Journal de la Société Française de Statistique
Vol. 151 No. 2 (2010)

Estimation récursive en régression inverse par tranches (sliced inverse regression)

Title: Recursive estimation for sliced inverse regression

Thi Mong Ngoc Nguyen [1,2] et Jérôme Saracco [1,2,3]

Résumé :

L'intérêt des méthodes récursives est de prendre en compte l'arrivée temporelle des informations et d'affiner ainsi au fil du temps les estimations. L'idée est d'utiliser les estimations calculées sur la base de données initiales et de les remettre à jour en tenant uniquement compte des nouvelles données arrivant dans la base. Le gain en terme de temps de calcul peut être très intéressant et les applications d'une telle approche sont nombreuses. Dans cet article, nous nous intéressons à la méthode SIR (sliced inverse regression, que l'on peut traduire par régression inverse par tranches) qui permet d'estimer le paramètre θ dans un modèle semi-paramétrique de régression du type $y = f(x'\theta, \varepsilon)$ sans avoir à estimer le paramètre fonctionnel f ni à spécifier la loi de l'erreur ε. Dans le cas particulier où l'on considère $H = 2$ tranches, il est possible d'obtenir une expression analytique de l'estimateur de la direction de θ. Nous proposons dans cet article une forme récursive pour cet estimateur. Nous donnons des propriétés asymptotiques de cet estimateur (convergence presque sûre et normalité asymptotique). Nous illustrons aussi sur des simulations le bon comportement numérique de l'approche récursive proposée. Un avantage majeur de l'utilisation de la forme récursive est que les temps de calculs des estimateurs sont beaucoup plus courts que ceux obtenus avec la forme non récursive, en particulier lorsque la dimension de x est grande.

Abstract: In this paper, we study a recursive estimation procedure for sliced inverse regression. When the number H of slices is equal to two, we obtain an analytic expression of the estimator of the direction of the parameter θ in the semiparametric regression model $y = f(x'\theta, \varepsilon)$, which does not require the estimation of the link function f. We propose a recursive estimation procedure for this estimator. We establish some asymptotic properties of the estimator. A simulation study illustrates the good numerical behavior of the estimator. The recursive approach has the advantage to be computationaly faster than the non recursive one.

Mots-clés : estimation récursive, modèle semi-paramétrique, Sliced Inverse Regression (SIR).
Keywords: recursive estimation, semiparametric regression model, Sliced Inverse Regression (SIR).
Classification AMS 2000 : primaire 62H12, 62J99 ; secondaire 62L12

[1] Université Bordeaux 1, Institut de Mathématiques de Bordeaux, UMR CNRS 5251
351 cours de la libération, 33405 Talence Cedex, France.
E-mail : Thi.Mong.Ngoc.Nguyen@math.u-bordeaux1.fr ; Jerome.Saracco@math.u-bordeaux1.fr
[2] Equipe CQFD, INRIA Bordeaux Sud-Ouest, France.
[3] Université Montesquieu-Bordeaux 4, GREThA, UMR CNRS 5113

Journal de la Société Française de Statistique, **Vol. 151** No. 2 19-46
http://www.sfds.asso.fr/journal
 ISSN: 2102-6238

1. Introduction

Les modèles de régression sont très utiles pour modéliser la liaison entre une variable à expliquer y et une variable explicative x. Ils sont appliqués à de nombreux domaines tels que l'économie, la biostatistique ou encore les sciences de l'environnement. Dans la littérature statistique, deux grandes classes de modèles de régression sont omniprésentes : les modèles paramétriques et les modèles non paramétriques. Ces deux types de modèles sont caractérisés par des avantages et des défauts spécifiques. Les modèles paramétriques permettent une interprétation claire de l'impact de la variable explicative sur la variable à expliquer. Cependant, le choix d'un bon modèle paramétrique au vu des données n'est pas toujours évident. Ainsi le modèle paramétrique choisi peut ne pas être en adéquation avec les données et peut donc parfois être très "éloigné" de la réalité des données. En conséquence les conclusions en découlant peuvent alors être erronées. Les modèles non paramétriques ont été proposés afin de surmonter ce problème de sélection du bon modèle paramétrique. Ils offrent davantage de flexibilité car aucune hypothèse paramétrique n'est imposée dans le modèle, seules des hypothèses de régularité de la fonction de lien entre x et y sont imposées. Cependant il faut estimer la fonction de lien le plus souvent au moyen de procédures de calculs intensifs (en particulier en ce qui concerne la recherche des paramètres de lissage), ce qui est beaucoup plus lourd en temps de calcul. De plus, l'interprétation de la fonction de lien n'est pas toujours évidente.

Les modèles semi-paramétriques ont été développés pour conjuger les avantages des approches paramétriques et non paramétriques, à savoir la capacité d'interprétation des modèles paramétriques et la souplesse des modèles non paramétriques. Ces modèles dépendent généralement d'un paramètre de dimension fini, noté θ, ainsi que d'une fonction de lien f à estimer. Dans cet article, nous nous intéressons au modèle semi-paramétrique de régression proposé par Duan et Li [3] lorsque la variable à expliquer y est à valeurs dans $\mathbb{R}$ et la covariable x appartient à $\mathbb{R}^p$:

$$y = f(\theta' x, \varepsilon), \tag{1.1}$$

où :

(i) le paramètre θ est un vecteur inconnu de $\mathbb{R}^p$;

(ii) le bruit ε est un terme d'erreur aléatoire indépendant de x, aucune hypothèse n'est faite sur la distribution de ε ;

(iii) la fonction de lien f est un paramètre fonctionnel à valeur dans $\mathbb{R}$, inconnu et arbitraire.

Dans le cadre de ce modèle, le paramètre θ n'est pas totalement identifiable, seule la direction de θ est identifiable. On parlera de direction EDR (comme "effective dimension reduction"). Notons que Li [7] a généralisé le modèle (1.1) en considérant une fonction f avec un nombre fini $d < p$ d'arguments $\theta_1' x$, ..., $\theta_d' x$. Dans ce modèle multi-indices, seul le sous-espace linéaire de $\mathbb{R}^p$ engendré par les vecteurs $\theta_1, \ldots, \theta_d$ est identifiable, on parle alors d'espace EDR.

Duan et Li [3] et Li [7] ont proposé une méthode permettant d'estimer la direction EDR (ou l'espace EDR) ne nécessitant pas l'estimation de la fonction de lien f du modèle semi-paramétrique (1.1) : leur méthode s'appelle SIR, acronyme de "Sliced Inverse Regression" que l'on peut traduire par régression inverse par tranches. Les techniques d'estimation développées juqu'à présent pour la méthode SIR ne sont pas récursives. D'une manière générale, l'avantage des méthodes récursives est de prendre en compte l'arrivée temporelle des informations et d'affiner ainsi au fil du temps les algorithmes d'estimation mis en œuvre. Un intérêt majeur de ces méthodes

 ISSN: 2102-6238

est qu'il n'est pas nécessaire de relancer le calcul de l'estimateur sur la totalité des données à chaque fois que la base de données est complétée par de nouvelles observations. L'idée est ici d'utiliser les estimations calculées sur la base de données initiales et de les remettre à jour en tenant uniquement compte des nouvelles données arrivant dans la base. Le gain en terme de temps de calcul peut être très intéressant et les applications d'une telle approche sont nombreuses, le lecteur pourra se reporter par exemple aux ouvrages de Duflo [4], [5] pour une présentation de méthodes récursives aléatoires.

Dans cet article, nous proposons une procédure d'estimation récursive de la direction de θ qui ne nécessite pas l'estimation du paramètre fonctionnel f du modèle semi-paramétrique (1.1) en adaptant la méthode SIR au cadre récursif. Dans la section 2, nous rappelons les fondements de la méthode SIR, puis nous donnons une expression analytique de l'estimateur de la direction de θ dans le cas particulier où l'on considère $H = 2$ tranches (pour construire l'estimateur). Nous proposons alors une forme récursive pour cet estimateur. Nous donnons ensuite à la section 3 quelques propriétés asymptotiques de l'estimateur proposé. Des simulations numériques illustrant le bon comportement de la méthode sont décrites à la section 4. Elles montrent que les temps de calculs obtenus avec la forme récursive de l'estimateur sont bien plus courts que ceux obtenus avec la forme non récursive. Enfin, pour conclure, nous donnons à la section 5 quelques remarques finales et extensions possibles de ce travail. Les démonstrations et développements théoriques sont donnés en Annexes.

2. Estimation récursive du paramètre θ

Nous présentons tout d'abord brièvement la méthode SIR et nous rappelons la définition de l'estimateur de la direction EDR (pour un nombre $H > 1$ quelconque de tranches) comme solution d'un problème aux valeurs propres d'une certaine matrice d'intérêt. Nous précisons ensuite une forme analytique de cet estimateur lorsque $H = 2$. Nous proposons alors dans un premier temps une forme récursive pour l'estimateur de la matrice d'intérêt. Puis nous donnons une forme récursive de l'estimateur de la direction EDR lorsque $H = 2$.

2.1. Méthode SIR

La régression inverse par tranches ou méthode SIR est une méthode de régression semi-paramétrique reposant sur un argument géométrique. Elle a été introduite par Li [7] et Duan et Li [3] . Cette méthode repose sur une propriété de la fonction de régression inverse (c'est-à-dire qu'au lieu de régresser y sur x, ce sont des propriétés de la régression de x sur y qui vont être utilisées). L'avantage de cette inversion de rôle est que la dimension du problème a été réduite : nous avons en effet maintenant p problèmes de dimension 1, la régression inverse permettant de régresser chaque coordonnée de x sur y. Le coût à payer est de rajouter une hypothèse probabiliste sur la distribution de la variable explicative x :

(H) *La variable explicative x possède une distribution de probabilité non dégénérée telle que, pour tout $b \in \mathbb{R}^p$, l'espérance conditionnelle $\mathbb{E}[b'x \mid \theta'x]$ est linéaire en $\theta'x$.*

Cette hypothèse, encore appelée condition de linéarité, est vérifiée lorsque la variable explicative x suit une distribution elliptique, en particulier lorsque la distribution de x est multinormale. Avant

 ISSN: 2102-6238

de poursuivre la présentation de la méthode SIR, il apparaît intéressant de discuter rapidement cette hypothèse **(H)** sur laquelle est fondée la théorie de SIR et qui n'est pas une hypothèse usuelle pour les modèle de régression. Remarquons tout d'abord qu'elle est très difficile à vérifier a priori vu qu'elle dépend de la direction inconnue θ. Par contre, il est possible de tester l'ellipticité de x. Notons que diverses techniques ont été proposées pour "forcer" les données à être elliptiques, voir par exemple la méthode proposée par Cook et Nachtsheim [2] . Lorsque la dimension de x est grande, il devient très difficile de forcer les données à être elliptiques avec ce type de méthodes. Cependant, Hall et Li [6] ont montré par un argument Bayésien que la condition fondamentale **(H)** est presque sûrement vérifiée lorsque $p \to \infty$. Ils ont alors précisé comment ce résultat permet d'élargir le champ d'applicabilité de la régression inverse par tranches. On peut aussi mentionner que Li [7] présente des résultats obtenus par simulations (sur des échantillons pour lesquels la variable explicative x n'est pas elliptique) montrant une robustesse de la méthode SIR à des écarts à l'hypothèse **(H)**. Enfin, pour conclure cette discussion, on renvoit le lecteur à l'article de Chen et Li [1] où cette hypothèse est plus longuement discutée.

Posons $\mu = \mathbb{E}[x]$ et $\Sigma = \mathbb{V}(x)$. Nous donnons maintenant le théorème de caractérisation de la direction de θ établi par Li [7] .

Theorem 2.1. *Dans le cadre du modèle (1.1) et sous l'hypothèse* **(H)***, la courbe de régression inverse centrée, $y \mapsto \mathbb{E}[x \mid \mathbb{T}(y)] - \mu$, appartient au sous-espace linéaire de $\mathbb{R}^p$ engendré par θ, où $\mathbb{T}$ est une transformation monotone de y (qui correspondra au "tranchage" précisé ultérieurement).*

Une conséquence directe de ce théorème est que la matrice de variances-covariances de cette courbe, $\Gamma = \mathbb{V}(\mathbb{E}[x \mid \mathbb{T}(y)])$, est dégénérée dans toute direction Σ-orthogonale à θ. Ainsi le vecteur propre $\tilde{\theta}$ associé à la valeur propre non nulle de $\Sigma^{-1}\Gamma$ est colinéaire à θ. Ce vecteur propre $\tilde{\theta}$ est donc une direction EDR.

Remarque 2.1. *Dans le cadre d'une généralisation du modèle (1.1) à un modèle à plusieurs indices, $\theta_1'x$, ..., $\theta_d'x$ avec $d < p$ (voir Li [7]), les vecteurs propres $\tilde{\theta}_1, \ldots, \tilde{\theta}_d$, associés aux d plus grandes valeurs propres non nulles de la matrice $\Sigma^{-1}\Gamma$ appartiennent au sous-espace linéaire engendré par les vecteurs $\theta_1, \ldots, \theta_d$, c'est à dire appartient l'espace EDR.*

En pratique, nous disposons d'un échantillon d'observations $\mathscr{S} = \{(x_i, y_i), i = 1, ..., N\}$. Nous allons maintenant rappeler la méthode d'estimation (non récursive) telle qu'elle a été introduite par Duan et Li [3] .

On note $\bar{x}_N$ et Σ_N la moyenne et la variance empirique de l'échantillon $\{x_i,\ i = 1, \ldots, N\}$, données respectivement par

$$\bar{x}_N = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{et} \quad \Sigma_N = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}_N)(x_i - \bar{x}_N)'.$$

Il reste alors à estimer la matrice Γ et donc à choisir une transformation $\mathbb{T}$ telle que cette matrice soit facilement estimable. Un choix particulier de $\mathbb{T}$ a été proposé par Duan et Li [3] : le "tranchage" qui est une discrétisation de y fondée sur un découpage du support de y en H tranches distinctes $s_1, \ldots, s_H$. Dans ce cadre, en notant $p_h = \mathbb{P}(y \in s_h)$ et $m_h = \mathbb{E}[x \mid y \in s_h]$ pour $h = 1, ..., H$, la matrice Γ s'écrit très simplement sous la forme :

$$\Gamma = \sum_{h=1}^{H} p_h(m_h - \mu)(m_h - \mu)'.$$

 ISSN: 2102-6238

L'estimation des différentes grandeurs intervenant dans la matrice Γ ne pose aucune difficulté : il suffit de substituer les moments empiriques aux moments théoriques. Nous obtenons alors l'estimateur Γ_N :

$$\Gamma_N = \sum_{h=1}^{H} p_{h,N}(m_{h,N} - \bar{x}_N)(m_{h,N} - \bar{x}_N)',$$

où

$$p_{h,N} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[y_i \in s_h] = \frac{N_h}{N} \quad \text{et} \quad m_{h,N} = \frac{1}{N_h}\sum_{i \in s_h} x_i$$

sont respectivement la proportion empirique des y_i tombant dans la tranche s_h, et la moyenne empirique des x_i tels que y_i appartienne à la tranche s_h. Ensuite, il suffit de déterminer le vecteur propre $\tilde{\theta}_N$ associé à la plus grande valeur propre de $\Sigma_N^{-1}\Gamma_N$. On dit que $\tilde{\theta}_N$ est un estimateur de la direction EDR.

Des résultats asymptotiques sont disponibles dans la littérature pour l'estimateur SIR non récursif de la direction de θ. Sans être exhaustif, on peut mentionner que la vitesse de convergence de la direction EDR estimée vers la vraie direction est $n^{-1/2}$, voir Li [7] et que la normalité asymptotique des éléments propres d'intérêt (projecteur propre sur l'espace EDR estimé, valeurs propres et vecteurs propres principaux, i.e. estimateurs des directions EDR) a été obtenue par Saracco [10] .

Etude du cas où $H = 2$ permettant d'obtenir une expression analytique de l'estimateur $\tilde{\theta}_N$. Dans cet article, nous nous intéresserons au cas particulier où l'on ne considère que deux tranches notées s_1 et s_2. La raison essentielle de ce choix est que nous pouvons obtenir facilement une expression analytique de l'estimateur de la direction de θ. En effet, lorsque $H = 2$, la matrice Γ s'écrit :

$$\Gamma = p_1 z_1 z_1' + p_2 z_2 z_2', \tag{2.1}$$

avec $z_h = m_h - \mu$ pour $h = 1, 2$. On peut alors montrer que la valeur propre non nulle λ de $\Sigma^{-1}\Gamma$ et le vecteur propre $\tilde{\theta}$ correspondant (colinéaire à θ) s'écrivent sous la forme :

$$\lambda = \frac{p_1}{p_2} z_1' \Sigma^{-1} z_1 \qquad \text{et} \qquad \tilde{\theta} = \Sigma^{-1}(z_1 - z_2). \tag{2.2}$$

Les calculs de $\tilde{\theta}$ et de λ sont détaillés en Annexe 1. Nous pouvons ainsi déduire des relations (2.2) des estimateurs λ_N et $\tilde{\theta}_N$ de λ et $\tilde{\theta}$:

$$\lambda_N = \frac{p_{1,N}}{p_{2,N}} z_{1,N}' \Sigma_N^{-1} z_{1,N} \qquad \text{et} \qquad \tilde{\theta}_N = \Sigma_N^{-1}(z_{1,N} - z_{2,N}), \tag{2.3}$$

où $z_{h,N} = m_{h,N} - \bar{x}_N$ pour $h = 1, 2$.

Dans la sous-section suivante, nous allons tout d'abord proposer une écriture récursive pour l'estimateur de la matrice d'intérêt $\Sigma^{-1}\Gamma$ pour un nombre H quelconque de tranches. Ensuite, nous déterminerons la forme récursive de l'estimateur de la direction θ lorsque $H = 2$.

 ISSN: 2102-6238

2.2. *Estimateur récursif de la matrice d'intérêt* $\Sigma^{-1}\Gamma$

Lorsque l'on s'intéresse à la forme non récursive des estimateurs, nous disposons d'un échantillon d'observations $\{(x_i,y_i), i=1,...,N\}$ de variables aléatoires (x,y) indépendantes et identiquement distribuées issues du modèle (1.1). Pour la forme récursive des estimateurs, on scinde cet échantillon en deux parties : le sous-échantillon $\{(x_i,y_i), i=1,...,N-1\}$ et l'observation (x_N,y_N).

Dans un premier temps, nous donnerons les formes récursives de $\bar{x}_N$, Σ_N et Σ_N^{-1}, puis celle de Γ_N. Ensuite, nous décrirons la forme récursive de la matrice d'intérêt $\Sigma_N^{-1}\Gamma_N$.

2.2.1. *Forme récursive de* $\bar{x}_N$, Σ_N *et* Σ_N^{-1}

La forme récursive de la moyenne empirique $\bar{x}_N$ des N observations $x_1,\ldots,x_N$ est la suivante :

$$\bar{x}_N = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{N-1}{N}\bar{x}_{N-1} + \frac{1}{N}x_N = \bar{x}_{N-1} + \frac{1}{N}\Phi_N \tag{2.4}$$

avec $\Phi_N = x_N - \bar{x}_{N-1}$. Elle fait intervenir, dans le premier terme, la moyenne empirique $\bar{x}_{N-1}$ des $N-1$ premières observations x_i, ainsi que la Nème observation x_N dans le second terme. De manière similaire, la forme récursive de la matrice des variances-covariances empiriques Σ_N est donnée par :

$$\begin{aligned}
\Sigma_N &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x}_N)(x_i-\bar{x}_N)' \\
&= \frac{N-1}{N}\Sigma_{N-1} + \frac{N-1}{N^2}(x_N-\bar{x}_{N-1})(x_N-\bar{x}_{N-1})' \\
&= \frac{N-1}{N}\Sigma_{N-1} + \frac{N-1}{N^2}\Phi_N\Phi_N'.
\end{aligned} \tag{2.5}$$

On peut montrer de même, via l'équation de Riccati, que l'inverse de Σ_N a pour forme récursive :

$$\Sigma_N^{-1} = \frac{N}{N-1}\Sigma_{N-1}^{-1} - \frac{N}{(N-1)(N+\rho_N)}\Sigma_{N-1}^{-1}\Phi_N\Phi_N'\Sigma_{N-1}^{-1} \tag{2.6}$$

avec $\rho_N = \Phi_N'\Sigma_{N-1}^{-1}\Phi_N$.

2.2.2. *Forme récursive de* Γ_N

On se place ici dans le cas général où le nombre H de tranches est quelconque ($H \geq 2$). Afin de donner la forme récursive de l'estimateur Γ_N de la matrice Γ, nous supposons que la nouvelle observation (x_N,y_N) est telle que y_N appartienne à la tranche s_{h^*}.

Notons tout d'abord que la forme récursive de l'estimateur $p_{h,N}$ de p_h est

$$p_{h,N} = \begin{cases} \frac{N-1}{N}p_{h^*,N-1} + \frac{1}{N} & \text{si } h = h^*, \\ \frac{N-1}{N}p_{h,N-1} & \text{sinon.} \end{cases} \tag{2.7}$$

 ISSN: 2102-6238

La forme récursive de $m_{h,N}$ est donnée, pour $\Phi_{h^*,N} = x_N - m_{h^*,N-1}$, par

$$m_{h,N} = \begin{cases} m_{h^*,N-1} + \frac{1}{N_{h^*,N-1}+1}\Phi_{h^*,N} & \text{si } h = h^*, \\ \\ m_{h,N-1} & \text{sinon.} \end{cases} \tag{2.8}$$

A partir des relations (2.4) et (2.8), on obtient la forme récursive de $z_{h,N} = m_{h,N} - \bar{x}_N$

$$z_{h,N} = \begin{cases} z_{h^*,N-1} - \frac{1}{N}\Phi_N + \frac{1}{N_{h^*,N-1}+1}\Phi_{h^*,N} & \text{si } h = h^*, \\ \\ z_{h,N-1} - \frac{1}{N}\Phi_N & \text{sinon.} \end{cases} \tag{2.9}$$

En écrivant Γ_N sous la forme

$$\Gamma_N = \sum_{h \neq h^*} p_{h,N} z_{h,N} z'_{h,N} + p_{h^*,N} z_{h^*,N} z'_{h^*,N},$$

nous pouvons déduire des relations (2.7) et (2.9) la forme récursive suivante pour Γ_N :

$$\begin{aligned} \Gamma_N &= \frac{N-1}{N}\Gamma_{N-1} - \frac{N-1}{N^2}\sum_{h=1}^{H} p_{h,N-1}\left(z_{h,N-1}\Phi'_N + \Phi_N z'_{h,N-1}\right) \\ &\quad + \frac{N-1}{N^3}\Phi_N\Phi'_N + \frac{N-1}{N}p_{h^*,N-1}A_{h*,N} + \frac{1}{N}B_{h*,N}B'_{h*,N} \end{aligned} \tag{2.10}$$

avec

$$\begin{aligned} A_{h^*,N} &= \frac{1}{N_{h^*,N-1}+1}\left(z_{h^*,N-1}\Phi'_{h^*,N} + \Phi_{h^*,N}z'_{h^*,N-1}\right) \\ &\quad - \frac{1}{N(N_{h^*,N-1}+1)}\left(\Phi_N\Phi'_{h^*,N} + \Phi_{h^*,N}\Phi'_N\right) + \frac{1}{(N_{h^*,N-1}+1)^2}\Phi_{h^*,N}\Phi'_{h^*,N} \end{aligned}$$

et $B_{h^*,N} = z_{h^*,N-1} - \frac{1}{N}\Phi_N + \frac{1}{N_{h^*,N-1}+1}\Phi_{h^*,N}$.

2.2.3. *Forme récursive de la matrice d'intérêt* $\Sigma_N^{-1}\Gamma_N$

A partir des relations (2.6) et (2.10), on peut obtenir une forme récursive pour $\Sigma_N^{-1}\Gamma_N$:

$$\Sigma_N^{-1}\Gamma_N = \Sigma_{N-1}^{-1}\Gamma_{N-1} + \Sigma_{N-1}^{-1}C_{h^*,N} - \frac{1}{N+\rho_N}\Sigma_{N-1}^{-1}\Phi_N\Phi'_N\Sigma_{N-1}^{-1}\left[\Gamma_{N-1} + C_{h^*,N}\right] \tag{2.11}$$

en posant

$$\begin{aligned} C_{h^*,N} &= -\frac{1}{N}\sum_{h=1}^{H} p_{h,N-1}\left(z_{h,N-1}\Phi'_N + \Phi_N z'_{h,N-1}\right) \\ &\quad + \frac{1}{N^2}\Phi_N\Phi'_N + p_{h^*,N-1}A_{h*,N} + \frac{1}{N-1}B_{h*,N}B'_{h*,N}. \end{aligned}$$

 ISSN: 2102-6238

2.3. *Estimateur récursif de la direction de* θ

Nous nous plaçons maintenant dans le cas où $H = 2$. Dans ce cas, nous avons donné en (2.3) l'expression analytique des formes non récursives des estimateurs λ_N et $\tilde{\theta}_N$. Nous allons maintenant en déduire leurs formes récursives. Nous supposons encore ici que la nouvelle observation (x_N, y_N) est telle que $y_N \in s_{h^*}$ avec $h^* = 1$ ou 2.

2.3.1. *Forme récursive de* $\tilde{\theta}_N$

D'après (2.6) et (2.9), nous obtenons la forme récursive de l'estimateur $\tilde{\theta}_N$ de $\tilde{\theta}$:

$$\begin{aligned} \tilde{\theta}_N &= \tfrac{N}{N-1}\tilde{\theta}_{N-1} - \tfrac{N}{(N-1)(N+\rho_N)}\Sigma_{N-1}^{-1}\Phi_N\Phi_N'\tilde{\theta}_{N-1} \\ &- \tfrac{(-1)^{h^*}N}{(N_{h^*,N-1}+1)(N-1)}\left(\Sigma_{N-1}^{-1} - \tfrac{1}{N+\rho_N}\Sigma_{N-1}^{-1}\Phi_N\Phi_N'\Sigma_{N-1}^{-1}\right)\Phi_{h^*,N}. \end{aligned} \tag{2.12}$$

2.3.2. *Forme récursive de* λ_N

Nous avons aussi la forme récursive de l'estimateur λ_N de λ :

$$\lambda_N = \frac{p_{1,N-1} + \frac{1}{N-1}\mathbb{I}[h^*=1]}{p_{2,N-1} + \frac{1}{N-1}\mathbb{I}[h^*=2]} v_N' \left(\frac{N}{N-1}\Sigma_{N-1}^{-1} - \frac{N}{(N-1)(N+\rho_N)}\Sigma_{N-1}^{-1}\Phi_N\Phi_N'\Sigma_{N-1}^{-1}\right) v_N,$$

avec $v_N = z_{1,N-1} - \frac{1}{N}\Phi_N + \frac{1}{N_{1,N-1}+1}\Phi_{1,N}\mathbb{I}[h^*=1]$. Il est possible d'écrire cet estimateur sous la forme

$$\lambda_N = \frac{N}{N-1}\lambda_{N-1} + F(x_N, y_N, N, N_{1,N-1}, N_{2,N-1}, p_{1,N-1}, p_{2,N-1}, \Phi_N, \Sigma_{N-1}^{-1}, z_{1,N-1}, \Phi_{1,N}, \Phi_{2,N}).$$

Nous ne donnons pas ici l'expression de $F(.)$ qui, malgré quelques simplifications, reste lourde à écrire et ne présente pas d'intérêt pour la suite de cet article.

3. Résultats asymptotiques

Nous nous focalisons dans cette section sur des résultats de convergence de l'estimateur $\tilde{\theta}_N$ de la direction EDR dans le cas où $H = 2$. Avant de donner ces résultats, nous posons les deux hypothèses suivantes :

(A1) Les observations $(x_i, y_i), i = 1, ..., N$, sont échantillonnées de manière indépendante à partir du modèle (1.1).

(A2) Le support de y est partionné en deux tranches fixes s_1 et s_2 telles que $p_h \neq 0$ pour $h = 1, 2$.

Nous présentons ci-après deux résultats de convergence : le premier précise la vitesse de convergence presque sûre de $\tilde{\theta}_N$ et le second la normalité asymptotique de cet estimateur. Il serait possible d'obtenir de manière similaire des résultats de convergence de λ_N vers λ.

 ISSN: 2102-6238

Theorem 3.1. *Sous les hypothèses* **(H)**, (A_1) *et* (A_2), *nous avons*

$$|| \tilde{\theta}_N - \tilde{\theta} ||= \mathscr{O}\left(\sqrt{\frac{\log(\log N)}{N}}\right) \quad p.s.$$

Theorem 3.2. *Sous les hypothèses* **(H)**, $(A1)$ *et* $(A2)$, *nous avons :*

$$\sqrt{N}(\tilde{\theta}_N - \tilde{\theta}) \xrightarrow{\mathscr{L}} \mathscr{N}(0, \Sigma^{-1}\Delta_3\Sigma^{-1}), \tag{3.1}$$

où l'expression de la matrice Δ_3 *est donnée en (C.4).*

Les démonstrations de ces deux théorèmes figurent respectivement en Annexes B et C.

4. Quelques résultats de simulations

Dans cette partie, nous allons étudier, sur des simulations, le comportement numérique de l'estimateur récursif $\tilde{\theta}_N$ que nous avons proposé lorsque l'on prend $H = 2$ tranches. Cet estimateur $\tilde{\theta}_N$ est un estimateur de la vraie direction de θ du modèle. Nous prendrons alors comme mesure de la qualité de l'estimation la valeur du cosinus carré entre $\tilde{\theta}_N$ et θ défini par

$$\cos^2(\tilde{\theta}_N, \theta) = \frac{(< \tilde{\theta}_N, \theta >)^2}{|| \tilde{\theta}_N || \times || \theta ||},$$

où $|| . ||$ désigne la norme associée au produit scalaire usuel $< ., . >$. Plus le $\cos^2(\tilde{\theta}_N, \theta)$ est proche de 1, meilleure est la qualité de l'estimation.

Nous présentons tout d'abord les modèles simulés et nous précisons les procédures d'estimation mises en œuvre. Nous donnons dans un second temps des résultats de comparaisons des temps de calculs entre les approches SIR récursive et non récursive, comparaison mettant en exergue la rapidité de l'approche récursive. Nous décrivons ensuite des résultats de simulations sur des échantillons particuliers et sur de nombreuses répliques d'échantillons, résultats montrant la convergence de l'estimateur vers la vraie direction θ pour différents jeux de paramètres (modèle, dimension p, matrice Σ, taille N d'échantillon). Nous terminons en illustrant le normalité asymptotique de l'estimateur $\tilde{\theta}_N$ proposé.

Modèles simulés et procédure d'estimation. Dans ces simulations, nous avons considéré les deux modèles de régression suivants :

$$\begin{aligned}(M1): &\quad y = (\theta' x)^3 + \varepsilon,\\ (M2): &\quad y = (\theta' x)\exp(-\theta' x/2) + (\theta' x)\varepsilon,\end{aligned}$$

où x suit la loi multinormale $\mathscr{N}_p(0,\Sigma)$, $\theta = (1,-1,0,\ldots,0)' \in \mathbb{R}^p$ et le terme d'erreur ε, indépendant de x, suit la loi normale $\mathscr{N}(0,1)$. Pour la matrice Σ, nous avons considéré le cas où $\Sigma = I_p$ et le cas où Σ est une matrice symétrique définie positive "quelconque" générée aléatoirement [1]. Nous avons fait varier la dimension p de x : $p = 5, 10, 20$ ou 40. Le modèle $(M1)$ est un modèle

[1] Pour générer une matrice Σ "quelconque", nous avons tout d'abord généré p^2 réalisations d'une loi uniforme sur $[-1,1]$ que nous avons rangées dans une matrice B. Puis nous avons pris $\Sigma = BB' + 0.1I_p$, le second terme permettant d'éviter des problèmes numériques dans l'inversion de Σ.

 ISSN: 2102-6238

homoscédastique ne présentant pas de difficultés particulières pour l'estimation de la direction de θ. Le modèle $(M2)$ est hétéroscédastique et l'estimation de la direction de θ est moins aisée avec des données issues d'un tel modèle. Ces deux modèles sont classiquement utilisés dans les simulations décrites dans la littérature sur la méthode SIR. Nous présentons à la Figure 1 deux nuages de $N = 200$ points $\{(x_i'\theta, y_i), i = 1, \ldots, N\}$ simulés à partir des modèles $(M1)$ et $(M2)$ pour $p = 10$ et $\Sigma = I_p$. Pour l'ensemble des simulations, nous avons choisi un ratio bruit sur signal (défini comme le quotient entre la variance des ε_i et la variance des y_i) d'environ 6% (resp. 25%) pour $(M1)$ (resp. pour $(M2)$), quelle que soit la matrice de variances-covariances Σ (matrice identité ou quelconque) et quelle que soit la dimension p.

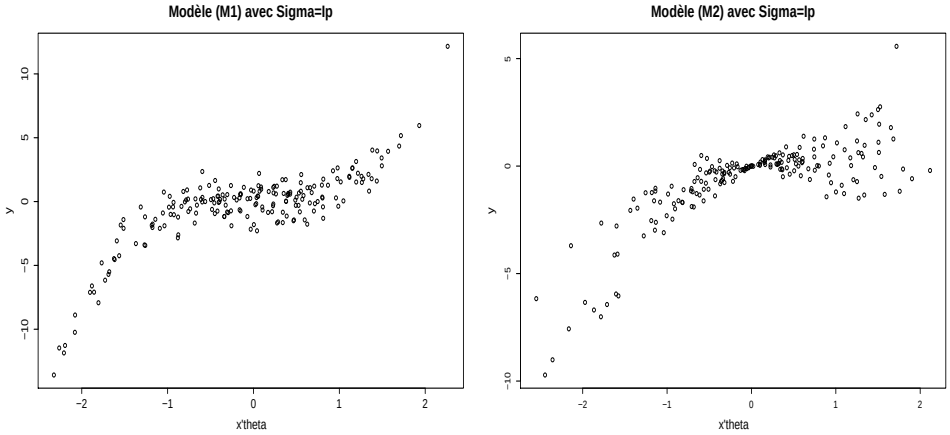

FIGURE 1: Nuage des points $\{(x_i'\theta, y_i), i = 1, \ldots, 200\}$ simulés à partir des modèles $(M1)$ et $(M2)$ pour $p = 10$ et $\Sigma = I_p$

Les simulations ont été faites à l'aide du logiciel R, les codes correspondants sont disponibles auprès des auteurs.

Nous avons simulé, à partir des modèles $(M1)$ et $(M2)$, $\mathscr{B} = 500$ échantillons de taille 1000 avec successivement $p = 5, 10, 20$ et 40, et pour $\Sigma = I_p$ puis pour des matrices Σ "quelconques". Pour chaque jeu de paramètres (modèle, dimension p, matrice Σ) et pour chaque échantillon simulé, nous avons calculé pour $N = N_0$ jusqu'à $N = 1000$, l'estimateur récursif $\tilde{\theta}_N$ ainsi que la qualité correspondante, à savoir $\cos^2(\tilde{\theta}_N, \theta)$, la valeur N_0 étant égale à $p+1$. Les N_0 premiers couples (x_i, y_i) de l'échantillon permettent de calculer les grandeurs nécessaires à l'initialisation de l'algorithme de calculs. Le choix de $N_0 = p+1$ s'explique par le fait que l'on a besoin de calculer la matrice $\Sigma_{N_0-1}^{-1}$ qui est de dimension (p, p). Le seuil permettant d'affecter une nouvelle observation (x_N, y_N) dans la première tranche (lorsque $y_N \leq$ seuil) ou dans la deuxième tranche (si $y_N >$ seuil) a été fixé à la valeur médiane des N_0 premières observations utilisées lors de la phase d'initialisation, ce qui permet automatiquement d'avoir des observations dans chacune des deux tranches. Notons que le choix du seuil n'a pas d'influence sur la qualité de l'estimation pourvu que l'on ait les deux fréquences $p_{h,N}$ non nulles, ceci était déjà vrai pour la version non récursive de l'estimateur SIR.

Comparaison des temps de calculs entre les approches SIR récursive et non récursive. Nous

 ISSN: 2102-6238

avons mesuré pour les approches SIR récursive et SIR non récursive (basées sur $H = 2$ tranches) les temps de calculs en secondes des estimateurs $\tilde{\theta}_n$ de la direction de θ pour n allant de N_0 à N. Dans la Table 1, nous donnons les valeurs des moyennes et des écarts-types calculés sur $\mathscr{B} = 500$ répliques d'échantillons issus du modèle $(M2)$ avec différentes valeurs de p et deux types de matrice Σ (identité ou "quelconque"). Les constats que l'on peut faire suite à la lecture de cette table sont les suivants :

- L'approche SIR récursive est bien plus rapide en moyenne pour calculer les $(N - N_0)$ estimateurs $\tilde{\theta}_n$ que l'approche SIR non récursive. Cet avantage est d'autant plus important que la dimension p de x est grande. Ceci provient très certainement du coût (en terme de temps de calculs) de l'inversion de l'estimateur de la matrice Σ de dimension $p \times p$ suivie du calcul des éléments propres d'une matrice de dimension $p \times p$.
  Lorsque $p = 5$, l'approche récursive est environ 10 fois plus rapide, lorsque $p = 20$, elle est plus de 15 fois plus rapide, et lorsque $p = 40$, elle est plus de 20 fois plus rapide.
- Lorsque la taille d'échantillon augmente, le temps de calcul semble croître proportionnellement pour la méthode SIR récursive et plus que proportionnellement pour l'approche SIR non récursive.
- La matrice de variances-covariances Σ (matrice identité ou quelconque) n'influence pas les temps des calculs des estimateurs.

Des calculs ont été faits de manière identique avec le modèle $(M1)$ et les temps de calculs qui en découlent sont similaires.

			$p = 5$	$p = 20$	$p = 40$
I_p	$N = 300$	SIR classique	1.00 (0.018)	1.99 (0.054)	4.03 (0.012)
		SIR récursif	0.11 (0.001)	0.13 (0.002)	0.17 (0.005)
	$N = 600$	SIR classique	2.01 (0.014)	4.55 (0.093)	10.09 (0.097)
		SIR récursif	0.22 (0.002)	0.26 (0.002)	0.35 (0.013)
	$N = 900$	SIR classique	3.35 (0.022)	7.55 (0.014)	16.57 (0.025)
		SIR récursif	0.33 (0.004)	0.39 (0.004)	0.53 (0.007)
Σ	$N = 300$	SIR classique	0.99 (0.008)	1.97 (0.014)	4.10 (0.014)
		SIR récursif	0.11(0.001)	0.13 (0.001)	0.17 (0.004)
	$N = 600$	SIR classique	2.11 (0.012)	4.57 (0.013)	9.81 (0.020)
		SIR récursif	0.22 (0.001)	0.26 (0.001)	0.35 (0.006)
	$N = 900$	SIR classique	3.35 (0.019)	7.66 (0.019)	16.76 (0.051)
		SIR récursif	0.33 (0.002)	0.39 (0.007)	0.54 (0.009)

TABLE 1

Temps de calculs en secondes des estimateurs $\tilde{\theta}_n$ de la direction de θ (pour n allant de N_0 à N) par la méthode SIR classique et par l'approche SIR récursive : moyennes et écarts-types entre parenthèses (calculés sur $\mathscr{B} = 500$ répliques d'échantillons issus du modèle $(M2)$ avec différentes valeurs de p et deux types matrice Σ (identité ou "quelconque")

Etude de quelques échantillons particuliers avec l'approche SIR récursive. A la Figure 2 (resp. Figure 3), nous considérons tout d'abord des échantillons particuliers simulés à partir du modèle $(M1)$ (resp. du modèle $(M2)$) avec $p = 10$ et $\Sigma = I_p$ ou $\Sigma = \Sigma_1$. Sur ces deux figures, nous avons représenté l'évolution de la mesure de qualité de l'estimateur $\tilde{\theta}_N$ en fonction de N. Nous voyons clairement que plus le nombre d'observations est important, plus la qualité d'estimation

 ISSN: 2102-6238

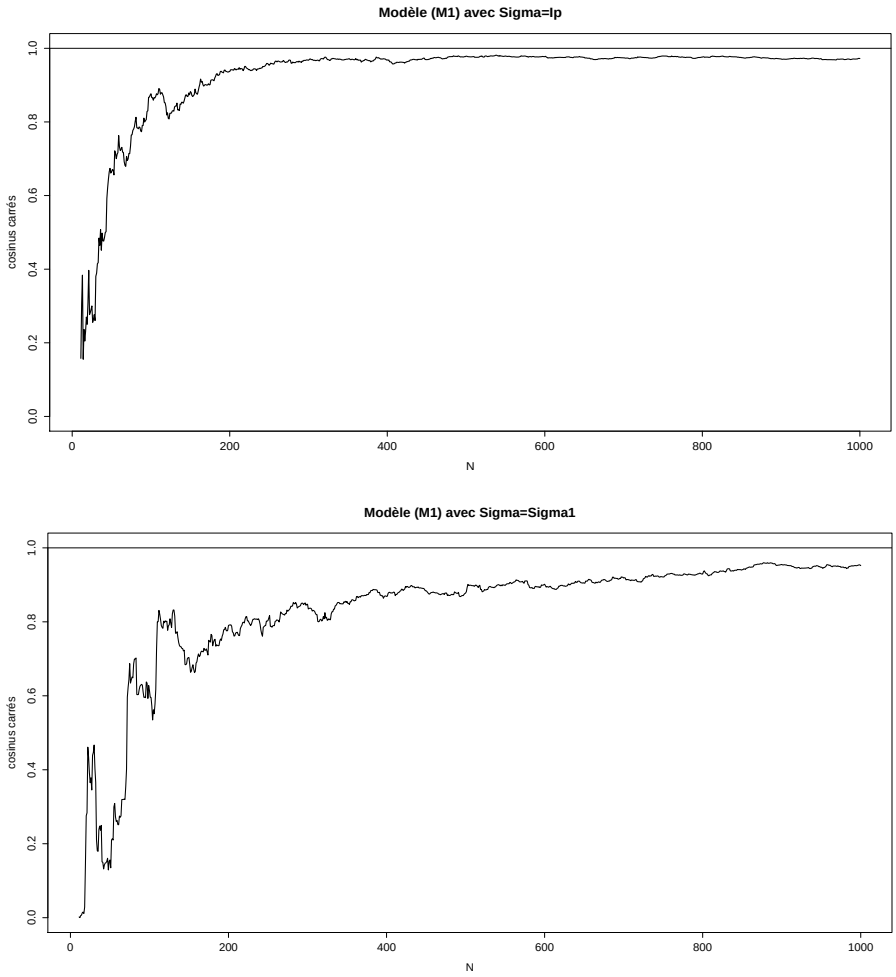

FIGURE 2: Evolution, en fonction de N, de la qualité de l'estimateur $\tilde{\theta}_N$ pour un échantillon issu du modèle $(M1)$ avec $p = 10$ et différentes matrices Σ (= I_p ou Σ_1)

est bonne. Pour le modèle $(M1)$ avec $\Sigma = I_p$, les valeurs du $\cos^2(\tilde{\theta}_N, \theta)$ se rapprochent de 1 très rapidement (voir à la Figure 2 le graphique du haut). Lorsque l'on change de matrice de variances-covariances en prenant $\Sigma = \Sigma_1$, la convergence de $\cos^2(\tilde{\theta}_N, \theta)$ vers 1 est plus lente, mais la qualité d'estimation est tout à fait correcte (supérieure à 0.85) lorsque la taille de l'échantillon N est supérieure 300 (voir à la Figure 2 le graphique du bas). Une explication est que, pour obtenir une bonne estimation de la matrice Σ_1 dont la structure est plus complexe que celle de la matrice I_p, une taille d'échantillon plus importante est nécessaire. A la Figure 3, on observe le même type de phénomèmes que ceux décrits pour le modèle $(M1)$. Il est à noter que le modèle $(M2)$ est plus complexe (modèle hétéroscédastique) que le modèle $(M1)$. Il apparaît ainsi assez naturel que la convergence de $\cos^2(\tilde{\theta}_N, \theta)$ vers 1 soit plus lente avec ces deux échantillons issus du modèle $(M2)$. Il est cependant utile de rappeler que ces graphiques ont été obtenus à partir d'échantillons

 ISSN: 2102-6238

simulés particuliers, et que, pour un même modèle et un même jeu de paramètres, les évolutions des cosinus carrés peuvent être sensiblement différentes pour de petites tailles N d'échantillon. Pour ces raisons-là, nous présentons ci-après une étude sur $\mathscr{B} = 500$ réplications d'échantillons.

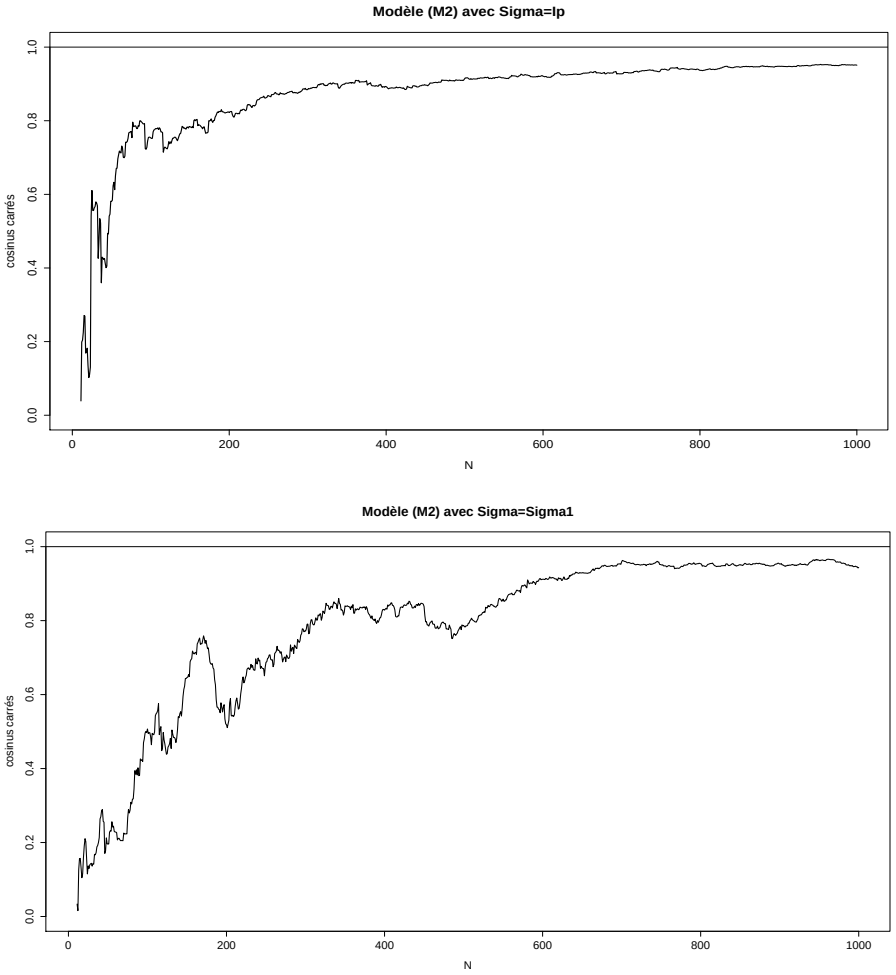

FIGURE 3: Evolution, en fonction de N, de la qualité de l'estimateur $\tilde{\theta}_N$ pour un échantillon issu du modèle $(M2)$ avec $p = 10$ et différentes matrices Σ (= I_p ou Σ_1)

Etude sur $\mathscr{B} = 500$ réplications d'échantillons avec l'approche SIR récursive. Nous avons représenté à la Figure 4 (resp. Figure 5) les boxplots, pour différentes valeurs de N (=50, 100, 150, 200, 300, 500, 700, 1000), des mesures de qualité de $\tilde{\theta}_N$ calculées sur les $\mathscr{B} = 500$ échantillons issus du modèle $(M1)$ (resp. du modèle $(M2)$) avec $p = 10$ et différentes matrices $\Sigma = I_p$ et $\Sigma = \Sigma_1$. Nous observons le même type de phénomème que ceux décrits dans le cas des échantillons particuliers aux Figures 2 et 3. Plus précisément, plus la taille N de l'échantillon est grande, meilleurs sont les estimations (avec des valeurs de $\cos^2(\tilde{\theta}_N, \theta)$ de plus en plus proche de 1 et une dispersion de plus en plus faible). De même, plus la structure de la matrice de variances-

 ISSN: 2102-6238

covariances Σ est complexe, plus il est nécessaire de disposer d'un échantillon de taille raisonnable pour estimer convenablement la direction de θ.

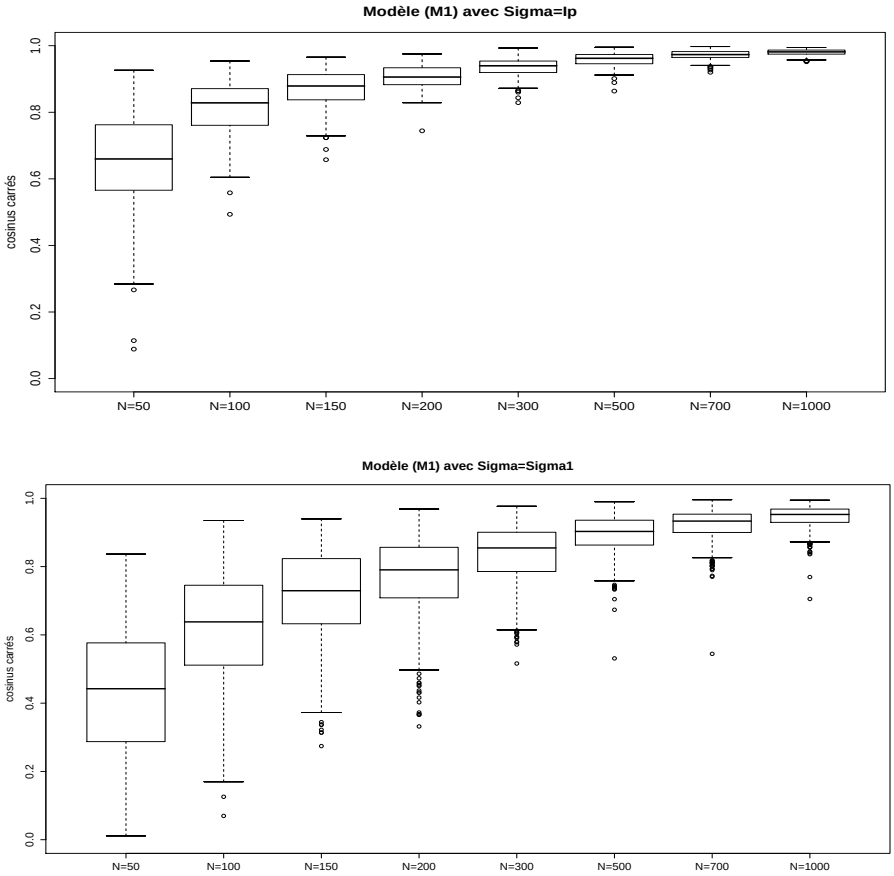

FIGURE 4: Boxplots, pour différentes valeurs de N, des mesures de qualité de $\tilde{\theta}_N$ calculées sur $\mathscr{B} = 500$ échantillons issus du modèle $(M1)$ avec $p = 10$ et différentes matrices Σ (= I_p ou Σ_1)

Etude de l'effet de la dimension p de la covariable x sur la qualité de l'estimation avec l'approche SIR récursive. La dimension p de la covariable a aussi naturellement une certaine influence sur la qualité d'estimation de la direction de θ. Pour illustrer ce point, nous avons représenté à la Figure 6 (resp. Figure 7), pour deux tailles d'échantillon ($N = 500$ et 1000), les boxplots, en fonction de la dimension p de x, des mesures de qualité de $\tilde{\theta}_N$ calculées sur $\mathscr{B} = 500$ échantillons issus du modèle $(M1)$ (resp. modèle $(M2)$) avec différentes matrices Σ (matrice identité ou "quelconque"). On observe clairement que plus la dimension de la covariable est importante, plus les valeurs des cosinus carrés $\cos^2(\tilde{\theta}_N, \theta)$ ont tendance à être plus petites. Ce phénomène est d'autant plus important que la taille N des échantillons est faible et que la structure de la matrice de variances-covariances Σ est complexe.

Illustration de la normalité asymptotique de l'estimateur. A la Figure 8, nous nous intéressons la normalité asymptotique de l'estimateur $\tilde{\theta}_N$. Pour cela, nous avons simulé 3000 réali-

 ISSN: 2102-6238

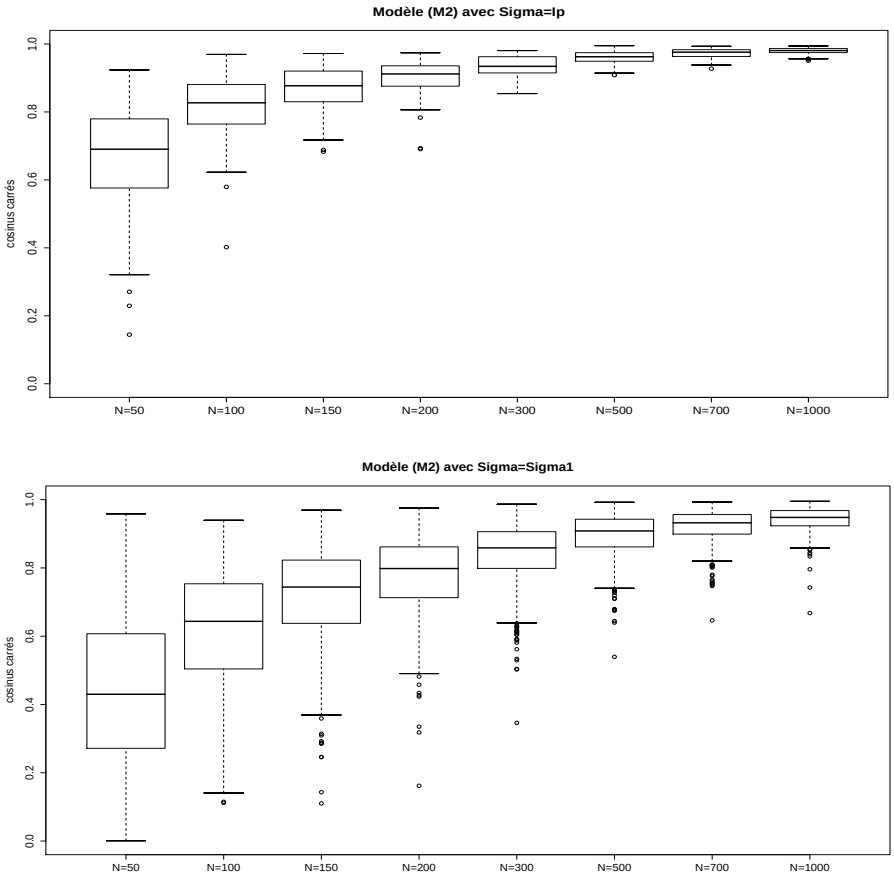

FIGURE 5: Boxplots, pour différentes valeurs de N, des mesures de qualité de $\tilde{\theta}_N$ calculées sur $\mathscr{B} = 500$ échantillons issus du modèle $(M2)$ avec $p = 10$ et différentes matrices Σ (= I_p ou Σ_1)

sations de $\tilde{\theta}_N$ à partir d'échantillons de taille $N = 1000$ issus du modèle $(M2)$ avec $p = 10$ et une matrice de variances-covariances Σ quelconque. Nous avons tracé sur un même graphique la densité de la loi normale centrée réduite (en pointillés) ainsi que l'histogramme des $\left\{ \sqrt{N}\left(\frac{\tilde{\theta}_{N,j}^{(l)} - \theta_j}{\gamma_j}\right),\ l = 1, \ldots, 3000 \right\}$ où $\tilde{\theta}_{N,j}^{(l)}$ correspond à la $j$ème composante de l'estimateur $\tilde{\theta}_N$ calculé sur le $l$ème échantillon, θ_j étant la $j$ème composante de θ, et γ_j le $j$ème élément diagonal à la puissance 1/2 de la matrice de variances-covariances donnée en (3.1). Nous avons choisi arbitrairement quatre composantes θ_j de θ ($j = 1$, 2, 3 et 10). Nous voyons clairement sur les quatre graphiques de la Figure 8 que la densité de la loi $\mathscr{N}(0,1)$ se superpose bien aux différents histogrammes, ce qui illustre bien la normalité asymptotique de l'estimateur proposé. Par souci de concision, nous n'avons pas fourni ici les résultats des simulations obtenus à partir de modèle $(M1)$ ou d'autres matrices de variances-covariances Σ (en particulier $\Sigma = I_p$) ou encore d'autres dimensions p. Pour ces différents cas, nous avons encore observé graphiquement une bonne superposition des histrogrammes avec la densité de la loi normale centrée réduite. Notons que

 ISSN: 2102-6238

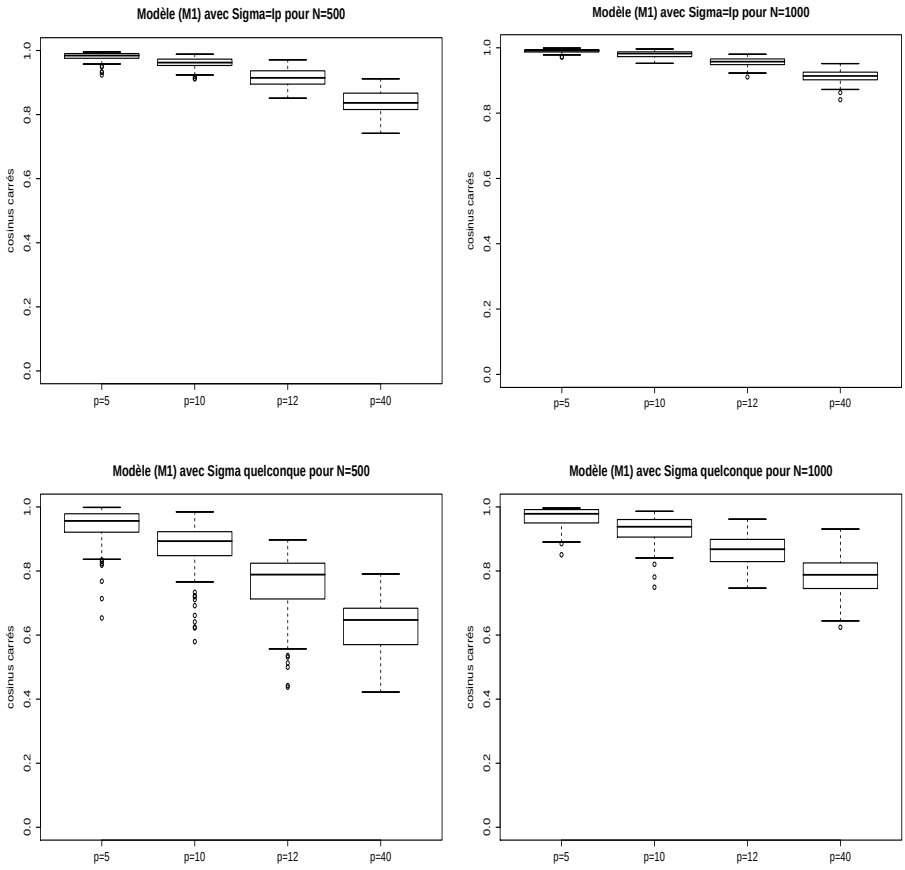

FIGURE 6: Boxplots, en fonction de la dimension p, des mesures de qualité de $\tilde{\theta}_N$ calculées sur $\mathscr{B} = 500$ échantillons issus du modèle $(M1)$ avec différentes matrices Σ (= I_p ou "quelconque"), pour deux tailles d'échantillons N=500 ou 1000

lorsque la dimension p de x est grande, il faut prendre une taille N plus importante (de l'ordre de 2000 pour $p = 40$) pour obtenir une bonne adéquation.

5. Conclusion

Dans cet article, nous nous sommes intéressés à une approche récursive de la méthode SIR dont l'objectif majeur est l'estimation de la direction du vecteur de paramètre θ du modèle (1.1). Dans un premier temps, nous avons donné la forme récursive de l'estimateur de la matrice d'intérêt, $\Sigma_N^{-1}\Gamma_N$, sur laquelle repose la méthode SIR pour un nombre H quelconque de tranches. Dans le cas particulier où l'on considère $H = 2$ tranches, nous avons précisé une expression analytique de l'estimateur $\tilde{\theta}_N$, et nous avons ensuite proposé une forme récursive pour cet estimateur de la direction de θ. Nous avons donné quelques propriétés asymptotiques de cet estimateur dont la convergence presque sûre (avec sa vitesse) et la normalité asymptotique. Nous avons enfin illustré

 ISSN: 2102-6238

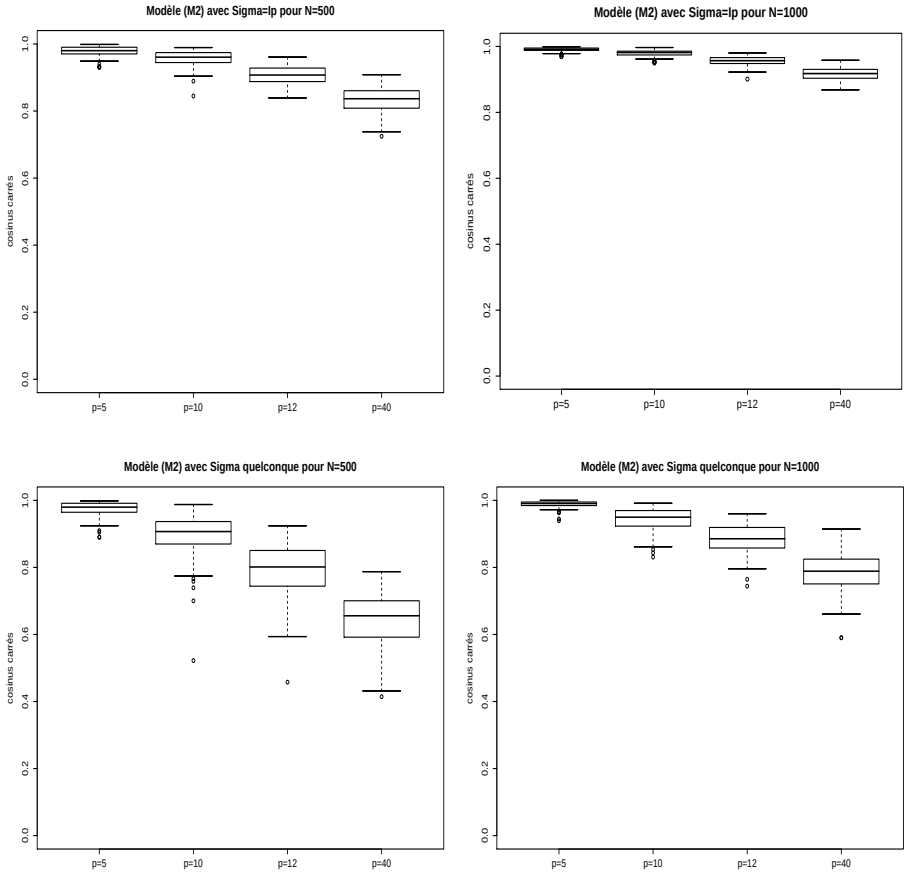

FIGURE 7: Boxplots, en fonction de la dimension p, des mesures de qualité de $\tilde{\theta}_N$ calculés sur $\mathscr{B} = 500$ échantillons issus du modèle $(M2)$ avec différentes matrices Σ (= I_p ou "quelconque"), pour deux tailles d'échantillons N=500 ou 1000

sur des simulations (en considérant différents modèles, différentes dimensions p et différentes matrices de variances-covariances Σ) le bon comportement numérique de l'approche récursive proposée. Un avantage majeur de cette approche récursive est sa rapidité par rapport à la méthode SIR non récursive. Les programmes utilisés ont été écrits en R et sont disponibles auprès des auteurs.

La suite naturelle de ce travail consiste en l'estimation simultanée de la direction de θ et de la fonction de lien f dans le cadre d'un modèle du type $y = f(\theta' x) + \varepsilon$. Pour cela, nous allons considérer l'estimateur de Nadaraya-Watson récursif (voir par exemple [5]) pour estimer la fonction de régression f, et le combiner à l'estimateur récursif $\tilde{\theta}_N$ de θ étudié dans cet article :

$$\hat{f}_N(z) = \frac{1}{\sum_{i=1}^{N} \frac{1}{h_i}\mathbf{K}\left(\frac{z-\hat{z}_i}{h_i}\right)} \sum_{i=1}^{N} \frac{1}{h_i}\mathbf{K}\left(\frac{z-\hat{z}_i}{h_i}\right) y_i,$$

 ISSN: 2102-6238

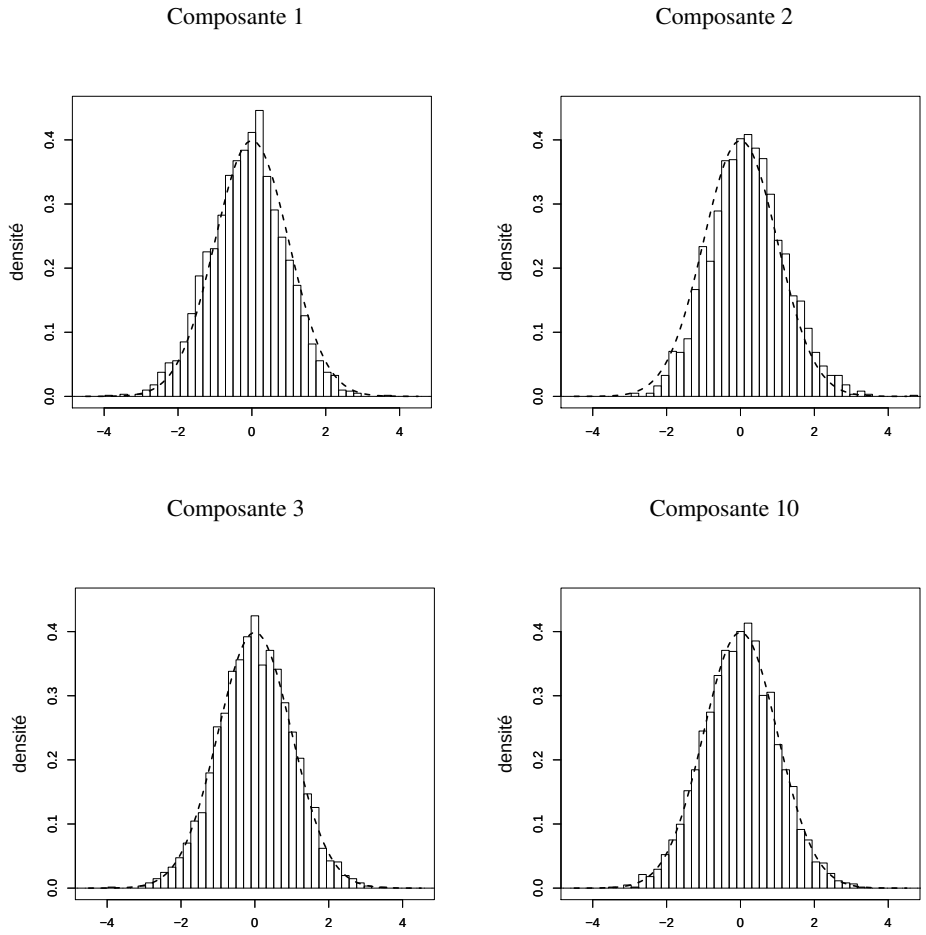

FIGURE 8: Illustration de la normalité asymptotique pour 4 composantes de $\tilde{\theta}_N$ (le graphe de la densité de la loi normale centrée réduite, en pointillé, est superposé à l'histogramme) pour le modèle $(M2)$ avec $p = 10$, $\Sigma = \Sigma_1$ et $N = 1000$

où **K** est un noyau, la suite (h_N) correspond au choix de la largeur de fenêtre, et où z et $\hat{z}_i$ sont respectivement les indices $z = \theta' x$ et $\hat{z}_i = \tilde{\theta}_i' x_i$. Ce travail est en cours d'investigation. Une autre extension possible et intéressante de ce travail est de sortir du cadre i.i.d. en considérant un cadre dépendant, comme par exemple le cadre autorégressif. Cette extension n'est pas directe et va nécessiter une nouvelle formulation de l'hypothèse de linéarité **(H)**. Notons qu'à notre connaissance, ceci n'a pas encore été étudié dans la littérature pour la version non récursive de la méthode SIR.

Remerciements. Qu'il nous soit permis de remercier l'Éditeur en Chef du Journal de la SFdS, ainsi que les deux relecteurs anonymes : leurs commentaires, leurs critiques et leurs suggestions constructives nous ont permis d'améliorer substantiellement la qualité de cet article.

 ISSN: 2102-6238

Références

[1] Chen, C .H. et Li, K. C. (1998). Can SIR be as popular as multiple linear regression ? *Statistica Sinica*, **8**, 289-316.

[2] Cook, R. D. et Nachtsheim, C. J. (1994). Re-weighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association*, **89**, 592-599.

[3] Duan, N. and Li, K. C. (1991). Slicing regression : a link-free regression method. *The Annals of Statistics*, **19**, 505-530.

[4] Duflo, M. (1990). *Méthodes récursives aléatoires*. Techniques Stochastiques, Masson, Paris.

[5] Duflo, M. (1997). *Random Iterative Models*, Springer-Verlag, Berlin.

[6] Hall, P. et Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, **21**, 867-889.

[7] Li, K. C. (1991). Sliced inverse regression for dimension reduction, with dicussion. *Journal of the American Statistical Association*, **86**, 316-342.

[8] Magnus, J. R. and Neudecker, H. (1979). The commutation matrix : some properties and applications. *Annals of Statistics*, *7*(2), 381-394.

[9] Magnus, J. R. and Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

[10] Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Communications in Statistics - Theory and methods*, **26**, 2141-2171.

[11] Stout, W.F. (1974). *Almost sure convergence* A Series of Monographs and Textbooks in Probability and Mathematical Statistics.

 ISSN: 2102-6238

Annexe A: expression analytique du vecteur propre $\tilde{\theta}$ et de la valeur propre λ de $\Sigma^{-1}\Gamma$ pour $H = 2$

Démonstration. Lorsque $H = 2$, nous avons

$$p_1 z_1 + p_2 z_2 = p_1 m_1 + p_2 m_2 - (p_1 + p_2)\mu = p_1 m_1 + p_2 m_2 - \mu = \mu - \mu = 0. \tag{A.1}$$

Vérifions que $\tilde{\theta} = \Sigma^{-1}(z_1 - z_2)$ est bien vecteur propre de $\Sigma^{-1}\Gamma$ et déterminons la valeur propre associée. En utilisant (2.1), on a

$$\begin{aligned} \Sigma^{-1}\Gamma\tilde{\theta} &= \Sigma^{-1}\Gamma\Sigma^{-1}(z_1 - z_2) \\ &= q_1 p_1 \Sigma^{-1} z_1 - q_{12} p_1 \Sigma^{-1} z_1 + q_{12} p_2 \Sigma^{-1} z_2 - q_2 p_2 \Sigma^{-1} z_2, \end{aligned}$$

où $q_1 = z_1'\Sigma^{-1}z_1$, $q_2 = z_2'\Sigma^{-1}z_2$ et $q_{12} = z_1'\Sigma^{-1}z_2 = z_2'\Sigma^{-1}z_1$.
De plus, en utilisant (A.1), nous avons : $p_1 q_1 = -p_2 q_{12}$ et $p_2 q_2 = -p_1 q_{12}$. Nous en déduisons que

$$\begin{aligned} \Sigma^{-1}\Gamma\tilde{\theta} &= p_2 q_{12}\Sigma^{-1}(z_2 - z_1) + p_1 q_{12}\Sigma^{-1}(z_2 - z_1) \\ &= -q_{12}\Sigma^{-1}(z_1 - z_2). \end{aligned}$$

Or en utilisant (A.1) à nouveau, nous avons : $-q_{12} = \frac{p_1}{p_2} z_1'\Sigma^{-1}z_1$. Finalement, en posant $\lambda = \frac{p_1}{p_2} z_1'\Sigma^{-1}z_1$, nous avons : $\Sigma^{-1}\Gamma\tilde{\theta} = \lambda\tilde{\theta}$. Le vecteur $\tilde{\theta}$ est donc bien vecteur propre associé à la valeur propre λ de la matrice $\Sigma^{-1}\Gamma$. □

Annexe B: démonstration du Théorème 3.1

Nous rappelons ici tout d'abord le théorème de Harman-Winter (appelé aussi loi du logarithme itéré de Hartman-Wintner, voir Stout [11] page 136 pour une démonstration de ce théorème).
Soit $\{Z_i, i \geq 1\}$ *une suite de variables aléatoires indépendantes et identiquement distribuées. Notons* $T_n = \sum_{i=1}^{n} Z_i$ *pour* $n \geq 1$, *si* $\mathbb{E}Z_1^2 < \infty$ *et* $\mathbb{E}Z_1 = 0$, *alors*

$$\limsup\ T_n/(2n\mathbb{E}Z_1^2\log(\log n))^{1/2} = 1 \quad p.s.$$

Démonstration. Lorsque $H = 2$, nous avons $\tilde{\theta}_N = \Sigma_N^{-1}(z_{1,N} - z_{2,N})$ et $\tilde{\theta} = \Sigma^{-1}(z_1 - z_2)$. On en déduit que

$$\tilde{\theta}_N - \tilde{\theta} = (\Sigma_N^{-1} - \Sigma^{-1})z_{1,N} - (\Sigma_N^{-1} - \Sigma^{-1})z_{2,N} + \Sigma^{-1}(z_{1,N} - z_1) - \Sigma^{-1}(z_{2,N} - z_2).$$

Nous obtenons alors la majoration suivante :

$$\begin{aligned} \| \tilde{\theta}_N - \tilde{\theta} \|^2 \leq\ & 4 \| z_{1,N} \|^2 \| \Sigma_N^{-1} - \Sigma^{-1} \|^2 + 4 \| z_{2,N} \|^2 \| \Sigma_N^{-1} - \Sigma^{-1} \|^2 \\ & + 4 \| \Sigma^{-1} \|^2 \| z_{1,N} - z_1 \|^2 + 4 \| \Sigma^{-1} \|^2 \| z_{2,N} - z_2 \|^2 . \end{aligned} \tag{B.1}$$

 ISSN: 2102-6238

Dans la suite de la démonstration, nous obtenons tout d'abord la vitesse de convergence de $(z_{h,N})$ en précisant la vitesse de convergence de $(\bar{x}_N)$ et de $(m_{h,N})$ où $h = 1,2$. Puis en appliquant l'équation de Riccati pour la matrice inverse, nous obtenons la vitesse de convergence de (Σ_N^{-1}). Enfin, nous en déduisons la vitesse de convergence de $(\tilde{\theta}_N)$.

Pour une matrice M, nous noterons par $\lambda_{max}(M)$ sa plus grande valeur propre et par $\lambda_{min}(M)$ sa plus petite valeur propre.

Étape 1 : étude de la vitesse de convergence de $(z_{h,N})$

Pour $h = 1,2$, nous avons :

$$z_{h,N} - z_h = (m_{h,N} - m_h) - (\bar{x}_N - \mu).$$

Utilisons la majoration suivante :

$$|| z_{h,N} - z_h ||^2 \leq 2 || m_{h,N} - m_h ||^2 + 2 || (\bar{x}_N - \mu) ||^2 \tag{B.2}$$

Précisons tout d'abord la vitesse de convergence de $(\bar{x}_N)$ et celle de $(m_{h,N})$.

Étude de la vitesse de convergence de $(\bar{x}_N)$

Nous pouvons déduire de la loi du logarithme itéré de Hartman-Wintner (le théorème 3.2.9 page 136 de Stout [11] que

$$\limsup \frac{|| S_N - N\mu ||^2}{2N\log(\log N)} \leq \lambda_{max}(\Sigma) \quad \text{p.s.}$$

où $S_N = \sum_{i=1}^{N} x_i$. Il en découle que

$$\limsup \left(\frac{N}{2\log(\log N)} \right) || \bar{x}_N - \mu ||^2 \leq \lambda_{max}(\Sigma) \quad \text{p.s.}$$

Ce qui entraîne

$$|| \bar{x}_N - \mu ||^2 = \mathcal{O}\left(\frac{\log(\log N)}{N} \right) \quad \text{p.s.} \tag{B.3}$$

Étude de la vitesse de convergence de $m_{h,N}$

Nous avons

$$m_{h,N} - m_h = m_{h,N} - \frac{N}{N_h}\mathbb{E}[X\mathbb{I}_{[Y\in s_h]}] + \frac{N}{N_h}\mathbb{E}[X\mathbb{I}_{[Y\in s_h]}] - m_h$$

où on rappelle que $m_{h,N} = \frac{1}{N_h}\sum_{i=1}^{N} x_i \mathbb{I}_{[y_i \in s_h]}$ et $N_h = \sum_{i=1}^{N} \mathbb{I}_{[y_i \in s_h]}$.

En majorant, nous obtenons

$$\|m_{h,N} - m_h\|^2 \leq 2 \left\| m_{h,N} - \frac{N}{N_h}\mathbb{E}[X\mathbb{I}_{[Y\in s_h]}] \right\|^2 + 2 \left\| \frac{N}{N_h}\mathbb{E}[X\mathbb{I}_{[Y\in s_h]}] - m_h \right\|^2. \tag{B.4}$$

Intéressons-nous tout d'abord au premier terme de la majoration. En remarquant que

$$N_h m_{h,N} = \sum_{i=1}^{N} x_i \mathbb{I}_{[y_i \in s_h]} = \sum_{i=1}^{N} \varepsilon_i(h) + N\mathbb{E}[X\mathbb{I}_{[Y\in s_h]}])$$

 ISSN: 2102-6238

où $\varepsilon_i(h) = x_i \mathbb{I}_{[y_i \in s_h]} - \mathbb{E}[x_i \mathbb{I}_{[y_i \in s_h]}] = x_i \mathbb{I}_{[y_i \in s_h]} - \mathbb{E}[X \mathbb{I}_{[Y \in s_h]}]$, nous avons

$$M_N(h) = N_h m_{h,N} - N\mathbb{E}[X \mathbb{I}_{[Y \in s_h]}] = \sum_{i=1}^{N} \varepsilon_i(h).$$

La suite $(\varepsilon_N(h))$ est constituée de vecteurs aléatoires indépendantes et de même loi avec $\mathbb{E}[\varepsilon_N(h)] = 0$ et

$$\mathbb{E}[\varepsilon_N(h)\varepsilon_N'(h)] \leq \Sigma + \mu\mu'.$$

Nous déduisons à nouveau de la loi du logarithme itéré de Hartman-Wintner que

$$\left\| M_N(h) \right\|^2 = \mathcal{O}\Big(N \log(\log N)\Big) \quad \text{p.s.}$$

Par suite

$$\left\| N_h m_{h,N} - N\mathbb{E}[X \mathbb{I}_{[Y \in s_h]}] \right\|^2 = \mathcal{O}\Big(N \log(\log N)\Big) \quad \text{p.s.}$$

donc

$$\left\| m_{h,N} - \frac{N}{N_h} \mathbb{E}[X \mathbb{I}_{[Y \in s_h]}] \right\|^2 = \mathcal{O}\left(\frac{N}{N_h^2} \log(\log N)\right) \quad \text{p.s.}$$

De plus, nous avons également

$$\frac{N_h}{N} \longrightarrow p_h \text{ p.s.} \quad \text{avec} \quad p_h = \mathbb{P}(Y \in s_h). \tag{B.a}$$

Sous l'hypothèse (A2) : $p_1 > 0$ et $p_2 > 0$. Donc $N_h \sim p_h N$ et nous obtenons finalement

$$\left\| m_{h,N} - \frac{N}{N_h} \mathbb{E}[X \mathbb{I}_{[Y \in s_h]}] \right\|^2 = \mathcal{O}\Big(\frac{\log(\log N)}{N}\Big) \quad \text{p.s.} \tag{B.5}$$

Étudions maintenant le second terme de la majoration. Comme $m_h p_h = \mathbb{E}[X \mathbb{I}_{[Y \in s_h]}]$, nous avons

$$\frac{N}{N_h} \mathbb{E}[X \mathbb{I}_{[Y \in s_h]}] - m_h = \frac{N}{N_h} m_h p_h - m_h = \frac{m_h}{N_h}\Big(N p_h - N_h\Big). \tag{B.b}$$

En outre, comme $N_h = \sum_{i=1}^{N} Z_i(h) + N p_h$, où $Z_i(h) = \mathbb{I}_{[y_i \in s_h]} - \mathbb{E}[\mathbb{I}_{[y_i \in s_h]}]$, nous avons

$$N_h - N p_h = M_N = \sum_{i=1}^{N} Z_i(h).$$

D'après la loi du logarithme itéré de Hartman-Wintner, nous avons

$$\left\| N_h - N p_h \right\|^2 = \mathcal{O}\Big(N \log(\log N)\Big) \quad \text{p.s.}$$

et en utilisant le fait que $N_h \sim p_h N$ (d'après (B.a)), nous déduisons de (B.b) que

$$\left\| \frac{N}{N_h} \mathbb{E}[X \mathbb{I}_{[Y \in s_h]}] - m_h \right\|^2 = \mathcal{O}\left(\frac{\log(\log N)}{N}\right) \quad \text{p.s.} \tag{B.6}$$

 ISSN: 2102-6238

Finalement, en utilisant les vitesses obtenues en (B.5) et (B.6), nous déduisons de (B.4) que

$$|| m_{h,N} - m_h ||^2 = \mathcal{O}\Big(\frac{\log(\log N)}{N}\Big) \quad \text{p.s.} \tag{B.7}$$

Obtention de la vitesse de $(z_{h,N})$
À partir des vitesses de convergence données en (B.3) et (B.7), nous trouvons par (B.2) que

$$|| z_{h,N} - z_h ||^2 = \mathcal{O}\Big(\frac{\log(\log N)}{N}\Big) \quad \text{p.s.} \tag{B.8}$$

Étape 2 : étude de la vitesse de convergence de (Σ_N^{-1})
D'après l'équation de Riccati pour l'inversion de matrice donnée à la page 96 de Duflo [5], nous pouvons écrire Σ_N^{-1} sous la forme

$$\Sigma_N^{-1} = \Sigma^{-1} - \Sigma^{-1}(\Sigma_N - \Sigma)\Sigma^{-1} + R_N$$

où $R_N = \Sigma^{-1}(\Sigma - \Sigma_N)\Sigma_N^{-1}(\Sigma - \Sigma_N)\Sigma^{-1}$.
Il est alors immédiat que

$$|| \Sigma_N^{-1} - \Sigma^{-1} ||^2 \leq 2 || R_N ||^2 + 2 || \Sigma^{-1}(\Sigma_N - \Sigma)\Sigma^{-1} ||^2 . \tag{B.9}$$

Étude de la vitesse de convergence de (Σ_N)
Comme

$$\begin{aligned} \Sigma_N &= \tfrac{1}{N}\textstyle\sum_{i=1}^N (x_i - \bar{x}_N)(x_i - \bar{x}_N)' \\ &= \tfrac{1}{N}\textstyle\sum_{i=1}^N (x_i - \mu)(x_i - \mu)' - (\bar{x}_N - \mu)(\bar{x}_N - \mu)' \end{aligned}$$

nous avons

$$\Sigma_N - \Sigma = \frac{1}{N}\sum_{i=1}^N e_i - (\bar{x}_N - \mu)(\bar{x}_N - \mu)',$$

où $e_i = (x_i - \mu)(x_i - \mu)' - \Sigma$.
Par conséquent, nous obtenons

$$|| \Sigma_N - \Sigma ||^2 \leq \frac{2}{N^2}\Big\| \sum_{i=1}^N e_i \Big\|^2 + 2 || \bar{x}_N - \mu ||^4 . \tag{B.10}$$

Posons $M_N(u) = u'\Big(\sum_{i=1}^N e_i\Big)u = \sum_{i=1}^N u' e_i u = \sum_{i=1}^N e_i(u)$, où $e_i(u) = u' e_i u = u'(x_i - \mu)(x_i - \mu)'u - u'\Sigma u$.
Nous avons

$$\mathbb{E}[e_i(u)] = \mathbb{E}[(u'(x_i - \mu)(x_i - \mu)'u - u'\Sigma u] = u'\Sigma u - u'\Sigma u = 0.$$

Nous avons également

$$\begin{aligned} \mathbb{E}[(e_i(u))^2] &= \mathbb{E}[(u'(x_i - \mu)(x_i - \mu)'u - u'\Sigma u)^2] \\ &= \mathbb{E}[u'(x_i - \mu)(x_i - \mu)'uu'(x_i - \mu)(x_i - \mu)'u] - (u'\Sigma u)^2 \\ &\leq \mathbb{E}[(u'(x_i - \mu)(x_i - \mu)'u)^2] \\ &\leq \mathbb{E}\Big[|| x_i - \mu ||^4 \Big] \times || u ||^4 . \end{aligned}$$

 ISSN: 2102-6238

Nous pouvons déduire une nouvelle fois par la loi du logarithme itéré de Hartman-Wintner que

$$\limsup \frac{|| \sum_{i=1}^{N} e_i(u) ||^2}{2N\log(\log N)} \leq \tau^4 \times || u ||^4 \quad \text{p.s.}$$

où $\tau^4 = \mathbb{E}\Big[|| x_i - \mu ||^4 \Big]$.
Ceci étant vrai pout tout u, il en découle que

$$\Big\| \sum_{i=1}^{N} e_i \Big\|^2 = \mathscr{O}\Big(N\log(\log N)\Big) \quad \text{p.s.} \tag{B.11}$$

À partir des vitesses de convergence obtenues en (B.3) et (B.11), nous déduisons de (B.10) que

$$\begin{aligned} || \Sigma_N - \Sigma ||^2 &= \mathscr{O}\left(\frac{1}{N^2} N\log(\log N)\right) + \mathscr{O}\left(\left(\frac{\log(\log N)}{N}\right)^2\right) \quad \text{p.s.} \\ &= \mathscr{O}\left(\frac{\log(\log N)}{N}\right) \quad \text{p.s.} \end{aligned} \tag{B.12}$$

Étude de la vitesse de convergence de (R_N)
Nous déduisons de l'expression de R_N que

$$\begin{aligned} || R_N || &\leq \lambda_{max}(\Sigma_N^{-1})\Sigma^{-1}(\Sigma - \Sigma_N)^2\Sigma^{-1} \\ &\leq \frac{1}{\lambda_{min}(\Sigma_N)}\Sigma^{-1}(\Sigma - \Sigma_N)^2\Sigma^{-1} \\ &\leq \frac{1}{\lambda_{min}(\Sigma_N)} \times \frac{||\Sigma_N - \Sigma||^2}{(\lambda_{min}(\Sigma))^2}. \end{aligned}$$

À partir de (B.12), nous obtenons

$$|| R_N || = \mathscr{O}\left(\frac{\log(\log N)}{N}\right) \quad \text{p.s.} \tag{B.13}$$

Obtention de la vitesse de (Σ_N^{-1})
À partir des vitesses de convergence obtenues en (B.12) et (B.13), il est immédiat par (B.9) que

$$|| \Sigma_N^{-1} - \Sigma^{-1} ||^2 = \mathscr{O}\left(\frac{\log(\log N)}{N}\right) \quad \text{p.s.} \tag{B.14}$$

Étape 3 : obtention de la vitesse de convergence de $\tilde{\theta}_N$
Finalement, à partir des vitesses de convergence obtenues en (B.8) et (B.14), nous déduisons de (B.1) que

$$|| \tilde{\theta}_N - \tilde{\theta} ||^2 = \mathscr{O}\left(\frac{\log(\log N)}{N}\right) \quad \text{p.s.}$$

d'où

$$|| \tilde{\theta}_N - \tilde{\theta} || = \mathscr{O}\left(\sqrt{\frac{\log(\log N)}{N}}\right) \quad \text{p.s.}$$

□

Ce qui achève la preuve du Théorème 3.1.

 ISSN: 2102-6238

Annexe C: démonstration du Théorème 3.2

Nous donnons ici tout d'abord quelques rappels sur la transformation de vectorisation, "vec", ainsi que sur le produit de Kronecker.

Soit $M = [m_1, \ldots, m_p]$ une matrice de dimensions $q \times p$ où les m_k , $k = 1, ..., p$, sont des vecteurs colonnes de dimension q. Alors $\text{vec}(M)$ est la transformation de M en un vecteur de dimension qp de la manière suivante :

$$\text{vec}(M) = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_p \end{pmatrix}.$$

Soient M et N deux matrices de dimensions respectives $q \times p$ et $r \times s$, alors le produit de Kronecker de M et N est la matrice partitionnée de dimensions $qr \times ps$ définie par : $M \otimes N = [m_{j,k}N]$ avec $j = 1, ..., q$ et $k = 1, ..., p$.
Une propriété importante reliant le produit de Kronecker et la transformation "vec" est :

$$vec(ABC) = (C' \otimes A)vec(B)$$

où les dimensions des matrices A, B et C sont telles que le produit matriciel est bien défini. D'autres propriétés sur la transformation "vec" et le produit de Kronecker peuvent être trouvées dans Magnus et Neudecker [8] , [9].

Démonstration. Pour débuter la démonstration, nous rappelons que $\tilde{\theta}_N = \Sigma_N^{-1}(z_{1,N} - z_{2,N})$ et $\tilde{\theta} = \Sigma^{-1}(z_1 - z_2)$. Nous cherchons la loi asymptotique de $\sqrt{N}(\tilde{\theta}_N - \tilde{\theta}) = \Sigma_N^{-1} Z$, où $Z = \sqrt{N}(z_{1,N} - z_{2,N} - \Sigma_N \tilde{\theta})$.
En utilisant la définition de $z_{1,N}$, $z_{2,N}$, z_1 et z_2, nous avons :

$$\begin{aligned} Z &= \sqrt{N}\{m_{1,N} - m_{2,N} - \Sigma_N \Sigma^{-1}(m_1 - m_2)\} \\ &= \sqrt{N}\{m_{1,N} - m_{2,N} - (\Sigma_N - \Sigma + \Sigma)\Sigma^{-1}(m_1 - m_2) \\ &= \sqrt{N}\{(m_{1,N} - m_1) - (m_{2,N} - m_2) - (\Sigma_N - \Sigma)\Sigma^{-1}(m_1 - m_2)\}. \end{aligned}$$

En utilisant la propriété reliant le produit de Kronecker et la transformation "vec", nous avons :

$$\begin{aligned} (\Sigma_N - \Sigma)\Sigma^{-1}(m_1 - m_2) &= \text{vec}(I_p(\Sigma_N - \Sigma)\Sigma^{-1}(m_1 - m_2)) \\ &= ((m_1 - m_2)'\Sigma^{-1} \otimes I_p)\text{vec}(\Sigma_N - \Sigma). \end{aligned}$$

Finalement, nous obtenons :

$$Z = \sqrt{N}\{(m_{1,N} - m_1) - (m_{2,N} - m_2) - D^* \text{vec}(\Sigma_N - \Sigma)\}, \tag{C.1}$$

où $D^* = (m_1 - m_2)'\Sigma^{-1} \otimes I_p$.

Dans la suite de la démonstration, nous obtenons tout d'abord la loi de Z en deux étapes en utilisant le théorème central limite et la Delta méthode. Puis, dans une troisième étape, nous en déduirons la loi de $\sqrt{N}(\tilde{\theta}_N - \tilde{\theta})$.

Étape 1 : application du théorème central limite

 ISSN: 2102-6238

Notons U_i le vecteur aléatoire de dimension $(2+2p+p+p^2)$ défini par :

$$U_i = (\mathbb{I}_{1(i)}, \mathbb{I}_{2(i)}, x_i' \mathbb{I}_{1(i)}, x_i' \mathbb{I}_{2(i)}, x_i', \text{vec}(x_i x_i')')',$$

où $\mathbb{I}_{h(i)} := \mathbb{I}[y_i \in s_h]$, avec $h = 1,2$.

Sous les hypothèses du théorème, les vecteurs U_i, $i = 1,...,N$ sont indépendants et identiquement distribués d'espérance μ_U et de matrice de covariances Σ_U de la forme :

$$\mu_U = \mathbb{E}[U_i] = (p_1, p_2, \widetilde{m}_1', \widetilde{m}_2', \mu', \text{vec}(\Sigma + \mu\mu')')',$$

et

$$\Sigma_U = \mathbb{V}(U_i) = \begin{pmatrix} B_{11} & B_{12} & B_{13} & B_{14} \\ B_{12}' & B_{22} & B_{23} & B_{24} \\ B_{13}' & B_{23}' & B_{33} & B_{34} \\ B_{14}' & B_{24}' & B_{34}' & B_{44} \end{pmatrix},$$

où

$\widetilde{m}_1 := \mathbb{E}[x\mathbb{I}[y \in s_1]]$ et $\widetilde{m}_2 := \mathbb{E}[x\mathbb{I}[y \in s_2]]$,

$$B_{11} = \begin{bmatrix} p_1(1-p_1) & -p_2 p_1 \\ -p_1 p_2 & p_2(1-p_2) \end{bmatrix},$$

$$B_{12} = \begin{bmatrix} (1-p_1)\widetilde{m}_1' & -p_1\widetilde{m}_2' \\ -p_2\widetilde{m}_1' & (1-p_2)\widetilde{m}_2' \end{bmatrix},$$

$$B_{13} = \begin{bmatrix} \widetilde{m}_1' - p_1\mu' \\ \widetilde{m}_2' - p_2\mu' \end{bmatrix},$$

$$B_{14} = \begin{bmatrix} \mathbb{E}[(x' \otimes x')\mathbb{I}[y \in s_1]] - p_1 \text{vec}(\Sigma + \mu\mu')' \\ \mathbb{E}[(x' \otimes x')\mathbb{I}[y \in s_2]] - p_2 \text{vec}(\Sigma + \mu\mu')' \end{bmatrix},$$

$$B_{22} = \begin{bmatrix} \mathbb{E}[xx'\mathbb{I}[y \in s_1]] - \widetilde{m}_1\widetilde{m}_1' & -\widetilde{m}_1\widetilde{m}_2' \\ -\widetilde{m}_2\widetilde{m}_1' & \mathbb{E}[xx'\mathbb{I}[y \in s_2]] - \widetilde{m}_2\widetilde{m}_2' \end{bmatrix},$$

$$B_{23} = \begin{bmatrix} \mathbb{E}[xx'\mathbb{I}[y \in s_1]] - \widetilde{m}_1\mu' \\ \mathbb{E}[xx'\mathbb{I}[y \in s_2]] - \widetilde{m}_2\mu' \end{bmatrix},$$

$$B_{24} = \begin{bmatrix} \mathbb{E}[x(x' \otimes x')\mathbb{I}[y \in s_1]] - \widetilde{m}_1 \text{vec}(\Sigma + \mu\mu')' \\ \mathbb{E}[x(x' \otimes x')\mathbb{I}[y \in s_2]] - \widetilde{m}_2 \text{vec}(\Sigma + \mu\mu')' \end{bmatrix},$$

$$B_{33} = \Sigma,$$

$$B_{34} = \mathbb{E}[x(x' \otimes x')] - \mu \text{vec}(\Sigma + \mu\mu')'.$$

$$B_{44} = \mathbb{E}[(xx') \otimes (xx')] - \text{vec}(\Sigma + \mu\mu')\text{vec}(\Sigma + \mu\mu')',$$

Définissons le vecteur $\overline{U} = \frac{1}{N}\sum_{i=1}^N U_i = (p_{1,N}, p_{2,N}, \widetilde{m}_{1,N}', \widetilde{m}_{2,N}', \bar{x}', \frac{1}{N}\sum_{i=1}^N \text{vec}(x_i x_i')')'$ où $\widetilde{m}_{h,N} = \frac{1}{N}\sum_{i=1}^N x_i \mathbb{I}_{h(i)}$ pour $h = 1,2$. Par le théorème central limite multidimensionnel, nous obtenons la distribution asymptotique suivante

$$\sqrt{N}(\overline{U} - \mu_U) \xrightarrow{\mathscr{L}} \mathscr{N}_{2+3p+p^2}(0, \Sigma_U).$$

 ISSN: 2102-6238

Étape 2 : application de la Delta méthode
A partir de $\overline{U}$, nous formons le vecteur colonne

$$\overline{U}_1 = (p_{1,N}, p_{2,N}, m'_{1,N}, m'_{2,N}, \vec{x}', \frac{1}{N}\sum_{i=1}^{N} \mathrm{vec}(x_i x_i')')',$$

où nous rappelons que $m_{h,N} = \frac{1}{N p_{h,N}} \sum_{i=1}^{N} x_i \mathbb{I}_{h(i)} = \frac{1}{p_{h,N}} \widetilde{m}_{h,N}$ pour $h = 1,2$. Par la Delta méthode, nous allons déduire la distribution asymptotique de $\overline{U}_1$ à partir de celle de $\overline{U}$.

Soit la fonction h_1 définie par

$$\begin{array}{rcccc} h_1 & : & \mathbb{R}^{2+3p+p^2} & \longrightarrow & \mathbb{R}^{2+3p+p^2} \\ & & u = (a_1, a_2, b'_1, b'_2, c', d')' & \longmapsto & (a_1, a_2, b'_1/a_1, b'_2/a_2, c', d')' \end{array}$$

En remarquant que $\overline{U}_1 = h_1(\overline{U})$ et que $\mu_{U_1} := h_1(\mu_U) = (p_1, p_2, m'_1, m'_2, \mu', \mathrm{vec}(\Sigma + \mu\mu')')$, nous obtenons par la Delta méthode

$$\sqrt{N}(h_1(\bar{U}) - h_1(\mu_U)) \xrightarrow{\mathscr{L}} \mathscr{N}_{2+3p+p^2}(0, J_1 \Sigma_U J_1') \tag{C.2}$$

avec $J_1 = \partial h_1 / \partial u \,|_{\mathbb{E}}$, où la notation $g\,|_{\mathbb{E}}$ désigne la valeur de la fonction g en l'espérance de son argument.
Notons I_s la matrice identité de dimensions $s \times s$ et $0_{s_1,s_2}$ la matrice nulle de dimension $s_1 \times s_2$. Après quelques calculs matriciels, nous obtenons

$$J_1' = \partial h_1' / \partial u \,|_{\mathbb{E}} = \begin{pmatrix} I_2 & A_1 & 0_{2,p} & 0_{2,p^2} \\ 0_{2p,2} & A_2 & 0_{2p,p} & 0_{2p,p^2} \\ 0_{p,2} & 0_{p,2p} & I_p & 0_{p,p^2} \\ 0_{p^2,2} & 0_{p^2,2p} & 0_{p^2,p} & I_{p^2} \end{pmatrix},$$

avec $A_1 = \begin{bmatrix} -\widetilde{m}'_1/p_1^2 & 0_{1,p} \\ 0_{1,p} & -\widetilde{m}'_2/p_2^2 \end{bmatrix}$ et $A_2 = \begin{bmatrix} \frac{1}{p_1} I_p & 0_{p,p} \\ 0_{p,p} & \frac{1}{p_2} I_p \end{bmatrix}$.

Nous en déduisons l'expression de la matrice de covariance de la distribution asymptotique de $\overline{U}_1$:

$$\Delta_1 = J_1 \Sigma_U J_1' = \begin{pmatrix} B_{11} & 0_{2,2p} & B_{13} & B_{14} \\ 0_{2p,2} & B^*_{22} & B^*_{23} & B^*_{24} \\ B'_{13} & B'^*_{23} & B_{33} & B_{34} \\ B'_{14} & B'^*_{24} & B'_{34} & B_{44} \end{pmatrix}.$$

où $B^*_{22} = A'_1 B_{11} A_1 + A'_1 B_{12} A_2 + A_2 B'_{12} A_1 + A_2 B_{22} A_2$; $B^*_{23} = A'_1 B_{13} + A_2 B_{23}$ et $B^*_{24} = A'_1 B_{14} + A_2 B_{24}$.
Considérons maintenant la fonction h_2 :

$$\begin{array}{rcccc} h_2 & : & \mathbb{R}^{2+3p+p^2} & \longrightarrow & \mathbb{R}^{p+p+p^2} \\ & & u = (a_1, a_2, b'_1, b'_2, c', \mathrm{vec}(d)')' & \longmapsto & (b'_1, b'_2, (\mathrm{vec}(d) - \mathrm{vec}(cc'))')'. \end{array}$$

 ISSN: 2102-6238

En remarquant que $h_2(\overline{U}_1) = \begin{pmatrix} m_{1,N} \\ m_{2,N} \\ \text{vec}(\Sigma_N) \end{pmatrix}$ et $h_2(\mu_{U_1}) = \begin{pmatrix} m_1 \\ m_2 \\ \text{vec}(\Sigma) \end{pmatrix}$, nous déduisons par la Delta méthode que

$$\sqrt{N}\left(\begin{pmatrix} m_{1,N} \\ m_{2,N} \\ \text{vec}(\Sigma_N) \end{pmatrix} - \begin{pmatrix} m_1 \\ m_2 \\ \text{vec}(\Sigma) \end{pmatrix}\right) \xrightarrow{\mathscr{L}} \mathscr{N}_{2p+p^2}(0,\Delta_2),$$

où

$$\Delta_2 = J_2 \Delta_1 J_2' \tag{C.3}$$

avec $J_2' = \partial h_2'/\partial u \,|_{\mathbb{E}}$. Après quelques calculs de dérivées vectorielles , nous obtenons

$$J_2' = \begin{pmatrix} 0_{2,p} & 0_{2,p} & 0_{2,p^2} \\ I_p & 0_{p,p} & 0_{p,p^2} \\ 0_{p,p} & I_p & 0_{p,p^2} \\ 0_{p,p} & 0_{p,p} & -(I_p \otimes \mu' + \mu' \otimes I_p) \\ 0_{p^2,p} & 0_{p^2,p} & I_{p^2} \end{pmatrix}.$$

Puisque $Z = \sqrt{N}\{(m_{1,N} - m_1) - (m_{2,N} - m_2) - D^* \text{vec}(\Sigma_N - \Sigma)\}$, introduisons la matrice de dimension $p \times (2p + p^2)$ suivante : $E = \begin{pmatrix} I_p & -I_p & -D^* \end{pmatrix}$. Nous avons alors

$$Z = E\sqrt{N}\left(\begin{pmatrix} m_{1,N} \\ m_{2,N} \\ \text{vec}(\Sigma_N) \end{pmatrix} - \begin{pmatrix} m_1 \\ m_2 \\ \text{vec}(\Sigma) \end{pmatrix}\right)$$

La loi asymptotique de Z est donc

$$Z \xrightarrow{\mathscr{L}} \mathscr{N}_{2p+p^2}(0,\Delta_3),$$

où

$$\Delta_3 = E \Delta_2 E'. \tag{C.4}$$

Étape 3 : loi asymptotique de $\sqrt{N}(\tilde{\theta}_N - \tilde{\theta})$
Nous avons $\sqrt{N}(\tilde{\theta}_N - \tilde{\theta}) = \Sigma_N^{-1} Z$. Vu que Σ_N^{-1} converge en probabilité vers Σ^{-1} et que Z est asymptotiquement normal de moyenne nulle et de matrice de covariances Δ_3, nous avons finalement

$$\sqrt{N}(\tilde{\theta}_N - \tilde{\theta}) \xrightarrow{\mathscr{L}} \mathscr{N}(0, \Sigma^{-1} \Delta_3 \Sigma^{-1}).$$

□

 ISSN: 2102-6238