}^{t-1})(i)}{\partial W_k}$: Soit un MLP à sortie vectorielle F_W, considérons le MLP extrait F_{W^i} qui a les mêmes poids que F_W avant la dernière unité cachée, mais qui ne garde que les

FIG. 3 – MLP extrait $F_W 3$ (lignes pleines) du MLP F_W (lignes pleines et lignes pointillés)



poids pointant sur la sortie i de F_W . La figure 3 décrit un MLP avec trois entrées et trois sorties et $i = 3$ (ce qui correspond à la troisième sortie).

Pour calculer $\frac{\partial F_W(y_{t-p}^{t-1})(i)}{\partial W_k}$, il suffit de calculer par rétro-propagation classique la dérivée [2] de la fonction représentée par le MLP extrait F_{W^i} : $\frac{\partial F_{W^i}(y_{t-p}^{t-1})}{\partial W_k^i}$. Finalement $\frac{\partial F_W(y_{t-p}^{t-1})(i)}{\partial W_k}$sera égale à $\frac{\partial F_{W^i}(y_{t-p}^{t-1})}{\partial W_k^i}$pour les poids en commun et sera nulle pour les autres poids.

On peut alors approcher le minimum du contraste associé à la vraisemblance par optimisation différentielle.

4.2 Identifiabilité de la fonction MLP

En général, un MLP vu comme une fonction paramétrique sur $\mathbb{R}^D$ n'est pas identifiable. Cependant en restreignant l'espace possible des paramètres, on obtient une classe de fonctions identifiables. Nous donnons dans la suite des conditions nécessaires et suffisantes pour que le modèle soit identifiable dans le cas d'un MLP à une couche cachée, avec des tangentes hyperboliques pour fonctions d'activation (Dans la suite on ne considère plus que cette famille de MLP).

Notation 7 *La notation des poids, adoptée ici, est similaire à celle de la figure 2.*

Notation 8 *Si $X = (X_1, \cdots, X_m)^T \in \mathbb{R}^m$ est un vecteur d'entrée, on note $v_i(X)$ l'impulsion de la i-ème unité cachée :*

$$\nu_i(X) = W_{i1} + \sum_{j=2}^{m+1} W_{ij} X_{j-1}$$

Fixons m. Le MLP avec C unités cachées est associé à C applications affines : $\begin{array}{c} \mathbb{R}^m \to \mathbb{R} \\ X \longmapsto \nu_j(X) \end{array}$. On dira que deux fonctions affines ν_1, ν_2 sont "signe-équivalentes" si $|\nu_1| = |\nu_2|$.

4.2.1 MLP réductible et irréductible

L'identifiabilité des MLP avec une sortie a été étudié dans [11], on rappelle ici les notions utilisées. On dira qu'un MLP est réductible s'il vérifie au moins une des conditions (**R**)

[2] la dérivée $\frac{\partial F_{W^i}(y_{t-p}^{t-1})}{\partial W_k^i}$ s'obtient en commençant la rétropropagation en fixant l'erreur égale à 1.

suivantes :

1. Il existe un indice $j \in \{1, \cdots, C\}$ tel que tous les poids A_{ij} , $1 \leq i \leq s$ sont nuls.
2. Il existe au moins deux indices différents $j_1, j_2 \in \{1, \cdots, C\}$ tels que les fonctions ν_{j_1}, ν_{j_2} soient signe-équivalentes
3. Il existe au moins un indice $j \in \{1, \cdots, C\}$ tel que la fonction ν_j est constante

On dira qu'un MLP est irréductible s'il n'est pas réductible. On note $\mathcal{N}_{m,C,s}$ l'ensemble des MLP avec m entrées, C unités cachées et s sorties qui sont irréductibles. $\mathcal{N}_{m,C,s}$ est isomorphe à $\mathbb{R}^D$, avec $D = (m+1) \times C + (C+1) \times s$.

Remarque 8 *Si $C = 0$, $\mathcal{N}_{m,0,s}$ représente les fonctions linéaires de $\mathbb{R}^m \longrightarrow \mathbb{R}^s$*

Espace des paramètres Il y a des transformations triviales qui ne changent pas la fonction MLP. Par exemple, si on choisit l'unité cachée i, que l'on change le signe de tous les poids W_{ij} pour $1 \leq j \leq C$ et que l'on change aussi le signe des A_{ij}, $1 \leq i \leq s$, comme la fonction tanh est impaire cela ne changera pas la fonction. Notons $\zeta_j(MLP)$ le MLP résultant de cette transformation.

Une autre possibilité est d'interchanger les deux unités cachées j_1 et j_2, ainsi que les poids correspondants, on note $\eta_{j_1,j_2}(MLP)$ le MLP correspondant. Les applications η_{j_1,j_2}, ζ_j , $j_1, j_2, j \in \{1, \cdots, C\}$ génèrent un groupe fini $\mathcal{G}_{m,C,s}$ de transformations sur l'ensemble $\mathcal{N}_{m,C,s}$ (de cardinal $2^C C!$).

On dira que deux MLP M_1 et M_2 sont équivalents ($M_1 \mathcal{R} M_2$), si et seulement si il existe une transformation $\phi \in \mathcal{G}_{m,C,s}$ telle que $M_1 = \phi(M_2)$. Pour que les fonctions MLP soient identifiables, on considère donc un ensemble de paramètre Θ^W inclus dans un ensemble quotient $\mathcal{N}_{m,C,s}/\mathcal{R}$, qui est évidement un espace polonais.

4.2.2 Identifiabilité des MLP dans $\bigcup_{C=0}^{M} \mathcal{N}_{m,C,s}/\mathcal{R}$, $M \in \mathbb{N}$

Soit $M \in \mathbb{N}$, [11] prouve que si $s = 1$ les MLP sont identifiables dans $\bigcup_{C=0}^{M} \mathcal{N}_{m,C,s}/\mathcal{R}$, i.e. $F_W = F_{W'} \Leftrightarrow W = W'$. On en déduit le théorème :

Théorème 7 *Soit deux MLP, $F_W, F_{W'}$, appartenant à $\bigcup_{C=0}^{M} \mathcal{N}_{m,C,s}/\mathcal{R}$ irréductibles, où $m, s \in (\mathbb{N}^*)^2$, alors $F_W = F_{W'} \Leftrightarrow W = W'$.*

Preuve Supposons que ce ne soit pas le cas, donc qu'il existe $W \neq W'$ tels que $F_W = F_{W'}$. Alors il existe deux MLP extraits (voir section 4.1) équivalents à deux MLP irréductibles (obtenus en enlevant tous les poids reliés aux unités cachées qui ne sont pas reliés à la sortie "i") $F_{W^i}, F_{W'^i}$, tels que $W^i \neq W'^i$ et $F_W i = F_{W'^i}$. Sinon au moins un des MLP vérifie (**R**)-1 et n'est pas irréductible. Mais, comme les MLP extraits ont une seule sortie, cela contredit le résultat de [11] ■

4.3 Identification presque-sûre du modèle

L'étude des propriétés statistiques de l'estimateur de minimum de contraste associé à la vraisemblance, nous permet de déduire le théorème suivant :

Théorème 8 *Soit un modèle correspondant au modèle (1) avec un MLP pour fonction F_W. Supposons satisfaites les conditions suivantes :*

1. $(\varepsilon_t)_{t\in\mathbb{N}^*}$ *est une suite de variables aléatoires de* $\mathbb{R}^d$ *centrées, de matrice de covariance* Γ_0 *définie positive, i.i.d. et indépendantes de l'état initial de la chaîne* $\left(Y_{t-p+1}^t\right)_{t\in\mathbb{N}}$. ε_1 *a une densité positive par rapport à la mesure de Lebesgue sur* $\mathbb{R}^d$*, avec, pour* $\eta > 0$*,* $E\left(\|\varepsilon_1\|^{12+\eta}\right) < \infty$.
2. $\theta = (W, \Gamma^{-1})$ *appartient à un sous-ensemble compact* Θ *:*

$$\Theta = \Theta^W \times \Theta^{\Gamma^{-1}} \subset \left(\bigcup_{C=0}^{M} \mathcal{N}_{d\times p,C,d}/\mathcal{R}\right) \times \mathbb{R}^{(d+1)d/2}$$

   où M *est le nombre maximal d'unités cachées pour les MLP considérés.*
3. *Pour toute matrice* $\Gamma^{-1} \in \Theta^{\Gamma^{-1}}$ *,* $\rho(\Gamma) > 0$.
4. *On suppose qu'il existe* $0 \leq m_0 \leq M$ *tel que le vrai modèle appartient à* $\Theta \cap \mathcal{N}_{d\times p,m_0,d}$.
5. *Les matrices* I_0 *et* J_0 *du théorème 4 sont supposées définies positives.*
6. *La vitesse de pénalisation* $c(n)$ *du contraste pénalisé* $CP_{n,\delta}(\theta)$ *est telle que :*

$$\lim_{n\longrightarrow\infty} \frac{c(n)}{n} = 0$$

   et

$$\liminf_{n\longrightarrow\infty} \frac{c(n)}{2\ln\ln n} > \frac{\lambda_{max} J_0}{2\lambda_{min} I_0}.$$

Alors le couple $\left(\hat{\delta}_n, \hat{\theta}_{n,\delta_n}\right)$ *converge* P_{θ_0}*-p.s. vers la vraie valeur* $(\delta_0, \theta_{0,\delta_0})$.

Preuve : Il est aisé de montrer que la fonction F_W est bornée pour tout $W \in \Theta^W$, on aura de plus pour tout $y_1^p \in \left(\mathbb{R}^d\right)^p$(cf Yao et Mangeas [12]) :

$$\|\nabla F_{W_0}(y_1^p)\| \leq Cte(1 + \|y_1^p\|)$$
$$\|HF_{W_0}(y_1^p)\| \leq Cte(1 + \|y_1^p\|^2)$$

et en notant TF_W la dérivée troisième par rapport aux paramètres de F_W , en remarquant que Θ^W est compact, il existera une constante γ telle que :

$$\forall W \in \Theta^W,\ \|TF_W(y_1^p)\| \leq \gamma(1 + \|y_1^p\|^3).$$

Comme on doit avoir $\|TF_W(y_1^p)\| \leq \gamma(1 + \|y_1^p\|^{a/2})$, avec un bruit ayant un moment d'ordre strictement supérieur à $2a$, il faut donc que le bruit possède un moment strictement supérieur à 12 dans le cas du perceptron. Il est facile de voir que les hypothèses relatives à la continuité sont satisfaites. De plus le modèle est identifiable grâce au théorème 7 du chapitre **??**. Le théorème 8 est alors une conséquence du théorème 6 ■

Remarque 9 *Si* γ *est une constante positive, un terme de pénalisation tel que* $c(n) = \gamma\ln(n)$ *satisfait les conditions du théorème 8. Si on choisit* $\gamma = 1$*, on aboutit alors à un critère de sélection de modèle du type (on rappelle que* B *est le nombre de paramètres du modèle) :*

$$CP_{n,\delta}(\theta) = U_n(\theta) + \frac{\ln(n)}{n}B = \frac{1}{2}\ln\det\left(\Gamma_n(W)\right) + \frac{\ln(n)}{n}B + Cte$$

ce qui revient à minimiser :

$$\widetilde{CP}_{n,\delta}(\theta) = \frac{1}{2}\ln\det\left(\Gamma_n(W)\right) + \frac{\ln(n)}{n}B$$

et cela correspond exactement au critère BIC.

5 Conclusion

On a montré que la fonctionnelle à minimiser pour tenir compte des termes non diagonaux de la matrice de covariance du bruit est le logarithme du déterminant de sa matrice de covariance empirique. Cet estimateur est celui du maximum de vraisemblance pour un bruit gaussien. Cependant, même si le bruit n'est pas gaussien, on a montré que cet estimateur avait de bonnes propriétés statistiques. En suivant la même démarche que Yao et Mangeas [12], nous avons montré qu'un contraste pénalisé de type BIC, convergeait p.s. vers le vrai modèle. Cela permet, par exemple pour les perceptrons multicouches, d'estimer et d'identifier le modèle grâce à un algorithme de type Step Wise descendant (cf Cottrell et al. [2]).

Références

[1] H. Cartan. *Calcul différentiel.* Herman, 1970.

[2] M. Cottrell, et al. Neural modeling for time series : a statistical stepwise method for weight elimination. *IEEE Transaction on Neural Networks*, 6 :1355–1364, 1995.

[3] M. Duflo. *Algorithmes stochastiques.* Springer-Verlag, 1996.

[4] M. Duflo. *Random iterative models.* Springer-Verlag, 1997.

[5] M. Duflo, R. Senoussi, and A. Touati. Sur la loi des grands nombres pour les martingales vectorielles et l'estimateur des moindres carrés d'un modèle de regression. *Annales de L'I.H.P.*, 26 :549–566, 1990.

[6] X. Guyon. *Random fields on a network-modeling.* Statistics and applications. Springer-Verlag, 1997.

[7] Roger A. Horn and Charles R. Johnson. *Matrix analysis.* Cambridge University Press, 1985.

[8] M. Mangeas. Propriétés statistiques des modèles paramétriques non-linéaires de prévision de série temporelles : Etude des réseaux de neurones à propagation directe. Thèse, Université de Paris 1, 1997.

[9] William H. Press, et al. *Numerical recipes in C : The art of scientific computing.* Cambridge University Press, 1992.

[10] R. Senoussi. Statistique asymptotique presque sûre de modèle convexe. *Ann. IHP. (Probabilités et Statistiques)*, 26 :19–44, 1990.

[11] H.J. Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output Map. *Neural Networks*, 5 :589–593, 1992.

[12] J. Yao and M. Mangeas. On least square estimation for stable nonlinear AR processes. Prépublication du SAMOS 67, Université Paris 1, 1996. à paraître dans Annals Inst. Math. Statist.