Rapport Final

Parcours Recherche

Inférence causale entre séries temporelles et extensions à l'analyse de sensibilité

Réalisé par:
Youssef Annaki

Encadré par: **Marianne Clausel**
en collaboration avec:
Georges Oppenheim
Aman Sinha

le 07 Juin 2021

Contents

Résumé

Nous nous intéressons à la notion de causalité entre séries temporelles. Cela s'inscrit dans le cadre de l'étude des interactions entre des séries temporelles et de l'analyse de l'influence d'une (ou plusieurs) série(s) sur une autre ce qui permet d'avoir des informations pertinentes sur l'évolution d'un ensemble de séries temporelles et donc d'améliorer les prédictions futures faites sur ces dernières. Nous cherchons aussi à définir la notion de causalité entre des séries temporelles et de mettre en place un nouveau cadre d'étude pour ces relations de causalité puisqu'il n'existe pas de définition exacte de la causalité. Nous recourant à des techniques d'analyse de sensibilité qu'on adapte au cadre des séries temporelles.
Nous développons une approche de la causalité en exploitant les **valeurs de Shapley**. Cette dernière permet de quantifier l'influence et l'importance des instants passés d'une série temporelle sur les valeurs futures prédites. On applique cette méthode à des séries financières relatives aux cotations du CAC40 en se plaçant sur différents intervalles de temps et pour différents pas des séries financières (15 secondes, 1 minute, 5 minutes,...).

Chapter 1

Introduction

1.1 Modèle Vecteur Autoregressif (VAR)

Definition 1 (Modèle VAR) Soit $y_t = (y_{1,t}, y_{2,t}, \ldots, y_{n,t}) \in \mathbb{R}^n$ un vecteur de taille $(n \times 1)$ où $(y_{i,t})_{t\in\mathbb{N}}, \forall i \in (1, \ldots, n)$ sont des séries temporelles dans $\mathbb{R}$.
Le modèle VAR de paramètre de retard p est donné par

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + \epsilon_t, t = 1, \ldots, T \tag{1.1}$$

où les A_i sont des matrices de coefficients de taille $(n \times n)$ et ϵ_t est un $n \times 1$ vecteur bruit blanc stationnaire (terme d'innovation) de moyenne nulle et de matrice de covariance Σ_ϵ.
En utilisant l'opérateur de shift , le VAR(p) s'écrit

$$A(L)y_t = c + \epsilon_t \tag{1.2}$$

avec $A(L) = I_n - A_1 L - \cdots - A_p L^p$.

Proposition 1 (Stabilité) *Soit $(y_t)_{t\in\mathbb{Z}}$ un processus VAR(p).*
$(y_t)_{t\in\mathbb{Z}}$ est dit stable si :

$$det(I_n - A_1 z - \cdots - A_p z^p) \neq 0, \forall z \text{ tel que } |z| > 1 \tag{1.3}$$

Il serait intéressant de développer l'idée dernière cette notion de stabilité pour les modèles VAR. En effet, considérons le processus VAR(1) suivant

$$y_t = c + A_1 y_{t-1} + \epsilon_t.$$

Si ce processus générateur commence à la date $t = 1$, on aura

$$y_t = (I_n + A_1 + \cdots + A_1^{t-1})c + A_1^t y_0 + \sum_{i=0}^{t-1} A_1^i \epsilon_{t-i}$$

De plus, il est parfois utile et plus commode (ce qui est le cas ici) de considérer que l'instant initial du processus VAR se trouve dans l'infini passé. Une autre façon de voir cela c'est de considérer que le temps t dans lequel on étudie le processus est

très éloigné de l'instant initial.
Sous cette hypothèse on a

$$y_t = (\sum_{i=0}^{\infty} A_1^j)c + \lim_{i\to\infty} A^i y_0 + \sum_{i=0}^{\infty} A_1^i \epsilon_{t-i}$$

Si toutes les valeurs propres de A_1 ont un module inférieur à 1, alors $(A_1^i)_{i\in\mathbb{N}}$ est absolument sommable. En effet, cela peut être prouvé en recourant à la forme canonique de Jordan de A_1. De plus, $\sum_{i=1}^{\infty} A_1^i \epsilon_{t-i}$ existe en moyenne quadratique (voir Appendice A, [19]). D'autre part,

$$(I_n + A_1 + \cdots + A_1^j)c \underset{j\to\infty}{\longrightarrow} (I_n - A_1)^{-1}c$$

En prenant en compte ces résultats, on peut écrire,

$$y_t = \mu + \sum_{i=0}^{\infty} A_1^i \epsilon_{t_i}, t = 0, \pm 1, \pm 2, \ldots \tag{1.4}$$

avec $\mu = (I_n - A_1)^{-1}c$.
Dans le cas d'un processus VAR(1), on dit qu'il est stable si

$$\det(I_n - A_1 z) \neq 0, \forall \text{ z tel que } |\text{z}|>1.$$

En d'autres termes, si les valeurs propres de la matrice A_1 ont un module strictement inférieur à 1, ce qui est bien l'hypothèse initiale faite sur la matrice A_1.

Le résultat précédent peut être généralisé facilement à des processus VAR(p) avec $p > 1$ car tout processus VAR(p) peut s'écrire sous la forme d'un VAR(1). Plus précisément, si $(y_t)_{t\in\mathbb{Z}}$ est un processus VAR(p) alors on a

$$Y_t = c + AY_{t-1} + \epsilon_t$$

avec

$$\boldsymbol{Y_t} = \begin{bmatrix} \boldsymbol{y_t} \\ \boldsymbol{y_{t-1}} \\ \boldsymbol{y_{t-2}} \\ \vdots \\ \boldsymbol{y_{t-p+1}} \end{bmatrix}, \boldsymbol{c_t} = \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{bmatrix},$$

$$\text{A}= \begin{bmatrix} \boldsymbol{A_1} & \boldsymbol{A_2} & \cdots & \boldsymbol{A_{p-1}} & \boldsymbol{A_p} \\ \boldsymbol{I_p} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I_p} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{I_p} & \boldsymbol{0} \end{bmatrix}, \boldsymbol{\epsilon_t} = \begin{bmatrix} \boldsymbol{\epsilon_t} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{bmatrix},$$

En suivant la même démarche que précédemment, Y_t est stable si

$$det(I_{n\times p} - Az) \neq 0 \text{ , } \forall z \text{ tel que } |z| > 1$$

Ce qui est la même condition de stabilité que dans la proposition (1).

Proposition 2 *En utilisant les notations précédentes, et pour un processus VAR(p) stable* $(Y_t)_{t\in\mathbb{Z}}$ *on a*

$$\begin{aligned}
i - \mathbb{E}(Y_t) &= \mu = (I_{np} - A)^{-1}c \\
ii - \Gamma_Y(h) &= \mathbb{E}[(Y_t - \mu)(Y_{t_h} - \mu)^{'}] \\
&= \lim_{n\to\infty} \mathbb{E}[(\sum_{i=0}^{n} A_1^i \epsilon_{t-i})(\sum_{j=0}^{n} \epsilon_{t-i}^{'} A^{i\prime}{}_1)] \\
&= \lim_{n\to\infty} \sum_{i=0}^{n}\sum_{j=0}^{n} A_1^i \mathbb{E}[\epsilon_{t_i}\epsilon'_{t_{hj}}] A^{j\prime}{}_1 \qquad (1.5) \\
&= \lim_{n\to\infty} \sum_{i=0}^{n} A_1^{h+i}\Sigma_\epsilon \boldsymbol{A}^{i\prime}{}_1 \\
&= \sum_{i=0}^{\infty} A_1^{h+i}\Sigma_\epsilon A^{i\prime}{}_1
\end{aligned}$$

Rappelons que $\mathbb{E}(\epsilon_t\epsilon'_s) = 0$ *pour* $s \neq t$ *et* $\mathbb{E}(\boldsymbol{\epsilon_t\epsilon'_t}) = \boldsymbol{\Sigma_\epsilon}$ *pour tout* t.
Γ_Y *est la fonction d'auto-covariance de* $(Y_t)_{t\in\mathbb{Z}}$.

Remarque

Ce qui motive le recourt aux modèles VAR stables est l'existence d'expressions explicites pour le moment de premier et deuxième ordre ainsi que l'existence d'une approche en terme de moyen mobile (MA) pour ce type de processus.

Mais un processus VAR peut être défini aussi en l'absence de condition de stabilité.

Definition 2 (Processus stationnaire) Un processus stochastique est stationnaire (au sens faible) si son moments de premier et deuxième ordre sont indépendants du temps. En d'autres termes, un processus $(y_t)_{t\in\mathbb{Z}}$ est stationnaire si

- $\mathbb{E}(y_t) = c, \forall t \in \mathbb{Z}$
- $\mathbb{E}[(y_t - \mu)(y_{t-h} - \mu)^{'}] = \Gamma_y(h) = \Gamma_y(-h), \forall t \in \mathbb{Z}$ and $h = 0, 1, \ldots$

Cela revient à dire que tous les y_t ont le même vecteur moyenne constant c et que l'auto-covariance du processus ne dépend pas de t mais uniquement de la longueur de l'intervalle de temps h entre y_t and y_{t-h}.

Proposition 3 *Soit* $(y_t)_{t\in\mathbb{N}}$ *un processus VAR stable, alors ce dernier est stationnaire.*

Proof Le raisonnement qui suit la proposition (1) ainsi que le résultat de la proposition (2) montrent que le moment de premier et deuxième ordre sont indépendants du temps, d'où la stationnarité. ■

1.2 Les modèles GARCH

Definition 3 Un processus GARCH de paramètres p et q (GARCH(p,q)) $(y_k)_{k\in\mathbb{Z}}$ est défini par

$$\begin{aligned} y_k &= \sigma_k \epsilon_k \\ \sigma_k^2 &= \omega + \sum_{i=1}^{p} \alpha_i y_{k-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{k-j}^2 \end{aligned} \tag{1.6}$$

avec $\omega > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, et le terme d'innovation $(\epsilon_i)_{i\in\mathbb{Z}}$ est un bruit blanc tel que $\mathbb{E}(\epsilon_0) = 0$ and $\mathbb{E}(\epsilon_0^2) = 1$.

Proposition 4 *σ_k^2 est la variance conditionnelle de y_k par rapport à la tribu définie par $\mathcal{F}_{k-1} = \sigma(\epsilon_t, -\infty < t \leq k-1)$.*

Proof On a,
$\forall (i \geq j) \in \mathbb{Z}$, $\epsilon_i \perp\!\!\!\perp \epsilon_j$ et donc $\mathbb{E}(\epsilon_i|\mathcal{F}_j) = \mathbb{E}(\epsilon_i)$. De plus, y_k est $\mathcal{F}_{k-1}$-mesurable car c'est une fonction mesurable de $(\epsilon_t, -\infty < t \leq k-1)$. D'autre part, on a

$$Var(y_k|\mathcal{F}_{k-1}) = \mathbb{E}(y_k^2|\mathcal{F}_{k-1}) - (\mathbb{E}(y_k|\mathcal{F}_{k-1}))^2,$$

et en remplaçant y_k par son expression (1.6) dans cette formule on obtient

$$Var(y_k|\mathcal{F}_{k-1}) = \sigma_k^2$$

■

La stationnarité et l'existence des moments pour un processus stochastique sont fondamentaux dans l'inférence statistique. C'est pour cette raison qu'il est impératif de trouver des conditions nécessaires et suffisantes pour la stationnarité et l'existence des moments pour un modèle GARCH(p,q). En effet, Bougeral et Picard (1992) ont proposé une condition nécessaire et suffisante pour la stationnarité (au sens faible) d'un processus GARCH. Min Chen et Hong Zhi An (1998) ont généralisé les travaux de Bougeral et Picard et ils ont proposé certaines conditions suffisantes pour la stationnarité (au sens faible) et l'existence des moments (au moins du premier et deuxième ordre) pour des processus GARCH.

Lemma 1 (Stationnarité et unicité) *Si*

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1, \tag{1.7}$$

alors le modèle GARCH(p,q) (1.6) a une unique solution stationnaire.

Proof Voir Min Chen and Hong Zhi An (1998), pages 507 to 509.

Proposition 5 *En utilisant la filtration $(\mathcal{F}_k)_k$ comme précédemment, et soit $(y_k)_{k\in\mathbb{Z}}$ un processus GARCH(p,q) stationnaire , on a*

- $\mathbb{E}(y_k) = 0$
- $\mathbb{E}(y_k y_{k+h}) = 0$, *pour* $h > 0$
- $\mathbb{E}(y_k^2) = \frac{\omega}{1-\sum_{i=1}^{R} \alpha_i+\beta_i}$

avec $R = max(p, q)$

PROOF k $\in \mathbb{Z}$ et $h > 0$,

- $\mathbb{E}(y_k) = \mathbb{E}(\sigma_k \epsilon_k) = \mathbb{E}[\mathbb{E}(\sigma_k \epsilon_k | \mathcal{F}_{k-1})] = \mathbb{E}(\sigma_k \mathbb{E}(\epsilon_k | \mathcal{F}_{k-1})) = \mathbb{E}(\sigma_k \epsilon_k) = 0$
- $\mathbb{E}(y_k y_{k+h}) = \mathbb{E}(y_k \sigma_{k+h} \epsilon_{k+h}) = \mathbb{E}[\mathbb{E}(y_k \sigma_{k+h} \epsilon_{k+h} | \mathcal{F}_{k+h-1})] = \mathbb{E}[y_k \sigma_{k+h} \mathbb{E}(\epsilon_{k+h} | \mathcal{F}_{k+h-1})] = 0$
- Considèrons le processus $(Z_k)_{k \in \mathbb{Z}}$ défini par

$$Z_k = y_k^2 - \sigma_k^2 = \sigma_k^2(\epsilon_k^2 - 1)$$

$(Z_k)_{k \in \mathbb{Z}}$ est une martingale difference process. En effet, pour tout k, Z_k est $\mathcal{F}_{k-1}$-adapté. De plus,

$$\mathbb{E}(Z_k | \mathcal{F}_{k-1}) = \sigma_k^2(\mathbb{E}(\epsilon_{k-1}^2 | \mathcal{F}_{k-1}) - 1) = 0$$

Ce qui signifie que, $(Z_k)_{k \in \mathbb{Z}}$ est une martingale difference process et que $\mathbb{E}(Z_k) = 0, \forall k \in \mathbb{Z}$.
On a aussi que,

$$\begin{aligned} y_k^2 &= \sigma_k^2 + Z_k \\ &= \omega + \sum_{i=1}^{p} \alpha_i y_{k-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{k-j}^2 + Z_k \\ &= \omega + \sum_{i=1}^{p} \alpha_i y_{k-i}^2 + \sum_{j=1}^{q} \beta_j y_{k-j}^2 - \sum_{j=1}^{q} \beta_j Z_{k-j} + Z_k \end{aligned}$$

Soit $R = max(p, q)$, $\alpha_i = 0$ tel que $i > p$, et $\beta_j = 0$ tel que $j > q$, alors y_k^2 peut s'écrire sous la forme

$$y_k^2 = \omega + \sum_{i=1}^{R} (\alpha_i + \beta_j) y_{k-i}^2 - \sum_{j=1}^{q} \beta_j Z_{k-j} + Z_k$$

Rappelons que $\mathbb{E}(y_k^2) = \mathbb{E}(y_{k+h}^2)$ pour tout k et h (Stationnarité de $(y_k)_{k \in \mathbb{Z}}$), la variance de $(y_k)_{k \in \mathbb{Z}}$ est donnée par:

$$\begin{aligned} \mathbb{E}(y_k^2) &= \omega + \sum_{i=1}^{R} (\alpha_i + \beta_j) \mathbb{E}(y_{k-i}^2) - \sum_{j=1}^{q} \beta_j \mathbb{E}(Z_{k-j}) + \mathbb{E}(Z_k) \\ &= \omega + \mathbb{E}(y_k^2) \sum_{i=1}^{R} (\alpha_i + \beta_i) \end{aligned} \tag{1.8}$$

et on a,

$$\mathbb{E}(y_k^2) = \frac{\omega}{1 - \sum_{i=1}^{R} \alpha_i + \beta_i}$$

■

1.3 La régression k-NN pour les séries temporelles

Dans cette section, on présente la méthode k-NN pour la prédictions de valeurs futurs de séries temporelles et qu'on utilisera par la suite dans le dernier chapitre de ce rapport.
k-NN est une méthode de classification non paramétrique basée sur la mesure de similarité d'un point à un ensemble de données d'entraînement étiquetées où k est un paramètre à fixer.

k-NN peut aussi être utilisé pour des problèmes de régression dans le cas de données unidimensionnelles. Dans cette approche, l'estimation d'une valeur x inconnue est la moyenne des k plus proche valeurs connues $(x_1, \ldots, x_k)$ de la base d'entraînement:

$$x = \sum_{i=1}^{k} \frac{x_i}{k} \tag{1.9}$$

Une meilleure prédiction est obtenue grâce à une moyenne pondérée des $(x_1, \ldots, x_k)$ en prenant en compte la proximité des k voisins de la valeur x qu'on cherche à prédire:

$$x = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=k}^{k} w_i} \tag{1.10}$$

Cette approche donne plus d'importance aux voisins les plus proches de x. A titre d'exemple, on peut utiliser des poids dépendant de la distance entre chaque voisin et la valeur à prédire,

$$w_i = \frac{1}{\lambda + d(x, x_i)}, i \in (1, 2, \ldots, k)$$

où λ est un paramètre.
On adaptera donc cette approche à la prédiction de séries temporelles. Pour se faire, on doit:

1. Choisir une fonction de similarité pour déterminer les plus proches voisins.
2. Fixer une méthodologie pour la prédiction.

Soit $(X_k)_{k\in\mathbb{N}}$ la séries temporelles sur laquelle on veut faire des prédictions. On voudrais, en se basant sur la séquence $A = (X_{n-h+1}, \ldots, X_{n-1}, X_n)$ avec $h > 1$, avoir un estimateur de X_{n+1}. On cherche les k plus proches séquences de longueur h de $(X_{n-h+1}, \ldots, X_{n-1}, X_n)$. Ces dernières sont de la forme:

$$(X_{q_1}, \ldots, X_{q_1+h-1}), \ldots, (X_{q_k}, \ldots, X_{q_k+h-1})$$

où $1 \leq q_i \leq n-h, i \in (1, \ldots, k)$. On peut utiliser par exemple la métrique/distance L^p pour évaluer les différentes distances. entre les séquences (de même longueur).
Après avoir déterminer les k plus proches séquences de A, on estime X_{n+1} par

$$X_{n+1} = \frac{\sum_{i=1}^{k} w_i X_{q_i+h}}{\sum_{i=k}^{k} w_i} \tag{1.11}$$

Algorithm 1 Algorithme k-NN pour la prédiction sur des séries temporelles

Require: k, horizon de prédiction p, Série univariée $(X_1, X_2, \ldots, X_n)$
Ensure: $(X_{n+1}, \ldots, X_{n+p})$
  for i $\in$ (1,...,p) **do**
    Déterminer les k plus proches séquences
    $(X_{q_1}, \ldots, X_{q_1+h-1}), \ldots, (X_{q_k}, \ldots, X_{q_k+h-1})$
    de la base de prédiction $(X_{n-h+i}, \ldots, X_{n+i-1})$
    Estimer X_{n+i} en se basant sur $X_{q_1^i+h}, \ldots, X_{q_k^i+h}$
    Ajouter X_{n+i} à la fin de la séquence $(X_1, \ldots, X_{n+i-1})$
  end for

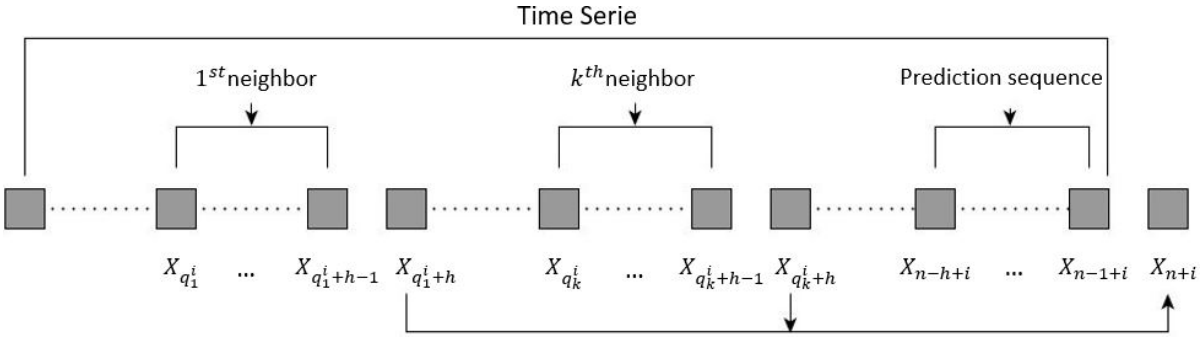

Figure 1.1

Chapter 2

Causalité de Granger

2.1 Introduction :

Dans cette section on va s'intéresser à la causalité au sens de Granger, introduite par W.J Granger en 1969. Cette notion repose sur la mesure de la dépendance statistique entre le passé d'un processus et le présent d'un autre. Dans ce sens, la définition proposée par Granger repose sur la perception intuitive qu'on a de la causalité, à savoir que la cause précède l'effet. Et on dit que : Un processus ne cause pas un autre (au sens de Granger) si la connaissance de l'historique du premier processus ne contribue pas à la prédiction des valeurs futures du deuxième.

Avant d'aller plus loin, il faut noter que la notion de causalité est d'abord philosophique et il n'existe aucune caractérisation exacte de cette notion. Cela a fait que la notion de causalité proposée par Granger soit fortement critiquée par les philosophes. Or, il faut savoir que cette causalité est définie dans le cadre des modèles économétriques et qu'elle ne visait pas à susciter des réflexions philosophiques. Dans les années 1950 et 1960, différents débats et controverses concernées la notion de causalité dans le cadre de l'économétrie, notamment entre Wold, Basmann et Strotz. Au sujet de cette difficulté de mettre en place une définition exacte de la causalité alliant économie et économétrie, Sargent (voir [28, Sargent, 1977]) a dit :

It is true that Granger's definition of a causal relation does not, in general, coincide with the economist's usual definition of one : namely, a relation that is invariant with respect to interventions in the form of imposed changes in the processes governing the causing variables.

Une autre caractéristique de la causalité de Granger qu'il ne faut surtout pas négliger, c'est que cette notion de la causalité est relative, au sens où elle est définie par rapport à l'ensemble d'information considéré et disponible au moment de l'étude. Par exemple, si on se concentre sur un ensemble de $d \in \mathbb{N}^*$ séries temporelles, la relation causale au sens de Granger entre deux processus de l'ensemble considéré est définie par rapport aux $d-2$ séries restantes.

2.1.1 Notations :

Toutes les variables, vecteurs et séries aléatoires sont définies sur un même espace de probabilité $(\Omega, \mathcal{B}, \mathbb{P})$. Ils prennent des valeurs dans $\mathbb{R}$ ou $\mathbb{R}^d$ avec d un entier strictement positif. Dans ce qui suit, les processus ou les séries temporelles seront représentés par $(A_t)_{t\in\mathbb{Z}}$, où A désigne le processus considéré. D'autre part, A^n désignera l'historique du processus $(A_t)_{t\in\mathbb{Z}}$ à partir de l'instant n. En d'autres termes, A^n= $(A_{n-1}, A_{n-2}, \dots)$.

2.2 Définitions

2.2.1 Définition probabiliste

Dans cette définition, on va se limiter au cas de trois séries temporelles. La définition dans le cadre de plusieurs séries temporelles est similaire. On se donne donc trois séries temporelles à temps discret $(X_t)_{t\in\mathbb{N}}$, $(Y_t)_{t\in\mathbb{Z}}$ et $(Z_t)_{t\in\mathbb{Z}}$, et on note X_t, Y_t et Z_t les réalisations à l'instant t de, resp, $(X_t)_{t\in\mathbb{Z}}$, $(Y_t)_{t\in\mathbb{Z}}$ et $(Z_t)_{t\in\mathbb{Z}}$.
D'autre part, $p(X_t)$ désignera la distribution de probabilité de la variable aléatoire X_t. De même pour Y_t et Z_t. En se basant sur la définition de la causalité de Granger présentée dans l'introduction, on peut voir que la non causalité au sens de Granger, par exemple de $(X_t)_{t\in\mathbb{Z}}$ vers $(Y_t)_{t\in\mathbb{Z}}$ relativement à (X^t, Y^t, Z^t) se traduit par $p(Y_t|X^t, Y^t, Z^t) = p(Y_t|Y^t, Z^t)$. Ces deux distributions sont identiques ssi X_t et Y_t sont indépendantes conditionnellement à Y^t et Z^t.

Definition 4 $(X_t)_{t\in\mathbb{N}}$ ne cause pas $(Y_t)_{t\in\mathbb{N}}$ au sens de Granger relativement à (X^t, Y^t, Z^t) ssi $Y_t \perp\!\!\!\perp X^t | Y^t, Z^t, \forall t \in \mathbb{Z}$.

Cela revient à dire que conditionnellement à Y^t, Z^t, l'historique de $(X_t)_{t\in\mathbb{Z}}$ ne contient pas plus d'information sur Y_t que l'historique de $(Y_t)_{t\in\mathbb{Z}}$.

Definition 5 (Couplage instantané) $(X_t)_{t\in\mathbb{N}}$ n'est pas instantanément couplé à $(Y_t)_{t\in\mathbb{N}}$ relativement à (X^t, Y^t, Z^t) ssi $Y_t \perp\!\!\!\perp X_t | X^t, Y^t, Z^t, \forall t \in \mathbb{Z}$

Contrairement à la définition de la causalité proposée précèdement, qui s'intéresse à l'impact de l'historique d'un processus sur le présent d'un autre, le couplage instantané mesure l'information commune entre deux processus à un instant donné, relativement à l'historique (ensemble d'information) des processus considerée pour l'étude.

2.2.2 Définition par modèle de prédiction

Dans cette définition, on supposera que $(X_t)_{t\in\mathbb{Z}}$, $(Y_t)_{t\in\mathbb{Z}}$, $(Z_t)_{t\in\mathbb{Z}}$ sont de variance finie. On se place toujours dans le cas de trois séries temporelles. Soit A^n=(X^n, Y^n, Z^n) et B^n un sous-ensemble de A^n: $B^n \subset A^n$. On note $P(Y_t|B^t)$ le meilleur estimateur, à l'instant t, sans biais de Y_t, basé sur les variables aléatoires dans B^t (l'information contenue dans B^t). On introduit aussi $\epsilon(Y_t|B^t) = Y_t - P(Y_t|B^t)$, le résidu relatif à la prédiction de Y_t par $P(Y_t|B^t)$. Et on a $\sigma^2(Y_t|B^t) = \mathbb{E}(\epsilon(Y_t|B^t)^2)$.

Definition 6 $(X_t)_{t\in\mathbb{Z}}$ cause $(Y_t)_{t\in\mathbb{Z}}$ ssi $\sigma^2(Y_t|B^t) < \sigma^2(Y_t|B^t \setminus X^t)$ pour au moins un instant $t \in \mathbb{Z}$.

De cette définition, on comprend bien que la causalité au sens de Granger équivaut au fait que l'historique d'un processus contribue à la prédiction d'un autre et améliore cette dernière. Cela signifie aussi que $(X_t)_{t\in\mathbb{Z}}$ contient une part de l'information sur $(Y_t)_{t\in\mathbb{Z}}$ et a donc un impact sur cette dernière.

2.3 Caractérisation de la causalité de Granger linéaire

Dans cette partie, on va introduire quelques méthodes d'inférence de la causalité de Granger, notamment l'indice de Granger qui est une mesure de causalité, ainsi qu'un test statistique permettant de mettre en évidence les différentes relations causales linéaires entre les séries temporelles étudiées.

On se donne $((\mathrm{y}_{1,t})_{t\in\mathbb{Z}},\ldots,(\mathrm{y}_{K,t})_{t\in\mathbb{Z}})$, un ensemble de K processus aléatoires. Et on pose $\mathrm{Y}_t=(y_{1,t},\ldots,y_{K,t})'$.
On introduit le modèle Vecteur Autorégressif (VAR) multivarié, bien défini, d'ordre p suivant (voir 1.1) :

$$\mathbb{Y}_t = \sigma + A_1\mathbb{Y}_{t-1} + A_2\mathbb{Y}_{t-2} + ... + A_p\mathbb{Y}_{t-p} + e_t$$

Avec $(A_i)_{i\in[1:\ldots:p]}$, les matrices coefficients du modèle VAR, de taille $\mathrm{K}\times K$. $(\mathrm{e}_t)_{t\in\mathbb{Z}}$ est un vecteur bruit blanc de taille $K\times 1$. p est le paramètre de retard qu'on détermine en se basant sur les critères AIC et BIC.

Hypothèse : (Stationnarité et normalité des résidus)
Pour chaque composante $y_{i,t}$ de $\mathbb{Y}_t$, les résidus $e_{i,t}$ sont indépendants et gaussiens avec $\mathbb{E}(e_{i,t}) = 0$ et $\mathbb{E}(e_{i,t}^2) = c\ \forall i \in (1,\ldots,K)$ et $\forall t \in \mathbb{N}$, où c est une constante strictement positive.
D'autre part, $\mathbb{Y}_t$ et e_t sont indépendants $\forall t \in \mathbb{Z}$. Cette dernière condition découle de la construction du modèle VAR.

2.3.1 Test de Granger

L'étude de la causalité linéaire par le biais du test de Granger va se baser sur la construction des modèles VAR. En particulier, on va déterminer l'existence, ou non, d'une relation causale au sens de Granger entre deux séries temporelles relativement à l'information disponible. On s'intéressera donc, dans ce qui suit, aux processus $(y_{i,t})_{t\in\mathbb{Z}}$ et $(y_{j,t})_{t\in\mathbb{Z}}$ avec $i \neq j$ et on se concentrera sur l'étude de la relation causale $(y_{i,t})_{t\in\mathbb{Z}} \Rightarrow (\mathrm{y}_{j,t})_{t\in\mathbb{Z}}$.

On détermine en premier lieu deux types de modèles VAR pour le processus $Effet$ $(y_{j,t})_{t\in\mathbb{Z}}$. En effet, la première représentation VAR va inclure l'historique de toutes les séries temporelles disponibles, y compris celui de $(y_{i,t})_{t\in\mathbb{Z}}$, c'est ce qu'on appellera un modèle Non-Restreint. Alors que la deuxième représentation, qu'on appellera un modèle Restreint, exclut l'historique de $(y_{i,t})_{t\in\mathbb{Z}}$ lors de la construction du modèle VAR. On a donc :

$$y_{t,j} = \sigma_j + \sum_{k=1}^{K}\sum_{m=1}^{p}(A_m)_{j,k}y_{t-m,k} + e_{t,j} \text{ (Modèle Non-Restreint)} \tag{2.1}$$

$$\hat{y}_{t,j} = \hat{\sigma}_j + \sum_{k=1,k\neq i}^{K}\sum_{m=1}^{\hat{p}}(\hat{A}_m)_{j,k}y_{t-m,k} + \hat{e}_{t,j} \text{ (Modèle Restreint)} \tag{2.2}$$

Le recours à ces deux modèles VAR différents permettra d'évaluer l'impact de $(y_{i,t})_{t\in\mathbb{Z}}$ sur la prédiction des valeurs futurs de $(y_{j,t})_{t\in\mathbb{Z}}$.

Maintenant qu'on a défini le cadre de base pour cette étude, nous allons procéder à un F-test (test de Fisher) pour la causalité de Granger à partir de (2.1) et (2.2). Le but de ce test de Fisher est de tester l'hypothèse

$$\mathcal{H}_0 : (A_m)_{j,i} = 0 \ \forall m \in (1,\dots,p)$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \exists m \in (1,\dots,p) \text{ tel que } (A_m)_{j,i} \neq 0$$

L'hypothèse $\mathcal{H}_0$ équivaut à dire que $(\mathrm{y}_{i,t})_{t\in\mathbb{N}}$ ne cause pas $(\mathrm{y}_{j,t})_{t\in\mathbb{N}}$.
On peut remarquer que dans cette étude, le test de Fisher équivaut à un test de significativité des paramètres de régression dans le modélisation VAR (2.1).

La statistique de test est donnée par :

$$T = \frac{\frac{SSR_R - SSR_F}{p}}{\frac{SSR_F}{n-2p-1}} \tag{2.3}$$

Où SSR_R (resp. SSR_F) est la somme des carrés des résidus du modèle Restreint (resp. Non Restreint) et n est le nombre de résidus considérés (nombre d'observations).

Sous l'hypothèse $\mathcal{H}_0$ on a :

$$T = \frac{\frac{SSR_R - SSR_F}{p}}{\frac{SSR_F}{n-2p-1}} \sim F(p, n-2p-1) \tag{2.4}$$

Cela est garanti par l'hypothèse de bruit blanc gaussien des résidus qui est essentielle dans cette étude.

De façon analogue, on peut aussi chercher à savoir si $(y_{j,t})_{t\in\mathbb{Z}}$ cause au sens de Granger $(y_{i,t})_{t\in\mathbb{Z}}$: $(y_{j,t})_{t\in\mathbb{Z}} \Rightarrow (\mathrm{y}_{i,t})_{t\in\mathbb{Z}}$.

Definition 7 (FeedBack Effect) Soit $(X_t)_{t\in\mathbb{Z}}$ et $(Y_t)_{t\in\mathbb{Z}}$ deux processus aléatoires. Si $(X_t)_{t\in\mathbb{Z}}$ cause au sens de Granger $(Y_t)_{t\in\mathbb{Z}}$ ainsi que $(Y_t)_{t\in\mathbb{Z}}$ cause au sens de Granger $(X_t)_{t\in\mathbb{Z}}$, on dit qu'on a une boucle rétroactive (feedback effect) et on notera $(X_t)_{t\in\mathbb{Z}} \leftrightarrow (Y_t)_{t\in\mathbb{Z}}$.

2.3.2 Indice de causalité de Granger

Toujours dans le cadre de la causalité linéaire et en utilisant les représentations VAR introduites précédemment (2.1), (2.2), on peut quantifier la relation causale entre séries temporelles en recourant à l'indice de causalité linéaire définie par :
Pour $(i, j) \in (1, \ldots, K)$

$$F_{y_{i,t} \Rightarrow y_{j,t}} = log(\frac{var(\hat{e}_{j,t})}{var(e_{j,t})}) \tag{2.5}$$

Ainsi, si $F_{y_{i,t} \Rightarrow y_{j,t}} > 0$ alors $var(\hat{e}_{j,t}) > var(e_{j,t})$. Cela revient à dire que l'historique de $(y_{i,t})_{t \in \mathbb{Z}}$ avant un instant t améliore la prédiction de $(y_{j,t})_{t \in \mathbb{Z}}$ en t et donc que $(y_{i,t})_{t \in \mathbb{Z}}$ cause $(y_{j,t})_{t \in \mathbb{Z}}$ au sens de Granger et relativement à l'information disponible. Dans le cas contraire, on aura $F_{y_{i,t} \Rightarrow y_{j,t}} \approx 0$ et $var(\hat{e}_{j,t}) \approx var(e_{j,t})$. Ce qui signifiera que $(y_{i,t})_{t \in \mathbb{Z}}$ ne cause pas $(y_{j,t})_{t \in \mathbb{Z}}$ au sens de Granger.

2.4 Caractérisation de la causalité de Granger non linéaire :

Notations :

Dans cette partie, on se limitera à l'inférence causale $((B_t)_{t \in \mathbb{Z}} \Rightarrow (A_t)_{t \in \mathbb{Z}})$ entre 2 séries temporelles $(A_t)_{t=1,\ldots,N}$ et $(B_t)_{t=1,\ldots,N}$ sur l'intervalle de temps $[1 : N]$, une généralisation au cas multivarié est donnée dans [16].
Soit p le paramètre de shift dans la représentation autorégressive (2.2). On pose $X_i = (A_i, \ldots, A_{i+m-1})^T$ et $Y_i = (B_i, \ldots, B_{i+m-1})^T$, ainsi que $x_i = A_{i+p}$ pour tout $i = 1, \ldots, N$. On considère ces quantités comme N réalisations des variables aléatoires X, Y et x.
On définit $\mathbb{X}$ comme étant une matrice de taile $m \times N$ ayant X_i comme colonnes et $\mathbb{Z}$ une matrice de taille $2m \times N$ ayant $Z_i = (X_i^T, Y_i^T)$ comme colonnes. On définit aussi $x = (x_1, \ldots, x_N)^T$. Sans perte de généralité, on suppose que chaque composante de X et Y est de moyenne nulle ainsi que x est normalisée et de moyenne nulle [16].

2.4.1 Causalité de Granger par la méthode des noyaux

Soit K une fonction noyau, ayant la représentation spectrale $K(X, X') = \sum_a \lambda_a e_a(X) e_a(X')$, on définit $H = Im(\mathbb{K})$, l'image de la matrice de Gram $\mathbb{K}$ d'éléments $K(X_i, X_j)$. On définie $\widetilde{x}$, la projection de x sur $\mathbb{H}$ et $\widetilde{x'}$ la projection de x sur $\mathbb{K}'$ la matrice de Gram d'éléments $K(Z_i, Z_j)$.
On pose $y = x - \widetilde{x}$. L'indice de causalité est défini par :

$$\delta_{B \Rightarrow A} = \sum_{i=1}^{m} r_i^2 \tag{2.6}$$

Avec r_i, le coefficient de corrélation de Pearson entre y et t_i, où t_i sont les valeurs propres de $\widetilde{\mathbb{K}} = \mathbb{K}' - \mathbb{K}'\mathbb{P} - \mathbb{P}(\mathbb{K}' - \mathbb{K}'\mathbb{P})$. $\mathbb{P}$ est la matrice de projection sur

$$\mathbb{H}$$

.

2.5 Exemples et applications

Dans cette partie, on va mettre en pratique les différentes méthodes d'inférence causale citées précédemment. On va tout d'abord s'intéresser à des données synthétiques qui vont nous permettre d'avoir une idée plus claire sur les points forts et aussi les points faibles des méthodes introduites, ainsi que la démarche générale suivie pour inférer les relations causales au sens de Granger, puis on va mener une étude de causalité entre des séries financières qui nous permettra d'identifier les faiblesses des méthodes d'inférence causale présentées et donc de motiver la mise en place de nouvelles approches plus efficaces.

2.5.1 Exemple de modèle linéaire

On considère trois processus aléatoires générés par chacun des processus couplés autoregressifs suivant :

$$\begin{aligned} x_1(t) &= 0.95x_1(t-1) - 0.9x_1(t-2) + \tau_1(t) \\ x_2(t) &= (1-a)x_1(t-1) + 0.5x_2(t-2) + \tau_2(t) \\ x_3(t) &= -0.7x_1(t-1) + \tau_3(t) \end{aligned} \tag{2.7}$$

Où $(\tau_i)_{i\in(1,2,3)}$ sont des bruit blancs gaussiens et $a \in [0:1]$ un paramètre du modèle. On s'intéressera dans ce qui suit à la relation causale $(x_1(t))_{t\in\mathbb{Z}} \Rightarrow x_2(t))_{t\in\mathbb{Z}}$.

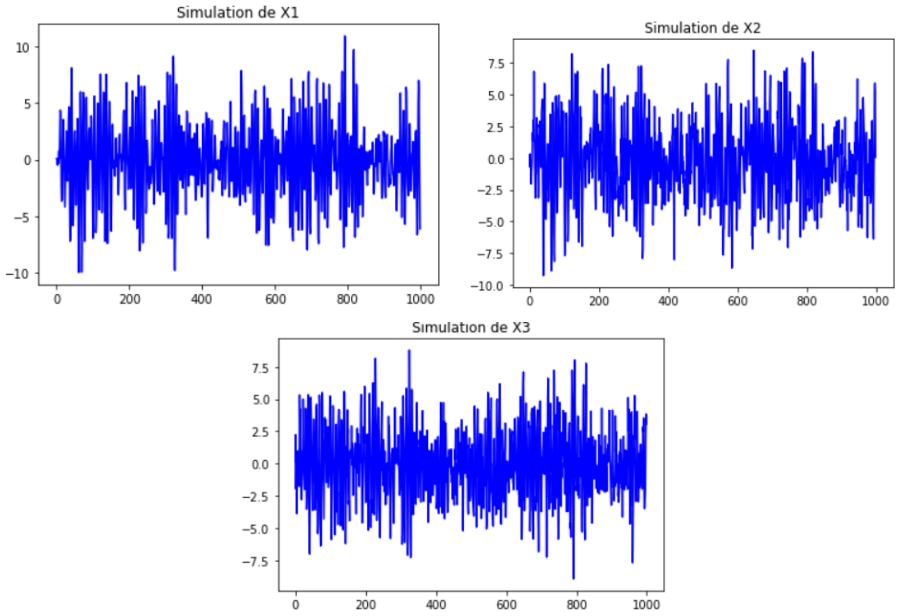

Figure 2.1

Etude des 3 processus

Avant de passer à une modélisation par VAR des processus précédents, il faut tout d'abord s'assurer que ces derniers sont bien stationnaires (dans notre étude, on va se limiter à la stationnarité faible). Pour cela, on va se baser sur deux tests statistiques distincts : le test de Dickey-Fuller Augmenté(ADF) et le test Kwiatkowski-Phillips-Schmidt-Shin(KPSS).

Le test ADF est un test de racine unitaire dont l'hypothèse nulle est que la série a été générée par un processus présentant une racine unitaire, et donc, qu'elle n'est

pas stationnaire. Alors que le test KPSS est basé sur l'hypothèse nulle H_0 que la série est stationnaire.

L'étape qui suit consiste à vérifier l'adéquation du modèle VAR en se basant sur une étude des résidus des représentations restreinte et non-restreinte de $(x_2(t))_{t\in\mathbb{N}}$, qu'on note respectivement $(\epsilon_2(t))_t$ et $(\epsilon'_2(t))_t$.
La première étape consiste à vérifier que les résidus sont stationnaires.

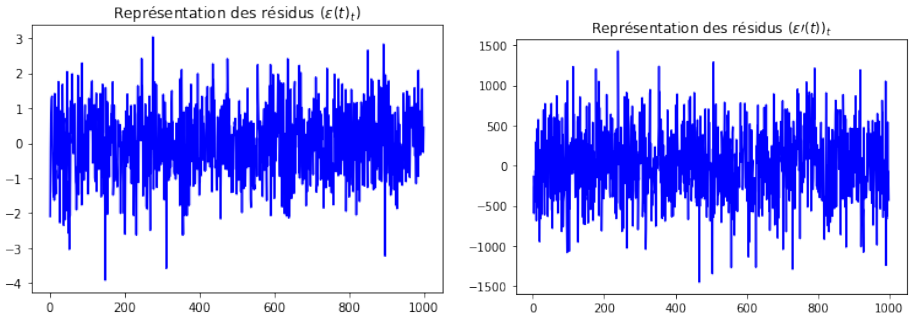

Figure 2.2: Visualisation des résidus

En effet, un test KPSS couplé avec un test ADF permet d'affirmer que les deux résidus sont bien stationnaires.

On va maintenant vérifier l'hypothèse d'absence d'auto-corrélations dans les processus $(\epsilon_2(t))_t$ et $(\epsilon'_2(t))_t$. On observera d'abord les diagrammes d'auto-corrélations:

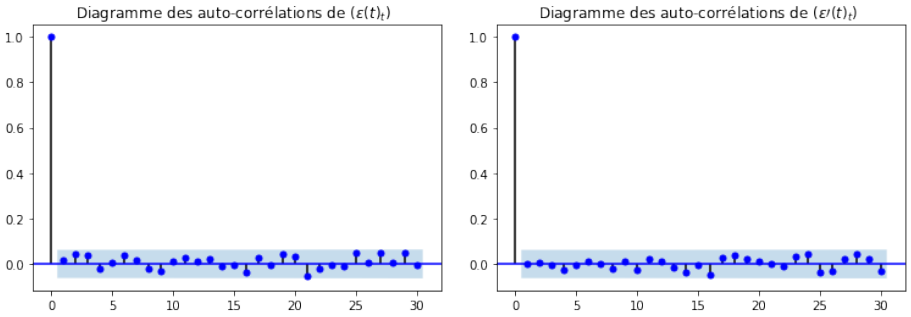

Figure 2.3: Diagrammes d'auto-corrélations de $(\epsilon_2(t))_t$ et $(\epsilon'_2(t))_t$

On peut donc facilement déduire à partir de la figure (2.3) que les résidus ne présentent pas des auto-corrélations significatives. Si besoin, on peut, parallèlement à cette approche graphique, utiliser le test de Ljung-Box dont l'hypothèse nulle est que les données sont distribuées indépendamment. On déduit, d'après ce qui précède que les résidus sont bien des bruits blancs. Le modèle VAR est donc bien défini et adapté à nos données.

Vérifions maintenant l'hypothèse de normalité des résidus par le biais d'un QQ-plot.

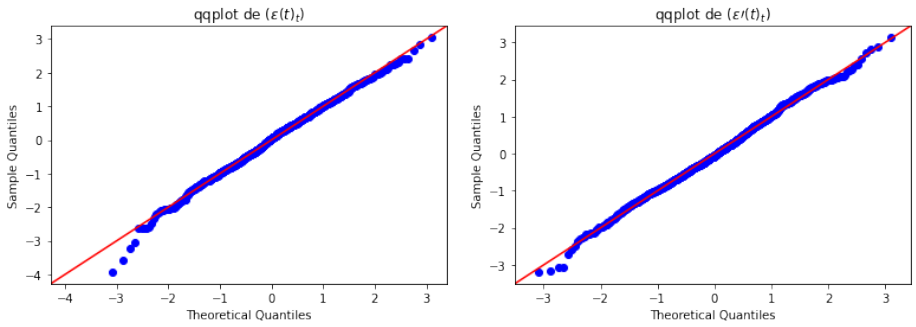

Figure 2.4: qqplots des résidus

On peut facilement déduire le type de distribution à partir de la figure précédente. Tout d'abord, une observation de l'étalement (skewness) de la distribution normale permet d'affirmer que la distribution des résidus est non biaisée (normale). D'autre part, on observe que l'aplatissement (kurtosis) est égale à 3. Ainsi, l'hypothèse de normalité des résidus est vérifiée.

Cette hypothèse de normalité est indispensable pour pouvoir appliquer le test de Granger qui repose sur une statistique de test qui suit à l'asymptotique une loi de Fisher à condition que les résidus soient de loi normale.

La calcul numérique de l'indice de Granger $F_{X1\Rightarrow X2}$ en fonction de a permet le traçage de la figure suivante :

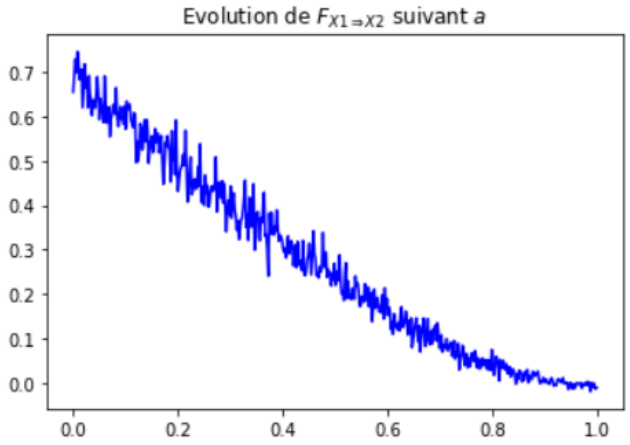

Figure 2.5: Évolution de l'indice de Granger en fonction de a (N=700 observations)

De la figure précédente, on peut remarquer que l'indice de causalité de Granger décroît quand a tend vers 1. Cela veut dire que plus a s'approche de la valeur 1 plus la relations causale $X1 \Rightarrow X2$ devient faible. En effet, cela peut facilement se déduire du modèle génératif : Le fait que a tend vers 1 implique que le coefficient relatif à $x_1(t-1)$ dans la représentation auto-régressive de $x_2(t)$ tend vers 0, et donc $x_2(t)$ ne dépend que de $x_2(t-2)$, d'où la décroissance de l'intensité de la relation causale.

Après avoir mis en application l'indice de Granger, on confirmera dans ce qui suit le résultat précédent grâce au teste Granger. Rappelons que l'hypothèse nulle du test de Granger équivaut à dire que x_1 ne cause pas x_2 relativement à x_3.

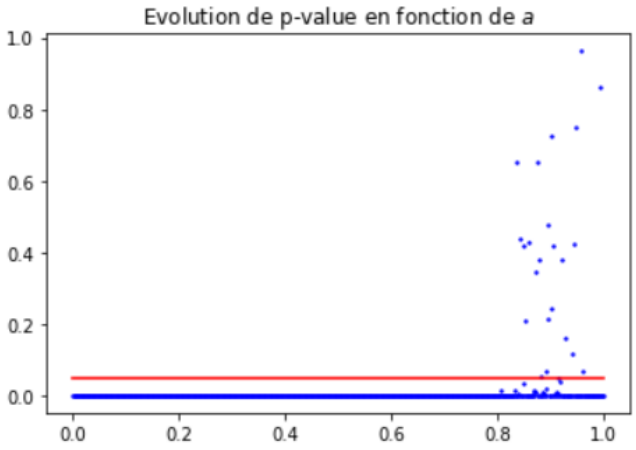

Figure 2.6: Évolution de la p-valeur en fonction de a (N=700 observations) et seuil de 5% (en rouge)

Avant d'aborder l'inférence de causalité non-linéaire, on va essayer de voir à quel point les méthodes précédentes d'inférence causale sont efficaces en présence de relations non linéaires entre les processus étudiés. Pour cela, on introduit le nouveau modèle génératif non linéaire :

$$\begin{aligned} x_1(t) &= 0.95x_1(t-1) - 0.9x_1(t-2) + \tau_1(t) \\ x_2(t) &= (1-a)x_1^2(t-1) + 0.5x_2(t-2) + \tau_2(t) \\ x_3(t) &= -0.7x_1(t-1) + \tau_3(t) \end{aligned} \tag{2.8}$$

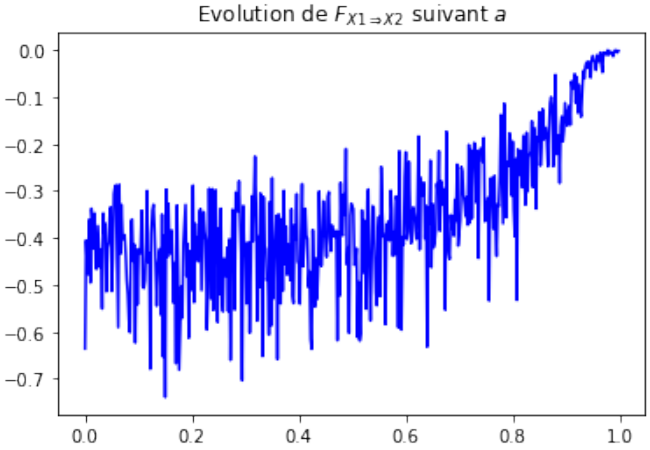

Figure 2.7: Évolution de l'indice de Granger en fonction de a (N=700 observations)

On peut immédiatement remarquer que dans ce cas, la fluctuation de l'indice de Granger est très importante et que l'indice prend des valeurs négatives ce qui est contre intuitif et incohérent avec la construction de l'indice. En effet, l'indice s'écrit sous la forme :

$$F_{y_{i,t} \Rightarrow y_{j,t}} = log(\frac{var(e'_{j,t})}{var(e_{j,t})})$$

Or, quelque soit la structure des processus étudiés on a $var(e'_{j,t}) \geq var(e_{j,t})$ et donc l'indice est toujours positif. Cela montre que l'indice de Granger "basique" n'est

valable que dans le cadre de la causalité linéaire et ne donne aucune information dans le cadre non linéaire, d'où la nécessité d'utiliser d'autres approches à fin d'évaluer la causalité non-linéaire.

2.5.2 Exemple de modèle non linéaire

On considère les trois processus aléatoires suivant :

$$\begin{aligned} x_1(t) &= 0.95\sqrt{2}x_1(t-1) - 0.9x_1(t-2) + \tau_1(t) \\ x_2(t) &= (1-a)x_1^2(t-1) + 0.5x_2(t-2) + \tau_2(t) \\ x_3(t) &= -0.7x_1(t-1) + \tau_3(t) \end{aligned} \tag{2.9}$$

Où $(\tau_i)_{i\in(1,2,3)}$ sont des bruit blancs gaussiens et $a \in [0{:}1]$ un paramètre du modèle.

Pour évaluer la relation causale $(x_1(t))_{t\in\mathbb{Z}} \Rightarrow x_2(t))_{t\in\mathbb{Z}}$, on va se baser sur la méthode des noyaux introduite précédemment. En particulier, on va utiliser un noyau gaussien de paramètre $\sigma = 1$. De plus, pour cette simulation, on utilisera le script Matlab développé par D. Marinazzo [6].

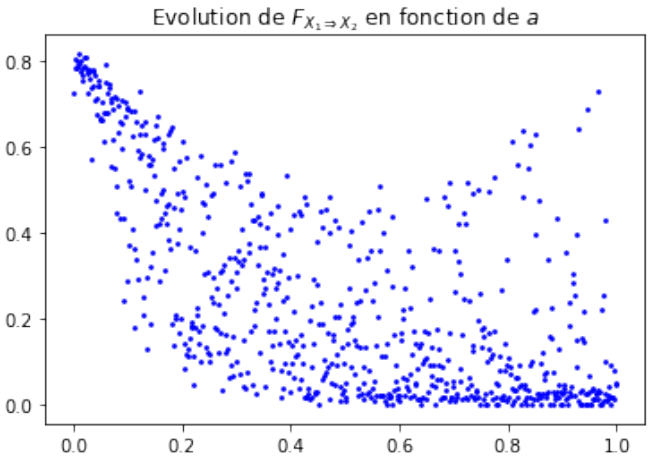

Figure 2.8: Évolution de l'indice de Granger non linéaire en fonction de a (N=700 observations)

2.5.3 Application à des séries financières:

Dans cette partie, on va chercher à inférer des relations de causalité linéaire entre 4 séries temporelles relatives à des indices financiers. On va s'intéresser en particulier au $CAC40$, $GDAX$, $FTSE100$ et au $S\&P500$ entre le 14/06/2010 et le 1/1/2021. (source : Yahoo finance)
Les historiques utilisés sont constitués de données journalières en considérant la valeur des indices à l'ouverture.

Dans la suite on va s'intéresser au logarithme des revenues de ces séries.

Definition 8 Soit le processus $(P_k)_{k\in\mathbb{N}}$ relatif à un prix d'action ou indice boursier, où P_k désigne la valeur à l'instant k. Alors le log-rendement $(r_k)_{k\in\mathbb{N}}$ est défini par :

$$r_k = \log(\frac{P_{k+1}}{P_k}) \; \forall k \in \mathbb{N}$$

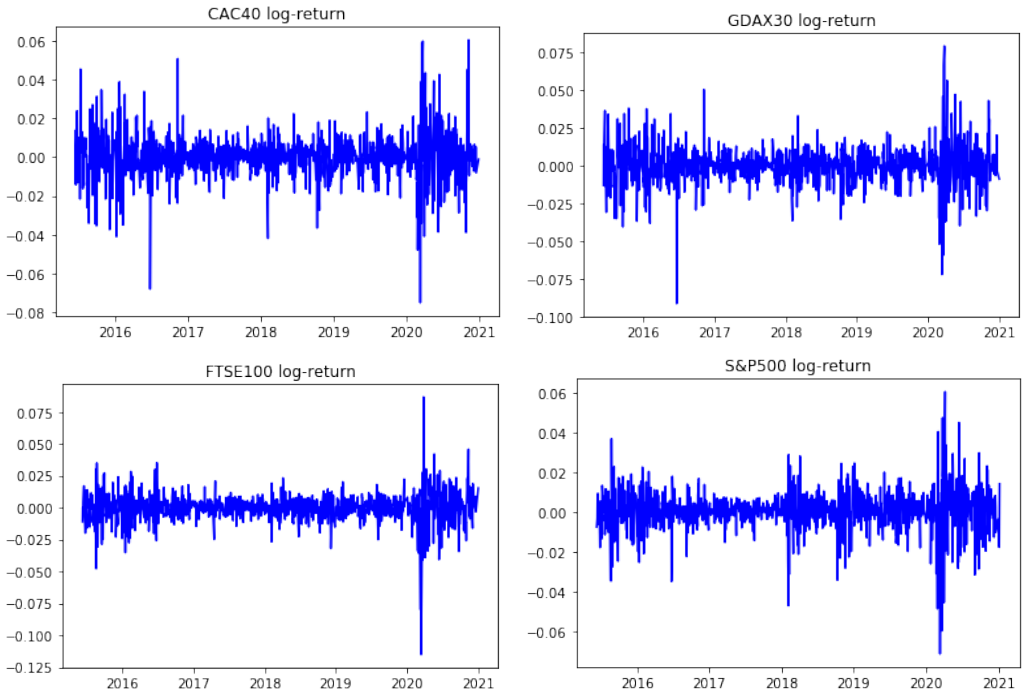

Figure 2.9: log-rendements étudiés

De première vue, on peut dire que les séries sont stationnaires en moyenne (ce qui est confirmé par un test ADF). De plus, elles ne présentent que de faibles auto-corrélations. Or, la variance de ces processus n'est pas stationnaire. En effet, on remarque bien la présence de cluster de volatilité (Hausse/Basse) dans chacune des figures.
Pour contourner ce problème, et donc appliquer un modèle VAR qui va nous permettre d'inférer les relations causales au sens de Granger, on peut opter pour deux approches différentes :

Période mobile (Rolling window)

Cette première approche permet de se placer sur des intervalles de temps réduits de telle sorte que les rendements des séries étudiées soient bien stationnaires. On considère donc dans notre application un intervalle mobile de 300 jours qui translate avec un pas d'une seule journée.
Sur chaque intervalle, on applique la même démarche que dans l'exemple précédent et on calcule les indices de Granger pour les différentes relations causales. C'est ainsi que pour chaque relation causale, on obtient une valeur de l'indice de Granger sur chaque intervalles de temps de longueur égale à 300 jours, et en moyennant ces dernieres on obtient un indice de causalité global. Cela permet de construire des graphes de causalité comme le montre la figure (2.10) dans lesquels chaque noeud représente une série temporelle et chaque arête représente une relation causale et est associé à la valeur de l'indice de Granger correspondant.

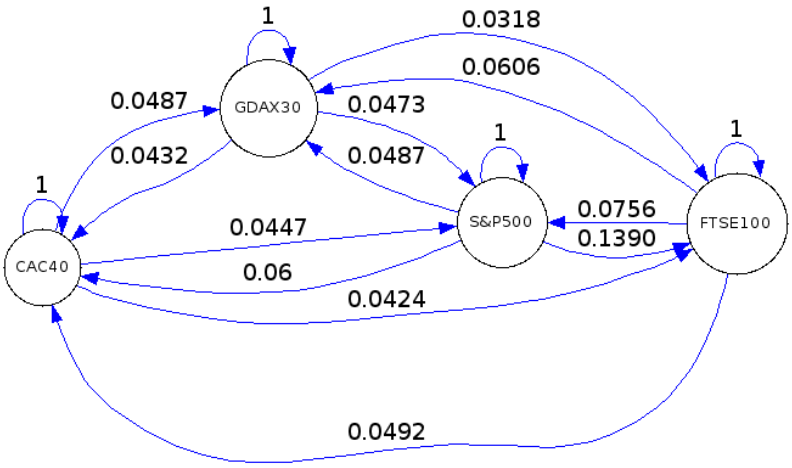

Figure 2.10: Graphe de causalité relatif aux indices CAC40, DAX30, FTSE100 et au SP500

Modèles GARCH

Cette approche repose sur deux étapes essentielles. La première consiste à construire le modèle VAR relatif aux séries temporelles étudiées exactement comme dans l'exemple précédent. Et la deuxième étape consiste à construire un modèle GARCH sur les résidus du VAR de telle sorte qu'on obtient à la fin des résidus stationnaires, gaussiens et ne présentant pas des autocorrelations, qu'on pourrait utiliser pour calculer les différents indices de Granger ou appliquer le test de Granger. Cette approche ne sera pas développer dans ce rapport.

Remark 1 Il faut remarquer que les méthodes d'inférence causales présentées précédemment imposent des conditions restrictives sur les séries temporelles (stationnarité, caractère gaussien, absence d'autocorrelations...), d'où la nécessité d'un long traitement des séries étudiées qui n'aboutit pas toujours à des résultats satisfaisant comme par exemple dans le cas des séries financières. C'est ainsi qu'on va se placer dans un cadre d'étude différent au sens où on ne va pas recourir à une modélisation approximative des séries temporelles et on imposera moins de conditions sur ces dernières.

Chapter 3

Extensions à l'analyse de sensibilité

L'objectif de cette deuxième partie est d'associer l'analyse de sensibilité au concept de causalité et de mettre en place des méthodes d'inférence causale entre séries temporelles. Dans un premier lieu, on abordera quelques concepts fondamentaux de l'analyse de sensibilité en introduisant les indices de Sobol et les valeurs de Shapley. Puis, on cherchera à généraliser ces indices à un cadre temporelles, ce qui nous permettra de les interpréter dans le cas de séries temporelle et puis d'adapter cette approche par analyse de sensibilité à l'inférence causale.

3.1 Méthodes d'analyse globale de sensibilité

L'analyse de sensibilité globale (AS) permet d'analyser un modèle mathématique en étudiant l'impact de la variabilité des facteurs d'entrée du modèle sur la variable de sortie. Elle permet en particulier d'étudier comment l'incertitude sur la variable de sortie est liée à différentes sources d'incertitudes dans les variables d'entrée. Les objectifs de l'analyse de variance (ANOVA) sont nombreux; on peut citer la vérification et la compréhension de modèle, la simplification de modèle et l'optimisation du choix des variables d'entrée.

Dans ce qui suit, les variables d'entrée serons représentés par un vecteur aléatoire $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$. On va se limiter dans ce qui suit à l'étude des modèles à sortie scalaire $Y \in \mathbb{R}$.
Un modèle est définie par une fonction $f(.)$ de tel sorte que

$$\boldsymbol{Y = f(X)} \tag{3.1}$$

3.1.1 Décomposition fonctionnelle de la variance : Indices de Sobol et Input indépendants

Theorem 1 (Hoeffding, 1948, Efron and Stein, 1981, Sobol, 1993) *Soit $X = (X_1, \ldots, X_d)$ un vecteur de variables Input indépendantes et intégrables, de distribution $\mu = \mu_1 \otimes \cdots \otimes \mu_d$, et soit $f : \mathbb{R}^d \to \mathbb{R}$ tel que $f \in L^2(\mathbb{R}^d, \mu)$.*

Il existe une unique représentation de f sous la forme

$$f(X) = f_0 + \sum_{i=1}^{d} f_i(X_i) + \sum_{1\leq i<j\leq d} f_{i,j}(X_i, X_j) + \cdots + f_{1,\ldots,d}(X_1, \ldots, X_d) \tag{3.2}$$

où $\mathbb{E}(f_I(X_I)|X_J) = 0$ *pour tout* $I \subseteq (1, \ldots, d)$ *et tout* $J \subsetneq I$. *De plus,*

$$\begin{aligned} f_0 &= \mathbb{E}(f(X)) \\ f_k(X_k) &= \mathbb{E}(f(X)|X_k) - f_0 \\ f_I(X_I) &= \mathbb{E}(f(X)|X_I) - \sum_{J\subsetneq I} f_J(X_J) \\ &= \sum_{J\subseteq I} (-1)^{|I|-|J|}\mathbb{E}(f(X)|X_J) \end{aligned}$$

Dans le cadre de l'analyse de la variance à entrées indépendantes, on se donne un vecteur aléatoire $X = (X_1, \ldots, X_d)$ où les variables $(X_i)_{i\in(1,\ldots,d)}$ sont mutuellement indépendantes, et une sortie $Y = f(X)$ d'un modèle déterministe $f(.)$. Ainsi, la décomposition de la variance, est donnée par :

$$Var(Y) = \sum_{i=1}^{d} V_i(Y) + \sum_{1<i<j\leq d} V_{i,j}(Y) + \cdots + V_{1,\ldots,d}(Y) \tag{3.3}$$

où

$$\begin{cases} V_i(Y) & = Var(\mathbb{E}(Y|X_i)) \\ V_{i,j}(Y) & = Var(\mathbb{E}(Y|X_i, X_j)) - V_i(Y) - V_j(Y) \\ \cdots & \\ V_{1,\ldots,d} & = Var(Y) - \sum_{i=1}^{d} V_i - \sum_{1\leq i<j\leq d}^{n} V_{i,j} - \cdots - \sum_{1\leq i_1<i_2<\cdots<i_{d-1}\leq d}^{d} V_{i_1,\ldots,i_{d-1}} \end{cases} \tag{3.4}$$

Les "indices de Sobol" sont donnés par:

$$S_i = \frac{V_i(Y)}{Var(Y)}, S_{i,j} = \frac{V_{i,j}(Y)}{Var(Y)} - S_i - S_j, S_{1,\ldots,p} = \frac{V_{1,\ldots,p}(Y)}{Var(Y)} - \sum_{i\in(1,\ldots,p)} S_i \tag{3.5}$$

Ces indices reflètent, **dans le cas d'entrées indépendantes**, la sensibilité de la variance de la variable de sortie Y à une variable entrée donnée ou à l'interaction entre la combinaison de plusieurs variables entrée. En effet, plus $S_{1,\ldots,k}$ est proche de 1 plus l'ensemble de variables $(X_1, \ldots, X_k)$ est important, au sens où les fluctuations des variable de cet ensemble ont un impact considérable sur la sortie du modèle. **Cette interprétation n'est plus valable dans le cas d'entrées dépendantes**.

Remark 2 Le nombre d'indices de Sobol est égale à $2^d - 1$, et lorsque le nombre de variables Input est grand, le nombre d'indices de Sobol explose. Ce qui rend l'interprétation et le calcul de toutes ces valeurs impossible.

C'est ainsi qu'on va introduire des indices de Sobol totaux, permettant d'évaluer la sensibilité totale de la sortie d'un modèle à une variable input. Ces indices incluent non seulement la sensibilité à une variable mais aussi à son interaction avec les autres variables input.

L'indice de Sobol total S_{T_i} relatif à la variable input X_i est défini comme la somme de tous les indices de sensibilité relatifs à la variable X_i :

$$S_{T_i} = \sum_{u \subseteq \{1,\dots,d\}, i \in u} S_u \tag{3.6}$$

Notons aussi que généralement, on se limite au calcul des indices de Sobol du premier et du deuxième ordre ainsi que les indices totaux.

Proposition 6 *D'après la formule* (3.3) *et toujours dans le cadre d'Input indépendants on a*

$$\sum_{u \subseteq \{1,\dots,d\}} S_u = 1. \tag{3.7}$$

Estimation non-paramétrique des indices de Sobol d'ordre 1:

Soit $V_l = \mathbb{E}(\mathbb{E}^2(Y|X_l))$, on a :

$$S_l = \frac{V_l - \mathbb{E}(Y)^2}{Var(Y)} \tag{3.8}$$

Cet indice de Sobol peut être approché par

$$\hat{S}_l = \frac{V_l - \bar{Y}^2}{\hat{\sigma}_Y^2} \tag{3.9}$$

où $\bar{Y} = \frac{1}{n}\sum_{j=1}^n Y_j$ and $\hat{\sigma}_Y^2 = \frac{1}{n}\sum_{j=1}^n (Y_j - \bar{Y})^2$ sont la moyenne et la variance empiriques de Y. En ce qui concerne l'approximation $\hat{V}_l$ de V_l, on utilisera l'estimateur de Nadaraya-Watson.

Definition 9 Supposons qu'on a n couples $(Y^i, X_1^i, \dots, X_d^i) \in \mathbb{R} \times \mathbb{R}^d$ pour $i \in \{1, \dots, d\}$ tel que $\forall i, Y^i = f(X_1^i, X_2^i, \dots, X_d^i)$.
Soit $K : \mathbb{R} \to \mathbb{R}$ un noyau tel que $\int_{\mathbb{R}} K(u)du = 1$ et soit $h > 0$ un pas. On pose aussi $K_h(x) = \frac{1}{h}K(\frac{x}{h})$. Un estimateur de S_l pour tout $l \in \{1, \dots, d\}$ est donné par:

$$\hat{S}_l^{(NW)} = \frac{\frac{1}{n}\sum_{i=1}^n (\frac{\sum_{i=1}^n Y_j K_h(X_l^j - X_l^i)}{\sum_{j=1}^n K_h(X_l^j - X_l^i)})^2 - \bar{Y}^2}{\hat{\sigma}_Y^2} \tag{3.10}$$

Cet estimateur est basé sur l'estimateur de Nadaraya-Watson de $\mathbb{E}(Y|X_l = x)$ donnée par :

$$\frac{\sum_{i=1}^n Y_j K_h(X_l^i - x)}{\sum_{j=1}^n K_h(X_l^j - x)}$$

Pour plus de détails, on renvoie vers [14].

Méthode d'estimation par Monte Carlo: Méthode Pick and Freeze

On considère le modèle précédent $Y = f(X)$. Posons $X = (U, V)$ où $U = (X^{i_1}, \dots, X^{i_k})$ regroupe k facteurs d'entrée et V le complément $(d - k$ facteurs).
La méthode Pick and Freeze repose sur le lemme de Sobol suivant,

Lemma 2 *Si U et V sont indépendantes alors $Var(\mathbb{E}(Y|U)) = Cov(Y, Y^U)$ où $Y^U = f(U, V')$, V' est une copie indépendante de V.*
PROOF On a

$$\begin{aligned} Var(\mathbb{E}(Y|U)) &= \mathbb{E}(\mathbb{E}(Y|U)^2) - \mathbb{E}(\mathbb{E}(Y|U))^2 \\ &= \mathbb{E}(\mathbb{E}(Y|U)^2) - \mathbb{E}(Y)^2 \\ &= \mathbb{E}(\mathbb{E}(Y|U)^2) - \mathbb{E}(Y)\mathbb{E}(Y^U) \\ Cov(Y, Y^U) &= \mathbb{E}(YY^U) - \mathbb{E}(Y)\mathbb{E}(Y^U) \end{aligned}$$

Or, Y et Y^U sont indépendantes conditionnellement à U et ont la même loi. Ce qui fait que,

$$\mathbb{E}(\mathbb{E}(Y|U)^2) = \mathbb{E}(\mathbb{E}(Y|U)\mathbb{E}(Y^U|U)) = \mathbb{E}(\mathbb{E}(YY^U|U)) = \mathbb{E}(YY^U)$$

D'où le résultat du lemme. ■

Alors, l'indice de Sobol peut se réécrire sous la forme suivant:

$$S_U = \frac{Cov(Y^U, Y)}{Var(Y)} \tag{3.11}$$

L'estimateur Pick and Freeze consiste à prendre un estimateur empirique du numérateur et du dénominateur.
Pour $U \subseteq \{1, \ldots, d\}$, on choisit un échantillon de taille N,$\{(Y_1, Y_1^U), \ldots, (Y_N, Y_N^U)\}$.
Un estimateur de S_U **dans le cas d'entrées indépendantes** est

$$\hat{S}_N^U = \frac{\frac{1}{N}\sum_{i=1}^N Y_i Y_i^U - (\frac{1}{N}\sum_{i=1}^N Y_i)(\frac{1}{N}\sum_{i=1}^N Y_i^U)}{\frac{1}{N}\sum_{i=1}^N (Y_i^2) - (\frac{1}{N}\sum_{i=1}^N Y_i)^2} \tag{3.12}$$

Remark 3 En utilisant une taille d'échantillon de Monte Carlo de N, il nous faut $2N$ simulations des variables Input puisqu'on génère deux échantillons différents. Le nombre d'évaluations de la fonction du modèle est donc $2N(k+1)$, où k est le nombre d'indices de Sobol calculés. Et pour d variables input, l'estimation de tout les indices de Sobol nécessite alors $2^d N$ évaluations de la fonction. Or, le fait de se limiter au calcul des indices de premiers ordres et les indices totaux nécessite $(2d+1)N$ évaluations. De plus, la vitesse de convergence de cet estimateur est en $O(\frac{1}{\sqrt{N}})$ et ne dépend pas du nombre de variables d'entrée.
On remarque donc que la méthode de Monte Carlo classique souffre d'un coût de calcul très élevé notamment en grande dimension. Pour remédier à ce problème, l'une des approches est de noter que dans (3.12) Y et Y^U ont la même loi. On peut donc estimer $\mathbb{E}(Y)$ et $\mathbb{E}(Y^2)$ par $\frac{1}{N}\sum_{i=1}^N \frac{Y_i + Y_i^U}{2}$ et $\frac{1}{N}\sum_{i=1}^N \frac{(Y_i)^2 + (Y_i^U)^2}{2}$ respectivement. Ceci donne lieu à un estimateur asymptotiquement efficace [15].

Proposition 7 (Normalité asymptotique) *Supposons que $\mathbb{E}(Y^4) < \infty$. Alors,*

$$\sqrt{N}(\hat{S}_N^U - S^U) \underset{N\to\infty}{\overset{Loi}{\to}} \mathcal{N}(0, \frac{Var((Y - \mathbb{E}(Y))[(Y^U - \mathbb{E}(Y)) - S^U(Y - \mathbb{E}(Y))])}{(Var(Y))^2}) \tag{3.13}$$

PROOF Remarquons d'abord que, d'après (3.12), S^U est invariante par translation de Y et Y^U. De même pour $\hat{S}_N^U$.
On pose donc $T = ((Y - \mathbb{E}(Y))(Y^U - \mathbb{E}(Y^U)), Y - \mathbb{E}(Y), Y^U - \mathbb{E}(Y^U), (Y - \mathbb{E}(Y))^2)$.
On défini la fonction suivante $\theta(w, x, y, z) = \frac{w - xy}{z - y^2}$. On a donc $S^U = \theta(\mathbb{E}(T))$.

D'autre part, soit $\bar{T}$ l'estimateur de Monte Carlo de $\mathbb{E}(T)$. Ainsi on a $\hat{S}_N^U = \theta(\bar{T})$. D'après le théorème central limite on a

$$\sqrt{N}(\bar{T} - \mathbb{E}(T)) \underset{N\to\infty}{\overset{Loi}{\to}} \mathcal{N}(0, \Gamma)$$

où Γ est la matrice de covariance de T. D'après la méthode Delta [theo 2.40 poly MSC Peyre], on a

$$\sqrt{N}(\bar{S}_N^U - S^U) \underset{N\to\infty}{\overset{Loi}{\to}} \mathcal{N}(0, f^t \Gamma f)$$

avec $f = \nabla\theta(\mathbb{E}(T)) = (\frac{1}{Var(Y)}, 0, 0, -\frac{S^U}{Var(Y)})$. D'où,

$$f^t \Gamma f = \frac{Var((Y - \mathbb{E}(Y))[(Y^U - \mathbb{E}(Y)) - S^U(Y - \mathbb{E}(Y))])}{(Var(Y))^2}$$

■

Exemple et application 1 [Modèle linéaire à deux entrées Gaussiennes]
On se place dans le cas où $d = 2$.
Soit le modèle $Y = \beta_0 + \beta X$. Où β_0 est une constante et $\beta = (\beta_1, \beta_2) \in \mathbb{R}^2$.
De plus, $X \sim \mathcal{N}(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix})$, $\rho \in [-1, 1], \sigma_1 > 0, \sigma_2 > 0$.
Dans ce cas on trouve,

$$\begin{aligned} Var(Y) &= \beta_1^2\sigma_1^2 + 2\rho\beta_1\beta_2\sigma_1\sigma_2 + \beta_2^2\sigma_2^2 \\ Var([\mathbb{E}(Y|X_1)) &= (\beta_1\sigma_1 + \rho\beta_2\sigma_2)^2 \\ Var([\mathbb{E}(Y|X_2)) &= (\beta_2\sigma_2 + \rho\beta_2\sigma_2)^2, \\ Var([\mathbb{E}(Y|X_{1,2})) &= Var(Y) \\ Var([\mathbb{E}(Y|X_\varnothing)) &= 0 \end{aligned} \tag{3.14}$$

Et donc,

$$\begin{aligned} \sigma^2 S_1 &= \beta_1^2\sigma_1^2 + 2\rho\beta_1\beta_2\sigma_1\sigma_2 + \rho^2\beta_2^2\sigma_2^2 \\ \sigma^2 S_2 &= \beta_2^2\sigma_2^2 + 2\rho\beta_1\beta_2\sigma_1\sigma_2 + \rho^2\beta_1^2\sigma_1^2 \\ \sigma^2 S_{T_1} &= \beta_1^2\sigma_1^2(1 - \rho^2) \\ \sigma^2 S_{T_2} &= \beta_2^2\sigma_2^2(1 - \rho^2) \end{aligned} \tag{3.15}$$

Étant dans le cadre d'Input à variables indépendantes, on pose $\rho = 0$ et on obtient,

$$\begin{aligned} \sigma^2 S_1 &= \beta_1^2\sigma_1^2 \\ \sigma^2 S_2 &= \beta_2^2\sigma_2^2 \end{aligned}$$

Exemple et application 2 [Modèle non linéaire à deux entrées]
On va s'interesser au modèle défini par :

$$f(X_1, X_2) = (X_1^2 - 1)X_2$$

où X_1 et X_2 sont deux variables aléatoires indépendantes de loi uniforme sur [0,1]. On va d'abord chercher à estimer les indices de Sobol S_1 et S_2 en recourant à aux librairies Python SALib et OpenTURNS qui se basent sur l'estimation par

échantillonage, puis on utilisera la méthode RBD-FAST (Tarantola et al. 2006) qui repose sur la décomposition de l'output Y par transformée de Fourier.

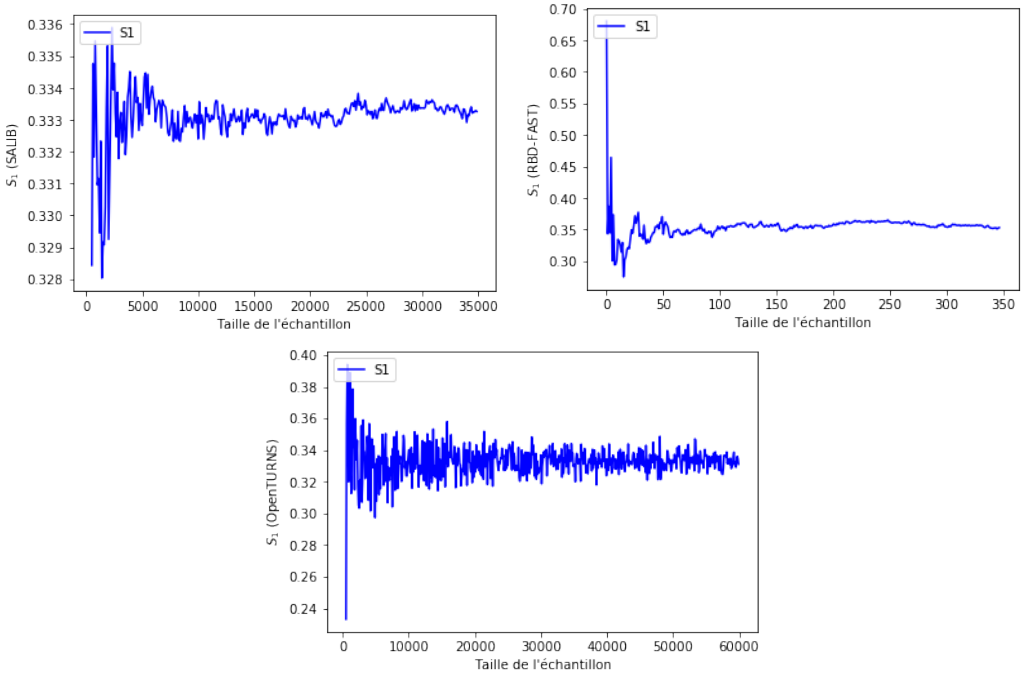

Figure 3.1: Évolution de l'estimateur de l'indice de Sobol S_1 en fonction de la taille N de l'échantillon en utilisant la librairie SALib et OpenTURNS ainsi que la méthode RBD-FAST.

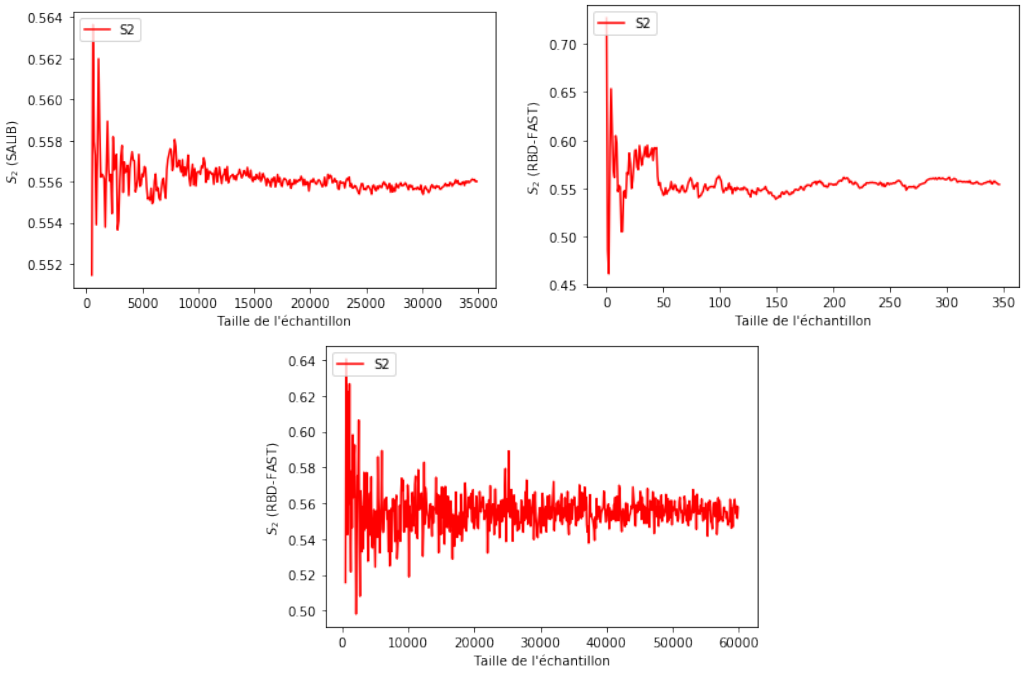

Figure 3.2: Évolution de l'estimateur de l'indices de Sobol S_2 en fonction de la taille N de l'échantillon en utilisant la librairie SALib et OpenTURNS ainsi que la méthode RBD-FAST.

Analytiquement on obtient $S_1 = \frac{1}{3}$ et $S_2 = \frac{5}{9}$, et donc $S_{1,2} = 1 - S_1 - S_2 = \frac{1}{9}$.
En comparant ces valeurs avec les résultats obtenus grâce aux estimations numériques,

on remarque bien que le calcul par le biais de SALib, OpenTURNS et de la méthode RBD-FAST permet d'avoir des estimations satisfaisantes des indices de Sobol (dans ce cas, du premier ordre). Néanmoins, la méthode RBD-FAST estime des indices de Sobol de premier ordre robustes et précis avec seulement un jeu d'échantillons de N simulations de l'ordre de quelques centaines d'évaluations, quel que soit le nombre de paramètres.

Indices	Valeur	Taille de l'échantillon
S_1	0.333	–
S_2	0.555	–
S_1 (RBD-FAST)	$0.3309_{+/-0.040751}$	$N = 350$
S_2 (RBD-FAST)	$0.553_{+/-0.04202}$	$N = 350$
S_1 (SALib)	$0.3331_{+/-0.0007548}$	$N = 35000$
S_2 (SALib)	$0.5561_{+/-0.0008215}$	$N = 35000$
S_1 (OpenTURNS)	$0.3318_{+/-0.05616}$	$N = 60000$
S_2 (OpenTURNS)	$0.55870_{+/-0.06956}$	$N = 60000$

Remarque L'ANOVA par SALib, OpenTURNS et RBD-FAST n'est valable que dans le cadre d'input indépendants.

3.1.2 Décomposition fonctionnelle de la variance : Indices de Shapley et Input dépendants

Dans cette partie, on va s'intéresser aux modèles à Input dépendants. L'objectif ici reste le même, à savoir, évaluer l'impacte d'une ou plusieurs variables Input sur l'output d'un modèle ainsi que l'importance de ces dernières.
Comme indiqué dans la partie précédente, les indices de Sobol ne permettent plus d'analyser l'impact de l'Input dans ce cas-là. C'est pour cette raison qu'on va s'intéresser aux indices de Shapley qui eux permettent de capturer les interactions entre les variables Input et de mieux expliquer la variance de l'output du modèle.
L'indice ou valeur de Shapley est un concept issu de la théorie des jeux coopératives, qui a été adapté à la mesure de l'importance d'une variable dans un modèle.
Dans le cadre des jeux coopératives, la valeur de Shapley Sh_i d'un jouer i est la valeur moyenne du gain qu'il apporte lorsqu'il intègre les équipes u constituées par les autres joueurs.
On pose comme fonction de gain, proposée par A.Owen(2014),

$$c(u) = S_u = \frac{Var(\mathbb{E}(Y|X_u))}{Var(Y)} \tag{3.16}$$

où u est un sous ensemble de $(1, \ldots, d)$. Dans ce cadre, u peut être vue comme une coalition.
En effet, dans le cadre de l'ANOVA, A.Owen proposa la fonction de gain (3.16) puisque pour $I \subseteq (X_1, \ldots, X_d)$, $c(I)$ mesure la part de variance de Y causée par l'incertitude des variables dans I. De plus, $c(\varnothing) = 0$ et $c((X_1, \ldots, X_d)) = Var(Y)$.
Definition 10 Les valeurs de Shapley correspondantes à la fonction de coût c(.) défini précédemment sont appelés les effets de Shapley. Les effets de Shapley Sh_i

sont définis par :

$$Sh_i = \sum_{u \subseteq -\{i\}} \frac{(d - |u| - 1)!|u|!}{d!}[c(u \cup \{i\}) - c(u)] \tag{3.17}$$

où $-\{i\}$ est l'ensemble des indices $(1, \ldots, d)$ ne contenant pas i.

Les effets de Shapley résultent d'un partage équitable de la variance de l'output du modèle entre les input de ce dernier.

Les valeurs de Shapley peuvent aussi être exprimées en terme de permutations de l'ensemble des variable entrées $(X_1, X_2, \ldots, X_d)$. Soit $\prod(X_1, \ldots, X_d)$ l'ensemble des permutations de $(X_1, \ldots, X_d)$. Pour $\lambda \in \prod(X_1, \ldots, X_d)$, on note $P_i(\lambda)$ les variables qui précédent la variable X_i dans λ.

Definition 11 Les valeurs de Shapley peuvent être réécrites telles que

$$Sh_i = \frac{1}{d!} \sum_{\lambda \in \prod(X_1, \ldots, X_d)} [c(P_i(\lambda) \cup \{i\}) - c(P_i(\lambda))] \tag{3.18}$$

Remark 4 D'après l'expression (3.17), on voit bien qu'il faut calculer dans chaque itération dans la boucle relatif à la somme, deux indices de Sobol. Ainsi, en utilisant une méthode d'estimation par échantillonnage du type Pick Freeze, la complexité combinatoire pour une seule itération est de l'ordre de $2N$ appels de la fonction modèle comme on la vu précédemment. De plus, le calcul total de la somme nécessite 2^d calcul ou estimation du gain $c(u)$. Ainsi, la complexité combinatoire relatif au calcul d'une seul valeur de Shapley est $2^{d+1}N$ appel de la fonction f du modèle.

Proposition 8 $\sum_{i=1}^{d} Sh_i = 1$ *par construction. Ainsi, par rapport aux indices de Sobol, les indices Shapley restent interprétables, qu'il y ait une structure de dépendance ou non, car ils somment toujours à un par construction.*

Dans le cadre d'entrées indépendantes, l'effet de Shapley Sh_i associé à l'entrée X_i évalue la part de variance expliquée par l'interaction entre un sous ensemble X_u de l'input, où $u \subseteq (1, \ldots, d)$ tel que $i \in u$ et chaque variable individuelle de l'input. En effet on a

$$Sh_i = \sum_{u \subseteq (1,\ldots,d), i \in u} \frac{S_u^2}{|u|} \tag{3.19}$$

Or dans le cadre d'entrées dépendantes, l'effet de Shapley associé à l'entrée X_i prend en compte non seulement les interactions, mais aussi les corrélations de X_i avec X_j, $1 \leq j \leq d, j \neq i$.

Exemple et application 3: [Modèle linéaire à deux entrées Gaussiennes]
On reprend l'exemple du modèle linéaire gaussien introduit précédemment.
En utilisant (3.17), on trouve :

$$\begin{aligned} Var(Y)Sh_1 &= \beta_1^2\sigma_1^2(1 - \frac{\rho^2}{2}) + \rho\beta_1\beta_2\sigma_1\sigma_2 + \beta_2^2\sigma_2^2\frac{\rho^2}{2} \\ Var(Y)Sh_2 &= \beta_2^2\sigma_2^2(1 - \frac{\rho^2}{2}) + \rho\beta_1\beta_2\sigma_1\sigma_2 + \beta_1^2\sigma_1^2\frac{\rho^2}{2} \end{aligned} \tag{3.20}$$

Remarquons d'abord que dans le cas de variables indépendantes ($\rho = 0$) on a, $Sh_1 = S_1$ et $Sh_2 = S_2$. D'autre part, $Sh_1 + Sh_2 = 1$.

Analysons maintenant le comportement de Sh_1 et Sh_2 en fonction du coefficient de corrélation ρ. Pour cela, on pose $\beta_1 = 1$ et $\beta_2 = 2$ et $\sigma_1 = 0.5$ et $\sigma_2 = 1$.

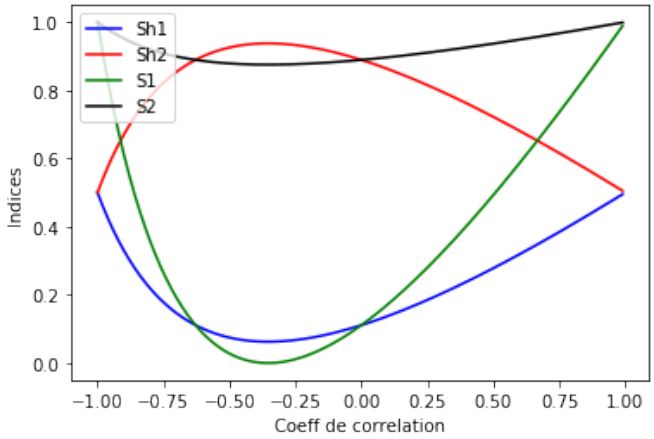

Figure 3.3: Évolution des indices de Shapley Sh_1 et Sh_2 et de Sobol S_1 et S_2 en fonction de ρ

En analysant la figure (3.3), on peut bien remarquer que l'effet de dépendance implique un équilibre entre les indices de Shapley des Inputs (l'effet de corrélation est partagé par Sh_1 et Sh_2). D'autre part, pour des Input parfaitement corrélés, les indices de Shapley sont égaux. De plus, remarquons la grande différence entre les valeurs de Shapley et les indices de Sobol dans le cas d'Input dépendant, ce qui reflète l'idée que les deux types d'indices n'ont pas la même interprétation dans le cadre de l'ANOVA avec entrées dépendantes. Néanmoins, ces indices sont égaux dans le cas d'indépendance, comme le montre la figure et aussi le calcul fait précédemment.

3.2 Quelques généralisations des valeurs de Shapley

3.2.1 Définitions, propriétés et lemmes

Dans la section précédente, on a introduit les valeurs de Shapley pour chaque variable input. Or, il serait intéressant et surtout très utiles d'introduire la notion de valeurs de Shapley pour un ensemble (coalition) de variables inputs. Cela va nous permettre de quantifier de façon plus exacte l'influence des variables d'entrées sur la sortie du modèle.
Comme précédemment, on considère une fonction modèle $f : \mathbb{R}^d \to \mathbb{R}$ qui prend $(X_1, X_2, \ldots, X_d)$ comme variables d'entrée et admet Y comme variable de sortie.

Definition 12 La valeur de Shapley, pour une fonction de coût v liée au modèle, relative à une coalition $I, I \subseteq (X_1, \ldots, X_d)$, est égale à la somme des valeurs de Shapley, pour la même fonction de coût, de chaque variable X_i, $i \in I$,

$$Sh_I((X_1, \ldots, X_d), v) = \sum_{i \in I} Sh_i((X_1, \ldots, X_d), v) \tag{3.21}$$

Cette définition permet d'introduire une approche plus générale pour les indices de Shapley ainsi que pour leur interprétation. En effet, pour $I \subseteq (X_1, \ldots, X_d)$, Sh_I

quantifie non seulement l'impact des variables dans I, prise individuellement, sur la sortie du modèle, mais aussi l'impact des interactions des variables au sein de la coalition I sur la sortie du modèle.
Dans le lemme suivant, on va introduire une nouvelle formulation des valeurs de Shapley ainsi qu'un nouveaux type de ces valeurs.

Lemma 3 *La valeur de Shapley $Sh_i((X_1, \ldots, X_d), v)$ relative à la variable X_i est décomposées en d différentes valeurs $Sh_{i,j}$, $j \in (1, \ldots, d)$, tel que*

$$Sh_i((X_1, \ldots, X_d), v) = \sum_{j=1}^{d} Sh_{i,j}((X_1, \ldots, X_d), v) \tag{3.22}$$

avec

$$Sh_{i,j}((X_1, \ldots, X_d), v) = \sum_{S \subseteq (1,\ldots,d),(i,j) \in S} \frac{(|S|-1)!(d-|S|)!}{d!}(Sh_j(S, v) - Sh_j(S \backslash (i), v)) \tag{3.23}$$

et, $Sh_i(S \setminus (i), v)) = 0$.
De la même façon on a,

$$Sh_j((X_1, \ldots, X_d), v) = \sum_{i=1}^{d} Sh_{i,j}((X_1, \ldots, X_d), v)$$

Ainsi, la valeur de Shapley $Sh_{i,j}$ quantifie l'importance de la variable X_i pour X_j dans l'influence sur la sortie Y du modèle. En d'autres termes, elle donne une idée sur ce que rapporte la présence de X_i pour X_j dans le modèle en terme d'impact sur Y.
On peut aussi introduire le même concept pour deux coalitions I et J.

Lemma 4 *La valeur de Shapley $Sh_{I,J}$ pour deux coalitions I et J, $I \subseteq (1, \ldots, d)$, $J \subseteq (1, \ldots, d)$ est*

$$Sh_{I,J}((X_1, \ldots, X_d), v) = \sum_{i \in I, j \in J} Sh_{i,j}((X_1, \ldots, X_d), v) \tag{3.24}$$

Dans ce qui suit, on va introduire une définition probabiliste des valeurs de Shapley qui va s'avérer fort utile pour l'approximation de ces dernières. Cette définition résulte de la méthode de l'extension multilinéaire des valeurs de Shapley introduite par Owen [ref].
Pour cette définition on aura besoin d'introduire une structure probabiliste sur l'espace des variables Input. Pour un $j = 0, 1, \ldots, d$ fixé et un sous-ensemble aléatoire U_j de $\{1, \ldots, d\} \backslash \{j\}$, la probabilité que X_i tel que $i \in \{1, \ldots, d\} \backslash \{j\}$ est égale à q avec $0 < q < x1$. En d'autres termes, $\mathbb{P}(X_i \in U_j) = q$ pour tout $i \in \{1, \ldots, d\} \backslash \{j\}$. D'autre part, on suppose que le tirage est tel que les événements $\{X_i \in U_j\}$ sont mutuellement indépendants.

Definition 13 (Extension multilinéaire des valeurs de Shapley) Le modèle précédent étant défini, une représentation probabiliste des valeurs de Shapley pour la fonction de coût c est donnée par

$$Sh_i = \int_0^1 e_i(q)dq \tag{3.25}$$

où

$$e_i(q) = \mathbb{E}(c(I^{(q)} \odot X + X^{(i)}) - c(I^{(q)} \odot X)). \tag{3.26}$$

$I^{(q)}$ est un vecteur aléatoire dont chaque composante suit une loi de Bernoulli de paramètre q. X est le vecteur des variables Input et $X^{(i)}$ est un vecteur dont toutes les composantes sont égales à 0 sauf la $i$ème composante qui est égale à X_i.

3.3 Estimation et calcul des valeurs de Shapley

L'une des contraintes majeures lors de l'utilisation des indices de Shapley est le coût de calcul qui est très important. C'est pour cette raison qu'on va recourir à des estimations et approximations qui vont réduire considérablement la complexité du calcul.

3.3.1 Estimation par tirage aléatoire uniforme des coalitions

Dans cette approche on recourt à un tirage aléatoire de m coalitions des variables input. Cela va permettre de réduire le nombre de termes dans la somme qui défini les valeurs de Shapley et par conséquent le nombre de fois qu'on fait appel à la fonction de coût. Ainsi, une approximation de la valeur de Shapley Sh_i relative à la variable X_i est donnée par

$$\hat{Sh}_i = \sum_{u \in \mathbf{A}} \frac{(d - |u| - 1)!|u|!}{d!}[c(u \cup \{i\}) - c(u)] \tag{3.27}$$

où $\mathbf{A}$, de cardinal m, est l'ensemble des coalitions tirées aléatoirement de $-\{i\}$.
La méthode de tirage la plus classique, au sens où elle ne permet pas d'inclure un biais ou un déséquilibre au niveau de la taille des coalitions choisies, est le tirage uniforme sur $\{1, \ldots, d\}$. Pour le tirage uniforme d'une coalition $S \subset \{1, \ldots, d\}$, on se base sur le schéma suivant :

1. On choisi un cardinal $s = 1, 2, \ldots, d$ de la coalition S uniformément.
2. On tire une coalition S de taille s uniformément de $\{1, \ldots, d\}$.

3.3.2 Estimation par tirage aléatoire structuré des coalitions

La méthode du tirage aléatoire structuré [20] est une variante de la méthode du tirage uniforme qui assure que la contribution marginale d'une variable input à une coalition de même taille (i.e., $c(u \cup \{i\}) - c(u)$) est calculée le même nombre de fois. Cela conduit à une meilleure approximation car le calcul des contributions marginales par rapport aux coalitions de même taille est distribué de façon égale. Ce tirage se fait en suivant la procédure suivante :

1. On tire uniformément r permutations de $\{1, \ldots, d\}$ tel que $r = t \times d$. On note $\mathbf{A}$ l'ensemble de ces r permutations.
2. On divise $\mathbf{A}$ en d groupes de taille t.
3. Pour chaque variable X_i :

(a) On échange X_i avec la variable dans la position j pour chacune des t permutations du groupe j, avec $j \in \{1, \ldots, d\}$. $\mathbf{A}$ devient donc $\mathbf{A}'$.

(b) On calcule la valeur de Shapley relative à la variable X_i :

$$\hat{Sh}_i = \frac{1}{d!} \sum_{\lambda \in \mathbf{A}'} [c(P_i(\lambda) \cup \{i\}) - c(P_i(\lambda))]$$

Comparaison entre l'estimation par tirage uniforme et par tirage structuré

Pour mener cette comparaison, on simulera N_1 jeux G_i composés de N_2 joueurs et tel que pour chaque coalition $U^j_{G_i}$ du jeu G_i, on associe une valeur $v_j = v(U^j_{G_i}) \in [0:1]$. On calcule ensuite la moyenne des erreurs absolues moyenne (AAAE) de chaque jeu simulé et qui est définie par

$$AAAE = \frac{1}{N_1} \sum_{j=1}^{N_1} \left(\frac{1}{N_2} \sum_{i=1}^{N_2} |\hat{Sh}_i(v_j) - Sh_i(v_j)| \right) \tag{3.28}$$

où $\hat{Sh}_i(v_j)$ est l'estimateur de la valeur de Shapley $Sh_i(v_j)$ pour le joueur i dans le jeu j. L'AAAE sera calculée pour toute taille d'ensemble de permutations tirées pour l'estimation des valeurs de Shapley ce qui permettra de suivre la convergence des estimateurs dans le cas de l'approche par tirage uniforme et par tirage structuré. Dans la figure (3.4), on a fixé $N_1 = 15$ et $N_2 = 6$.

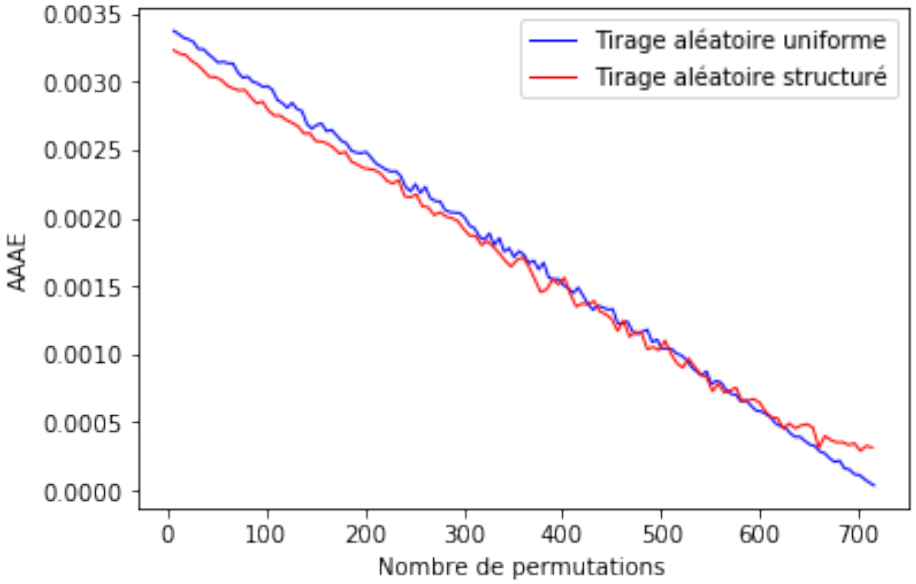

Figure 3.4

	AAAE_{Unif}	AAAE_{struc}
5 permutations	0.003371	0.003229
156 permutations	0.002691	0.002559
566 permutations	0.00073641	0.0007168
671 permutations	0.0002334	0.0003687
716 permutations	$3.49288e-05$	0.00031342

D'après le tableau précédent et la figure (3.4), on remarque que pour un nombre N de permutations tirées aléatoirement tel que $N << d!$ où d est le nombre total de joueurs, le tirage aléatoire uniforme est plus performant que le tirage uniforme. Néanmoins, la différence de performance reste petite. D'autre part, le temps de calcul de la méthode de tirage structuré est un peu plus grand que celui de la méthode du tirage uniforme. Cela est du aux opérations d'échange de joueurs dans les permutations choisies.

3.3.3 Estimation par Owen Sampling

Dans cette approche on va se baser sur l'expression (3.25) des valeurs de Shapley. L'idée est de discrétiser (3.25). Comme première approche, on utilisera la méthode des rectangles pour le calcul de l'intégrale et on va recourir à un estimateur empirique de l'espérance. Cela va nous amener à introduire deux nouveaux paramètres q et m qui représentent respectivement le pas de discrétisation de l'intégrale et la taille de l'échantillon pour l'estimation de l'espérance.
L'algorithme de calcul est basé sur une boucle externe qui approche l'intégrale de la formule (3.25) et une boucle interne qui utilise un estimation par méthode de Monte Carlo pour l'estimation de (3.26). D'autre part, le tirage des coalitions se fait de façon uniforme, ce qui est assuré par le biais du vecteur $I_m^{(q)}$ qui est composé de d résultats de tirages indépendants de Bernoulli. Ce vecteur est utilisé pour calculer la contribution marginale $h_{m,j}^{(q)}$. Dans la ligne 6, l'opérateur $\odot$ désigne la multiplication terme à terme et est appliqué aux vecteurs X et $I_m^{(q)}$. $X^{(j)}$ dans la ligne 6 représente le vecteur $(0, 0, \ldots, x_j, 0, \ldots, 0)$. L'approximation non normalisée du vecteur des valeurs de Shapley est actualisée dans la ligne 9. Pour normaliser ces valeurs de Shapley, on divise par le nombre total d'itérations QM.

Algorithm 2 Owen sampling simple pour l'estimation des $\hat{Sh}_i$:

Require: $c, X = (x_0, x_1, \ldots, x_n), Q, M$
Ensure: $\hat{Sh} = (\hat{Sh}_0, \hat{Sh}_1, \ldots, \hat{Sh}_d)$
  $\hat{S}_j = 0$ for all j
  for q = 0,1/Q, 2/Q,...,1 **do**
    $e_j = 0$ for all j
    for m = 1,2,...,M **do**
      $I_m^{(q)} \leftarrow (b_j : b_j \sim Bernoulli(q)$, for all j)
      $h_{m,j}^{(q)} \leftarrow c(I_m^{(q)} \odot X + X^j) - c(I_m^{(q)} \odot X)$
      $e_j \leftarrow e_j + h_{m,j}^{(q)}$, for all j
    end for
    $\hat{Sh}_j \leftarrow \hat{Sh}_j + e_j$, for all j
  end for
  $\hat{Sh}_j \leftarrow \hat{Sh}_j / QM$, for all j

La convergence de l'algorithme est garantie par la proposition suivante :

Proposition 9 *Pour un j fixé, si la fonction $q \to e_j(q)$ est intégrable au sens de Riemann (ou Lebesgue), alors les estimateurs de l'algorithme (3) convergent vers les vraies valeurs de Shapley quand $Q, M \to \infty$.*

PROOF En recourant au modèle probabiliste sur lequel se base la définition (13), (3.26) peut s'écrire sous la forme équivalente suivante :

$$e_j(q) = \mathbb{E}(c(I^{(q)} \odot X + X^{(j)}) - c(I^{(q)} \odot X)),$$

où $I^{(q)}$ est un vecteur aléatoire dont chaque composante suit une loi de Bernoulli de paramètre q.
Puisque $q \to e_j(q)$ est Riemann intégrable, (3.25) peut s'écrire sous la forme:

$$Sh_j = \lim_{Q\to\infty} \sum_{i=0}^{Q} \frac{1}{Q} e_j(i/Q). \tag{3.29}$$

D'autre part, pour un i,$0 \leq i \leq Q$ fixé, $e_j(i/Q)$ peut être approchée par l'estimateur de Monte-Carlo suivant,

$$\hat{e}_j(i/Q) = \frac{1}{M} \sum_{m=1}^{M} (c(I_m^{(i/Q)} \odot X + X^{(j)}) - c(I_m^{(i/Q)} \odot X))),$$

où $I_m^{i/Q}$, pour un m fixé, est un vecteur de variables aléatoires de Bernoulli de paramètre i/M et $\hat{e}_j(i/Q)$ converge presque sûrement vers $e_j(i/Q)$.
Ainsi (3.29) est égale à

$$\frac{1}{Q \times M} lim_{Q\to\infty} lim_{M\to\infty} \sum_{i=0}^{Q} \sum_{m=1}^{M} (c(I_m^{(i/Q)} \odot X + X^{(j)}) - c(I_m^{(i/Q)} \odot X))).$$

On définit donc l'estimateur convergeant $\hat{Sh}_j$ du vecteur valeurs de Shapley des variables Input, qui coïncide avec celui utilisé dans l'algorithme (.), par

$$\hat{Sh}_j = \frac{1}{Q \times M} \sum_{i=0}^{Q} \sum_{m=1}^{M} (c(I_m^{(i/Q)} \odot X + X^{(j)}) - c(I_m^{(i/Q)} \odot X))). \tag{3.30}$$

■

Remark 5 L'avantage majeur de l'estimation des valeurs de Shapley par Owen sampling est la possibilité de réduire considérablement le temps de calcul et d'améliorer la précision de l'estimateur en menant des calculs avec Q et M grands et vus comme des hyper-paramètres.

Dans (3), l'idée était de discrétiser l'intégrale par la méthode des rectangle. Or, il est possible d'opter pour d'autre méthodes d'intégration numérique, moyennant un temps de calcul plus important, chose qu'on n'abordera pas dans ce rapport.
Une autre approche différente et qui mérite d'être citée est le recours à une méthode double Monte Carlo pour estimer (3.25). En effet, on a

$$Sh_i = \int_0^1 e_i(q)dq = \mathbb{E}^{Q\sim Uniform([0,1])}(e_i(Q)) \tag{3.31}$$

D'après (3.31), on remarque bien qu'on peut estimer les deux espérances définissant la valeur de Shapley par méthode de Monte Carlo. L'avantage de cette approche est qu'elle permet de mettre en oeuvre les différents outils de la réduction de la variance pour améliorer la précision et diminuer le temps de calcul.

Algorithm 3 Owen sampling par double MC pour l'estimation des $\hat{Sh}_i$:

Require: $c, X = (x_0, x_1, \ldots, x_n), Q, M$
Ensure: $\hat{Sh} = (\hat{Sh}_0, \hat{Sh}_1, \ldots, \hat{Sh}_d)$
  $\hat{S}_j = 0$ for all j
  for q = 1,2,...,Q **do**
    $e_j = 0$ for all j
    $p \leftarrow Uniform([0:1])$
    for m = 1,2,...,M **do**
      $I_m^{(p)} \leftarrow (b_j : b_j \sim Bernoulli(p)$, for all j)
      $h_{m,j}^{(p)} \leftarrow c(I_m^{(p)} \odot X + X^j) - c(I_m^{(p)} \odot X)$
      $e_j \leftarrow e_j + h_{m,j}^{(p)}$, for all j
    end for
    $\hat{Sh}_j \leftarrow \hat{Sh}_j + e_j/M$, for all j
  end for
  $\hat{Sh}_j \leftarrow \hat{Sh}_j/Q$, for all j

Remark 6 En pratique, on couple les méthodes d'estimation présentées précédemment avec l'estimation de la fonction de coût $c(.)$ puisque cette dernière est généralement difficile à calculer exactement ou il n'existe tout simplement pas de formule explicite pour cette derniere.

Chapter 4

Application des valeurs de Shapley à des données du CAC40

4.1 Présentation des données

Les données financières sont des cotations des 7 plus grandes valorisations du CAC40 rassemblées par M. Georges Oppenheim. Ces séries temporelles sont échantillonnées au pas de temps constant de 15 secondes du 01/07/2020 à 09 : 00 : 00 au 30/10/2020 à 17 : 30 : 00. La présente étude s'intéresse au prix de l'action de d'Air Liquide. La figure (4.1) montre l'évolution du prix de cette action.

Figure 4.1: Evolution du prix de l'action d'Air Liquide du 01/07/2020 à 09 : 00 : 00 au 30/10/2020 à 17 : 30 : 00

La stationnarité des séries temporelles étudiées ne pose pas de problèmes dans la présente étude. De ce fait, aucune transformation (par exemple de type Box-Cox) ne sera appliquée aux données.

4.2 Présentation de l'objectif

On a vu dans le chapitre précédent que les valeurs de Shapley permettent de quantifier l'importance des variables d'entrée d'un modèle mathématique sur la sortie de ce dernier et donc d'expliquer et d'interpréter le résultat d'un modèle. Cette idée va être exploitée. De plus, elle permet de déterminer d'éventuelles relations de causalité entre les différents instants d'un processus.
Soit G une fonction de $\mathbb{R}^d \to \mathbb{R}$ qui représente le modèle de prédiction et d un paramètre. Soit $(X_t)_{t\in\mathbb{N}}$ une série temporelle à valeurs réelles. G définit le prédicteur $\hat{X}_t$ par $\hat{X}_t = G(X_{t-1}, X_{t-2}, \ldots, X_{t-d})$. G est un modèle de prédiction pouvant appartenir à la famille AR ou ARIMA ou SARIMA,... ou encore être un modèle de machine learning comme la régression K-NN ou la régression par random forest ou encore un modèle de réseaux de neurones récurrents (GRU, LSTM,...). Ces modèles étant souvent paramétrés. on dira alors que G est définie sur $\mathbb{R}^{d+s}$ avec $s \in \mathbb{N}$, et que $\hat{X}_t = G(X_{t-1}, X_{t-2}, \ldots, X_{t-d}, \theta)$ où $\theta \in \mathbb{R}^s$ est un paramètre.

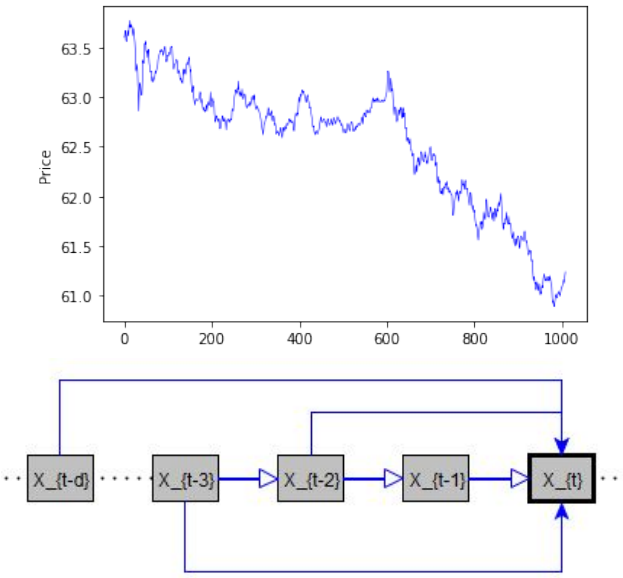

Figure 4.2

G étant définie, l'objectif est de calculer la valeur de Shapley pour chaque instant $t-p$, $p \in (1, \ldots, d)$, contribuant à la prévision de l'instant t: les indices de Shapley pour chaque $X_{t-p}, p \in (1, \ldots, d)$ (2.10). Les valeurs de Shapley sont évaluées sur chaque fenêtre temporelle de la série.
Cette approche nous permettra d'identifier des comportements propre à chaque séries temporelles, d'expliquer des phénomènes de variation brutale ou des tendances de longues durées et de faire des prédictions plus ciblées.

4.3 Approche adoptée

Pour des raisons de complexité, en temps et mémoire, des calculs [1], et comme première étape, le modèle de prédiction choisi est la régression kNN (1.3) qui a le bon goût d'être facile à mettre en oeuvre, de ne faire appel qu'à un seul paramètre k et qui souvent produit des résultats de prédictions à horizon court satisfaisants. Dans la suite de l'analyse, on pose $k = 2$ et $G = G_{kNN}$.
D'autre part, on choisit les $d = 9$ instants précédents l'instant t pour prédire l'instant t.

On a vu précédemment (3.17), (13) que les valeurs de Shapley sont définies en fonction d'un coût, noté $c(.)$. Dans le cadre des séries temporelles et de la prévision, on choisit:

$$c = 1 - \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} |y_i^{test} - \hat{y}_i^{test}| \tag{4.1}$$

où N_{test} est le cardinal de la base de données test, $(\hat{y}_i)_{1 \leq i \leq N_{test}}$ représente les valeurs prédites et $(y_i)_{1 \leq i \leq N_{test}}$ les valeurs à prédire et effectivement observées.

Il faut aussi noter qu'il est possible d'estimer les valeurs de Shapley sur tout intervalle de temps. Les estimations seront de bonne qualité si l'intervalle est assez large.

[1] Dans tout ce chapitre, les calculs sont fait sur l'environnement d'exécution de Google Colab

4.4 Résultats

Cette partie étudie des valeurs de Shapley selon l'approche (3) puis selon (.).

4.4.1 Estimation des valeurs de Shapley d'après (3)

Tout d'abord, on fixe $Q = 50$ et $M = 50$ qui correspondent respectivement au pas de discrétisation de l'intégrale par la méthode des rectangles et à la taille de l'échantillon de l'estimateur de Monte Carlo de la définition (13). Dans un premier temps, les valeurs de Shapley sont estimées sur des intervalles de temps de longueur égale à 1010 ($\sim$ 252 minutes). A titre d'exemple, on obtient les figures suivantes:

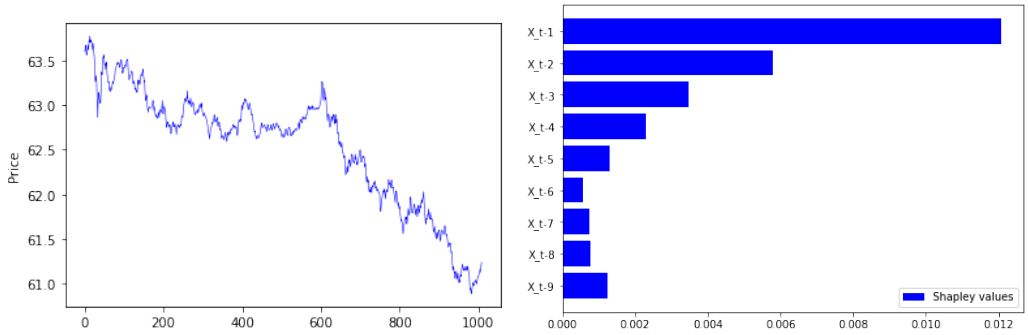

Figure 4.3: A droite, l'estimation selon (3) des valeurs de Shapley relatives à l'évolution du prix (à gauche) de 2020-01-07 09:00:00 à 2020-01-07 13:12:15

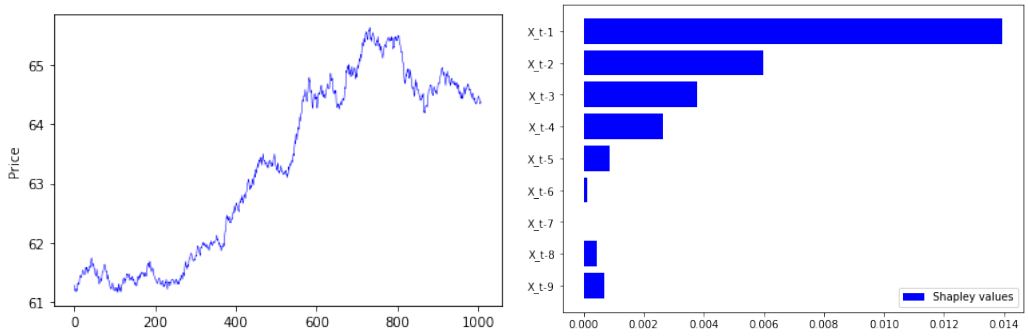

Figure 4.4: A droite, l'estimation selon (3) des valeurs de Shapley relatives à l'évolution du prix (à gauche) de 2020-01-07 13:12:30 à 2020-01-07 17:24:45

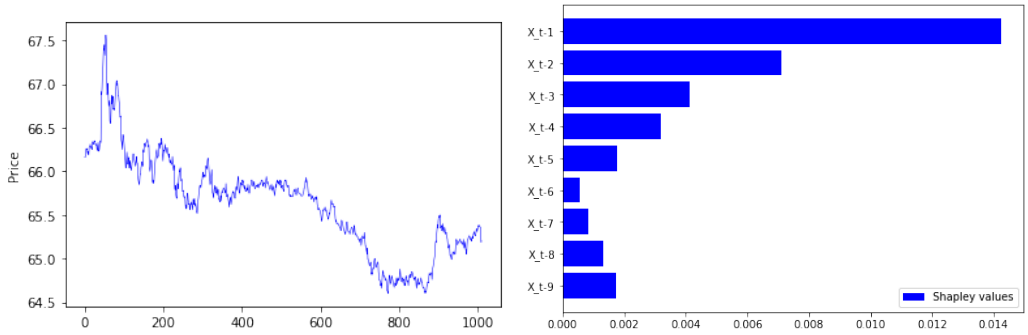

Figure 4.5: A droite, l'estimation selon (3) des valeurs de Shapley relatives à l'évolution du prix (représentés à gauche) de 2020-02-07 17:19:45 à 2020-03-07 13:01:45

On remarque le comportement global suivant: la valeur de Shapley relative à l'instant $t-1$ qui précède immédiatement l'instant t est très importante comparée aux autres valeurs qui décroissent au fur et à mesure que l'on s'intéresse à des instants plus lointains dans le passé. Cela revient à dire que plus en s'éloigne de l'instant t de prédiction, moins sont importantes pour la prédiction les valeurs. Ce comportement est intuitif: on a tendance à valoriser le passé proche. On remarque néanmoins qu'après une forte décroissance, les valeurs de Shapley ont tendance à croître lentement pour les instants les plus éloignés.

Remark 7 Le choix des paramètres Q et M fait précédemment est loin d'être optimal puisqu'on s'est limité à de petites valeurs insuffisantes pour assurer la convergence de l'estimateur (3.30) des indices de Shapley. On assiste à une perte importante de précision. Ce point est repris plus loin.
De plus le temps de calcul est grand. Cette durée représente une contrainte majeure dans ces applications. Prendre de grandes valeurs pour Q et M risque de rendre les calculs trop long, impossible à mener avec les moyens traditionnels. Il serait utile de recourir au calcul distribué et au calcul GPU.

A ce stade, il serait intéressant d'étudier l'influence de la taille de l'intervalle temporel de calcul sur les indices de Shapley. L'intervalle de longueur 435 ($\sim$ 108 minutes) est étudié.

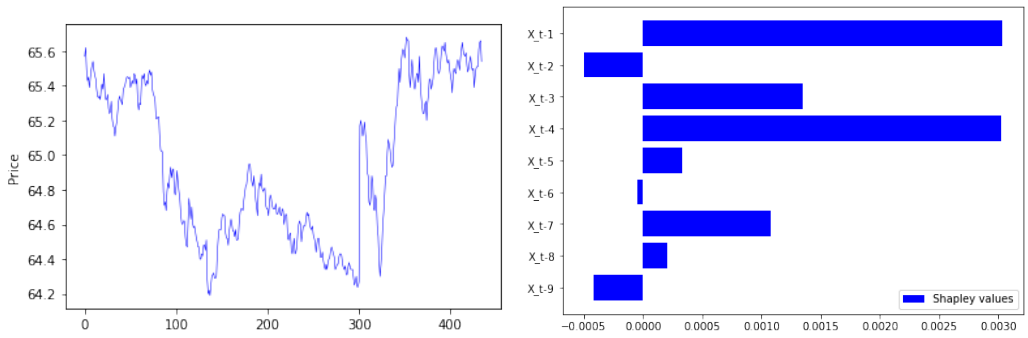

Figure 4.6: A droite, l'estimation selon (3) des valeurs de Shapley relatives à l'évolution du prix (à gauche) de 2020-01-07 16:15:00 à 2020-02-07 09:33:15

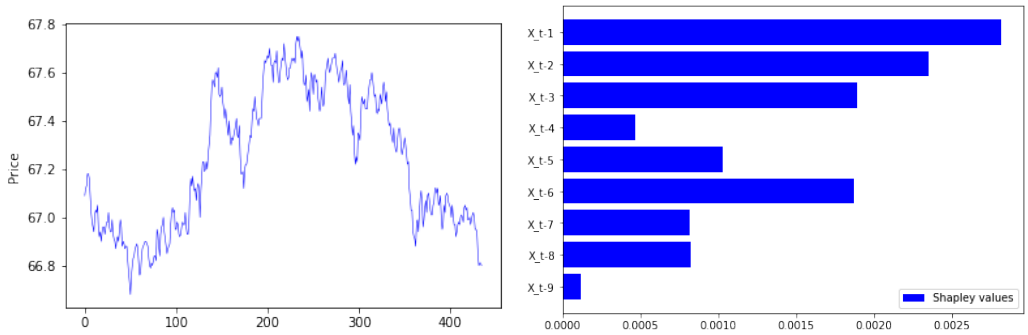

Figure 4.7: A droite, l'estimation selon (3) des valeurs de Shapley relatives à l'évolution du prix (à gauche) de 2020-02-07 17:19:45 à 2020-03-07 13:01:45

D'après les figures (4.6) et (4.7), les répartition des valeurs de Shapley sont très différentes. Les fenêtres temporelles plus petites induisent une réduction de la taille

de l'ensemble d'apprentissage rendant la prévision sensible aux différents instants considérés.

Passons à la formulation (3.31) de ces valeurs avec les mêmes paramètres que précédemment.

4.4.2 Estimation des valeurs de Shapley d'après (3.31)

On obtient les figures suivantes pour les mêmes situations que précédemment:

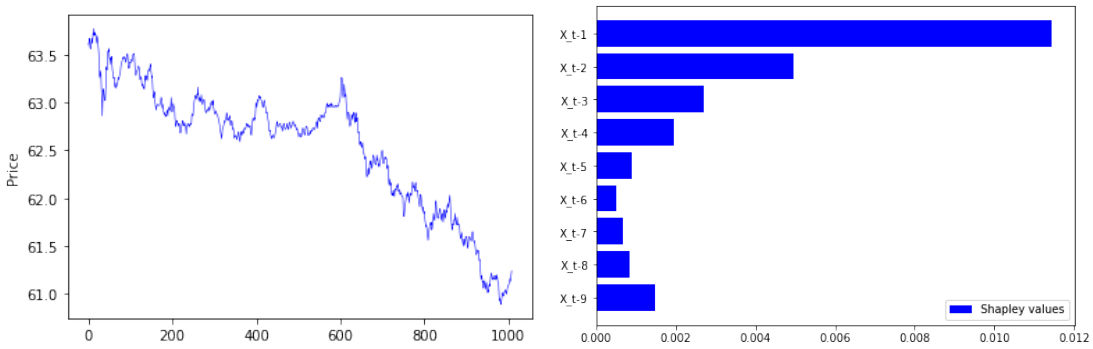

Figure 4.8: A droite, l'estimation selon (3.31) des valeurs de Shapley relatives à l'évolution du prix (à gauche) de 2020-01-07 09:00:00 à 2020-01-07 13:12:15

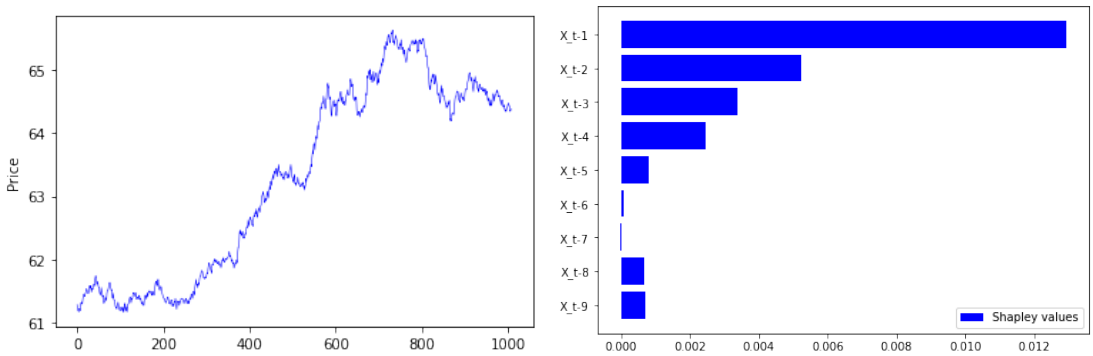

Figure 4.9: A droite, l'estimation selon (3.31) des valeurs de Shapley relatives à l'évolution du prix (à gauche) de 2020-01-07 13:12:30 à 2020-01-07 17:24:45

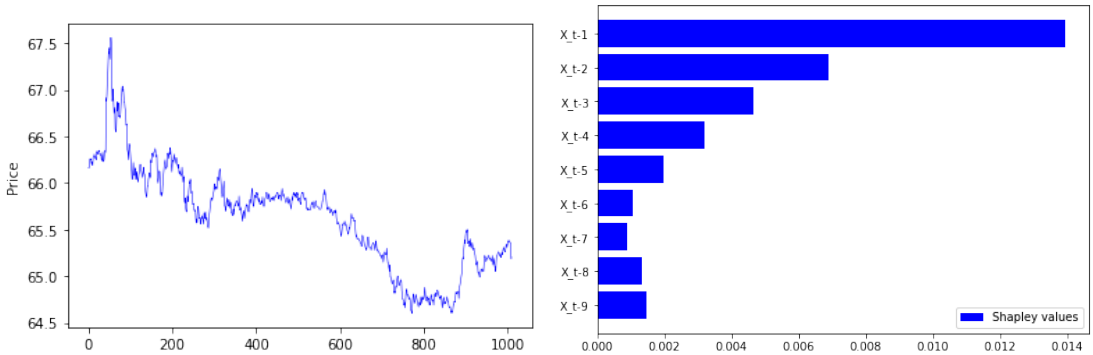

Figure 4.10: , A droite, l'estimation selon (3.31) des valeurs de Shapley relatives à l'évolution du prix (à gauche) de 2020-02-07 17:19:45 à 2020-03-07 13:01:45

En comparant les figures (4.8), (4.9) et (4.8) avec les figures (4.3), (4.4) et (4.5), on voit que les deux méthodes aboutissent à une même allure de la répartition des valeurs de Shapley mais pas exactement aux mêmes valeurs.

Dans la section qui va suivre, on analyse plus en détails les résultats de ces deux méthodes.

4.4.3 Analyse des résultats

On a pu voir dans la section précédente quelques résultats de simulation et d'estimation des indices de Shapley sur différentes fenêtres temporelles. On procède à une analyse plus fine de ces valeurs.
Tout d'abord on s'intéresse à l'intervalle de temps allant de 2020-01-07 09:00:00 à 2020-01-07 13:12:15, réduisant le temps de calcul. On garde le cadre d'étude présenté dans (4.3). On procède à l'estimation des indices de Shapley (dans notre cas 9 valeurs) un grand nombre de fois [2] ce qui permet de construire des distributions statistiques.

	X_t-9	X_t-8	X_t-7	X_t-6	X_t-5	X_t-4	X_t-3	X_t-2	X_t-1
1	0.001291	0.000716	0.000502	0.000514	0.001199	0.002267	0.003247	0.005518	0.012179
2	0.001520	0.000796	0.000598	0.000316	0.001024	0.002270	0.003424	0.005748	0.012025
3	0.001228	0.000763	0.000738	0.000568	0.001300	0.002304	0.003468	0.005786	0.012083
4	0.001155	0.000659	0.000862	0.000625	0.001293	0.002516	0.003389	0.005550	0.011855
5	0.001371	0.000713	0.000751	0.000686	0.001549	0.002708	0.003772	0.006379	0.012451
...	...	...	...	...	...	...	...	...	...
96	0.001366	0.000798	0.000634	0.000608	0.001181	0.002318	0.003504	0.005771	0.011691
97	0.001473	0.000795	0.000695	0.000512	0.001162	0.002433	0.003471	0.005977	0.012350
98	0.001253	0.000862	0.000659	0.000654	0.001026	0.002526	0.003562	0.006118	0.012148
99	0.001234	0.000546	0.000488	0.000584	0.000951	0.002108	0.003131	0.005316	0.011422
100	0.001345	0.000838	0.000510	0.000714	0.001193	0.002386	0.003639	0.006108	0.012558

100 rows × 9 columns

Figure 4.11: Exemple de tableau des estimations des valeurs de Shapley

[2] Dans cette application on se limite à 100 simulations (∼12 heures de calcul)

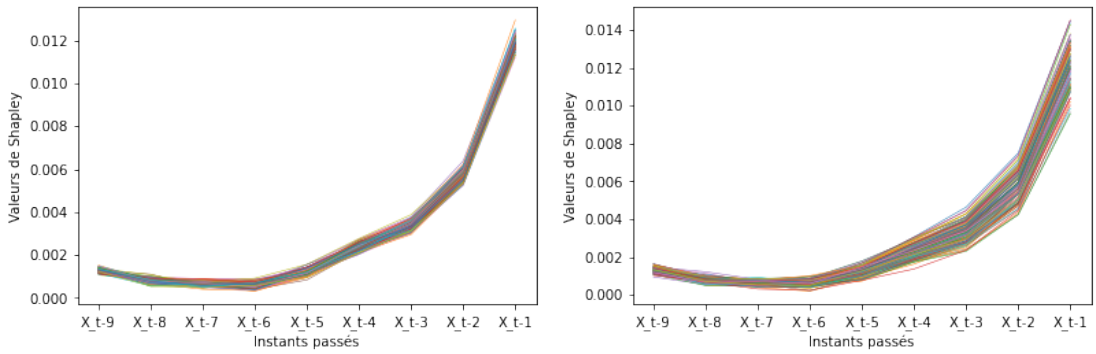

Figure 4.12: A gauche, les 100 valeurs de Shapley estimées selon (3) et à droite selon (3.31) pour chaque instant $X_{t-p}, p \in (1, \ldots, 9)$

.

La figure (4.12) montre que les deux approches ont des résultats du même ordre de grandeur et le même comportement. Néanmoins l'estimateur des valeurs de Shapley par (3.31) a une variabilité et une incertitude beaucoup plus grande due au faite qu'on a deux sources d'aléa (les deux boucles de Monte Carlo). Voir aussi la figure (4.13)

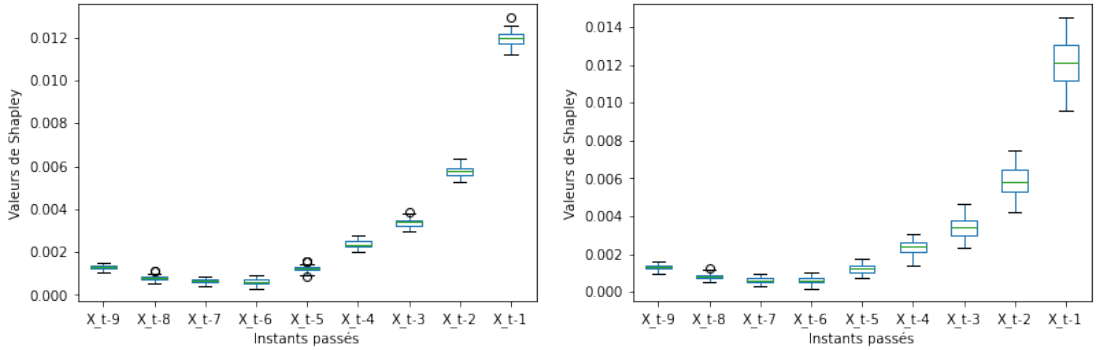

Figure 4.13: A gauche, les 100 valeurs de Shapley estimées selon (3) et à droite selon (3.31) pour chaque instant $X_{t-p}, p \in (1, \ldots, 9)$ présentées en Box Plot

.

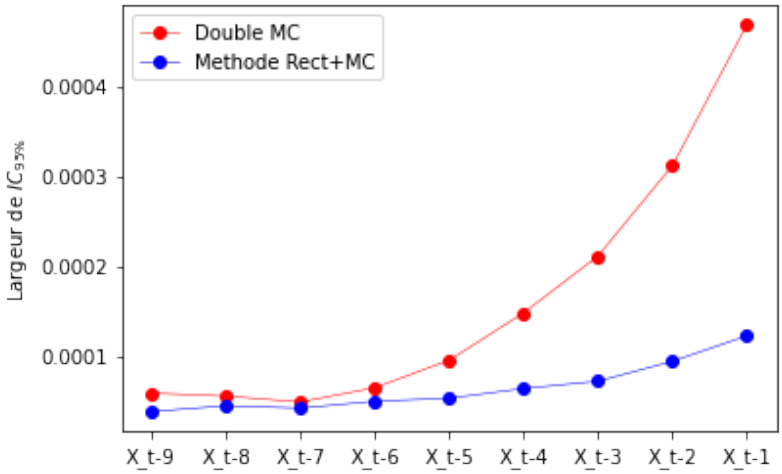

Figure 4.14

D'autre part, l'approche par double Monte Carlo (3.31) présente une perte de précision considérable. Pour remédier à ce problème, il est nécessaire de recourir aux méthodes de réduction de la variance notamment l'introduction d'une variable de contrôle paramétrée. Cette approche nécessite un temps de calcul très important [3] et n'a pas été mise en oeuvre. De plus la précision du calcul n'est pas l'object dans ce chapitre. On se concentre dans la suite sur l'approche basé sur (3).

Sensibilité des valeurs de Shapley à des perturbations des données initiales

Dans cette partie, on bruite de façon aléatoire la série temporelle; On étudie l'effet du bruitage sur la sensibilité des indices de Shapley à des perturbations des séries temporelles.
La perturbation est multiplicative ($X_{perturbe}(t) = X(t)(1 + \epsilon(t))$) où $\epsilon(t)$ est une suite de variables aléatoires identiquement distribuées de loi $Uniform([-10^{-4} : 10^{-4}])$.

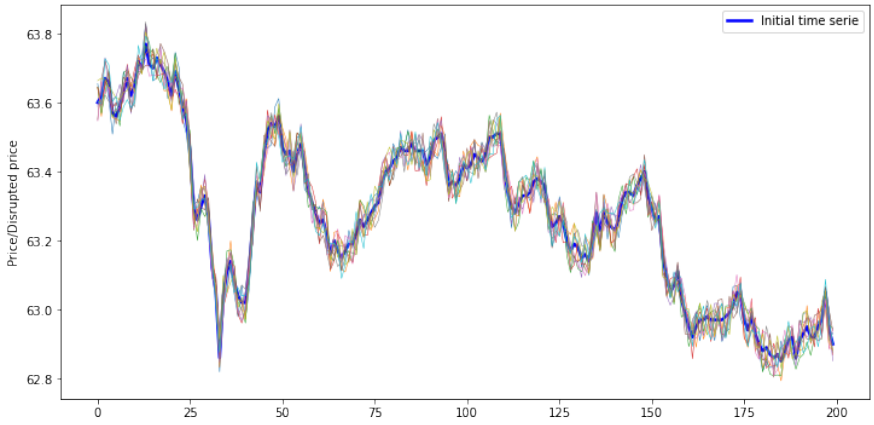

Figure 4.15: Perturbation de la série temporelle initiale selon ϵ

Comme précédemment, on estime les valeurs de Shapley après chaque perturbation des données initiales. Le calcul est répété un grand nombre de fois. [4]

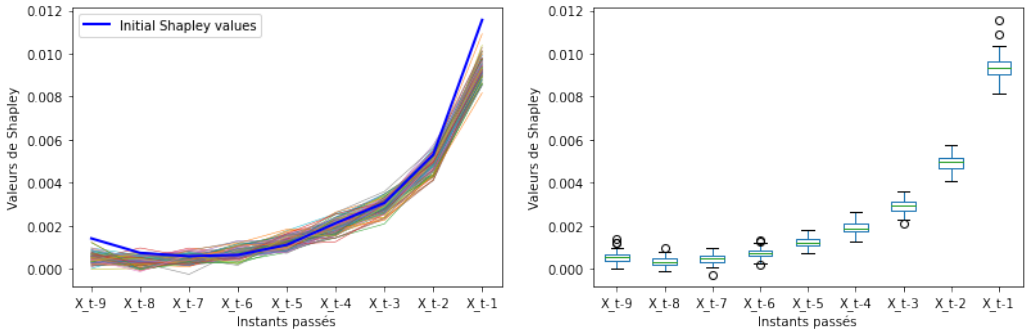

Figure 4.16

.

[3] L'introduction d'une variable de contrôle paramétrées nécessite l'estimation du paramètre de cette variable. Ce dernier est généralement exprimé en fonction d'un terme de covariance qu'on va donc devoir estimer, en plus de l'estimation de la quantité d'intérêt.

[4] Là encore, on fait 100 simulations.

Le pourcentage de variable des moyennes des valeurs calculées après perturbation est présentée dans la figure (4.17)

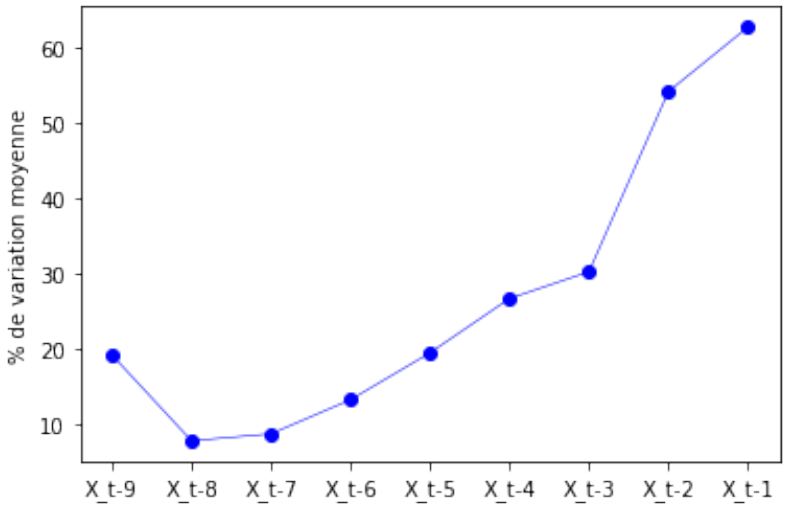

Figure 4.17

On remarque donc que pour des perturbations relativement petites de l'ordre de 10^{-4}, les variations des valeurs de Shapley peut atteindre 65% de la valeur initiale. On conclut à une forte sensibilité des indices aux (perturbations) des données initiales. Cette propriété présente des avantages et des inconvénients. La forte sensibilité permet d'affirmer que les Shapley caractérisent bien la série temporelle ainsi que l'intervalle de temps étudié et que ces valeurs capturent le comportement fin de la série. L'ambiguïté vient de la remarque suivante: si deux individus observent le même phénomène au fil du temps suivant des niveaux de précisions différents, les valeurs de Shapley seront différentes influençant l'interprétation des prédictions futures et des liens entre les différents instants.

Exemple d'évolution temporelle des valeurs de Shapley

Dans la section précédente on s'est limité à l'intervalle de temps allant de 2020-01-07 09:00:00 à 2020-01-07 13:12:15. Il est intéressant de suivre l'évolution des Shapley et de de leur distribution en fonction du temps.

Les mêmes calculs sont effectués sur un intervalle de temps glissant. On fixe un intervalle d'une *heure* de données, le translatant avec un pas d'une *heure*.

La figure(4.18) représenté l'évolution des Shapley sur une durée d'un mois.

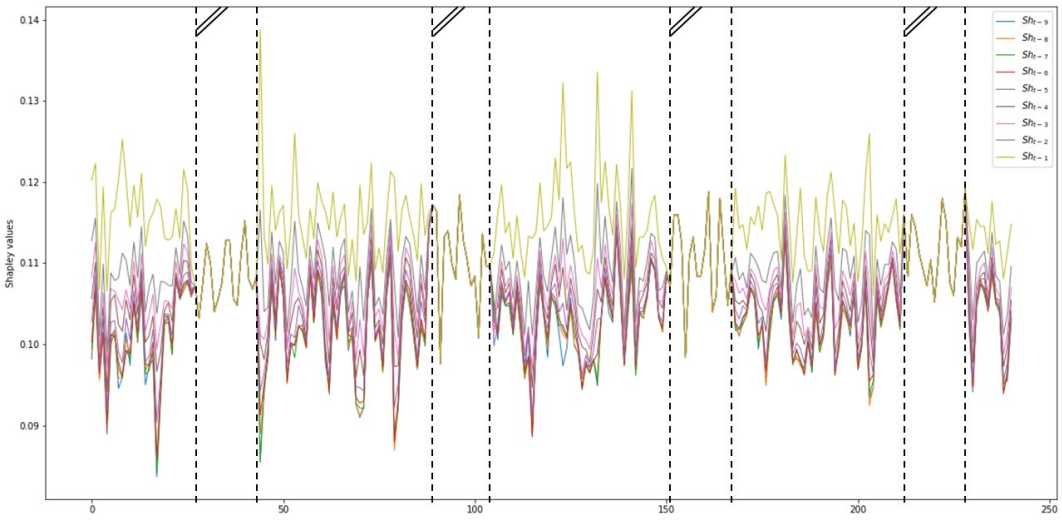

Figure 4.18: Evolution des différentes valeurs de Shapley sur une durée d'un mois. Les intervalles de temps marqués représentent les week-ends.

Détection des ruptures dans l'évolution temporelle des Shapley

La détection des ruptures identifie les changements dans une série temporelle. Les ruptures partitionnent les données en plusieurs segments, chacun représentant une période du temps au cours de laquelle les paramètres gouvernant le processus restent stables. Parmi ces paramètres on trouve notamment : la tendance, la moyenne et/ou la variance [Aminikhanghahi and Cook, 2017].

On se concentre sur les méthodes dites $Offline$ notamment l'algorithme $PELT$ (Pruned Exact Linear Time) et on s'intéresse aux ruptures de l'évolution temporelle des valeurs de Shapley.

La méthodologie générale de détection de ruptures $Offline$ consiste à se donner une séquence $(Y_1, \ldots, Y_N)$ d'une série temporelle $(Y_t)_{t\in\mathbb{N}}$ sur un intervalle de temps donné. On se donne aussi une suite de vecteurs aléatoires X_t à valeurs dans $\mathbb{R}^p$. On suppose aussi que pour tout $t \in (1, \ldots, N)$, Y_t est une fonction paramétrée de X_t:

$$Y_t = f_\theta(X_t, \epsilon_t), \tag{4.2}$$

où f_θ est une fonction déterministe de $\mathbb{R}^{p+1}$ dans $\mathbb{R}$ de paramètre θ ($\theta \in \Theta \subset \mathbb{R}^q$) et $(\epsilon_t)_{t\in\mathbb{N}}$ est un bruit blanc centré de variance σ^2 tel que les ϵ_t et les X_t sont des suites indépendantes.

L'objectif est donc de détecter un nombre K^* inconnu de ruptures dans $(Y_t)_{t\in\mathbb{N}}$. En d'autres termes, déterminer une suite d'instants $(\tau_1^*, \ldots, \tau_{K^*}^*)$ tels que $1 < \tau_1^* < \cdots < \tau_{K^*}^* < N$ et $\theta_j^* \neq \theta_{j+1}^*$, pour tout $j \in (0, 1, \ldots, \tau_{K^*}^* - 1)$. Dans les intervalles, le paramètre garde une valeur constante qui change d'un intervalle au suivant.

La méthode de détection de rupture $offline$ consiste donc à minimiser en $(K, (\tau_i), (\theta_i), \sigma^2)$

la fonction de coût suivante:

$$F(Y, X, (\theta_i), \sigma^2, K, (\tau_i)) = \sum_{i=0}^{K} \sum_{t=\tau_i+1}^{\tau_{i+1}} C(Y_t, X_t, \theta_i, \sigma^2) + \beta f(K), \tag{4.3}$$

où la fonction d'erreur C peut être définie de différentes manières (une erreur quadratique, une log-vraisemblance,...). Le terme $\beta f(K)$ est un terme de pénalité qui permet d'éviter un nombre excessif de ruptures. Dans cette section, on considère que $\beta f(K) = \beta K$ avec $\beta \geq 0$.
Le partitionnement optimal est assuré grâce à l'algorithme de la partition optimale (OP) (Jackson, 2005). Cet algorithme scanne de gauche à droite la série temporelle et calcule itérativement la valeur minimale $Z(u)$ de la fonction coût F sur l'intervalle de temps $[1, u], \forall u \leq N$. Le minimum $Z(u)$ se calcule par récurrence comme ci-dessous:

$$Z(u) = \underset{K_u, 1<\tau_1<\tau_2<x\cdots<\tau_{K_u}<u, \sigma^2}{min} \sum_{i=0}^{K_u} [\sum_{t=\tau_i+1}^{\tau_{i+1}} C(Y_t, X_t, \theta_i, \sigma^2) + \beta] \tag{4.4}$$

L'algorithme $PELT$ [16] est décrit ci-dessous dans (4)

Algorithm 4 Pruned Exact Linear Time [PELT]

Require: T : Time serie, $c(.)$: Cost function, β : penality value
Ensure: $L[T]$: Estimated set of change points
// Initialization
C ← (T+1)-long array with $C[0] = -\beta$
L ← (T+1)-long array with $L[0] = 0$
K ← set of time indexes initialized with $\{0\}$
// Change point detection
for t = 1,2,...,T **do**
$t^* \leftarrow \underset{s \in K}{argmin}(C[s] + c(T_{[s:t]}) + \beta)$
$C[t] \leftarrow C[s] + c(T_{[t^*:t]} + \beta)$
$L[t] \leftarrow L[t^*] \cup \{t^*\}$
$K \leftarrow \{s \in K, C[s] + c(T_{[s:t]}) \leq C[t]\} \cup \{t\}$
end for
return L[T]

Dans ce qui suit, on utilise la librairie **Ruptures** [17] sur Python. En fixant $\beta = 2$ on obtient les figures suivantes

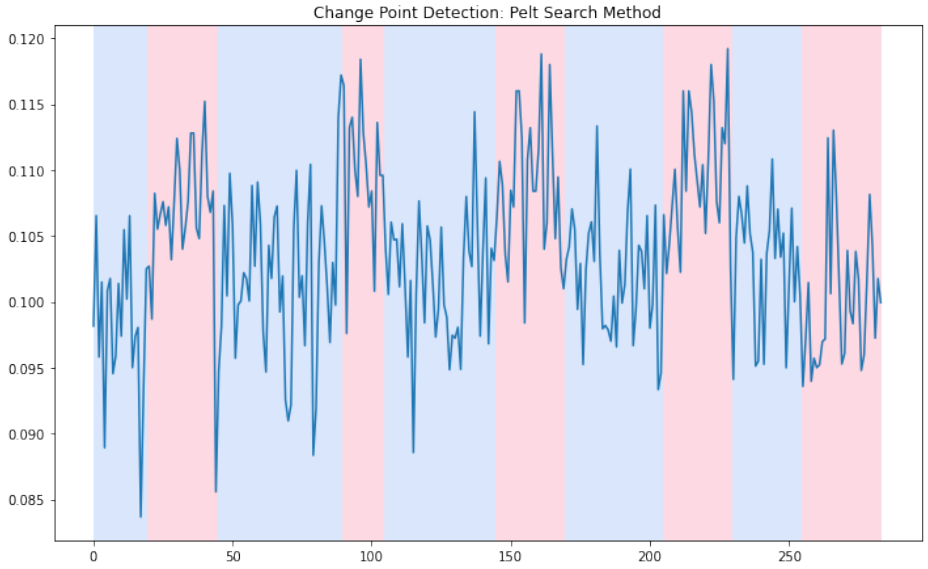

Figure 4.19: Méthode $PELT$ avec $\beta = 2$ appliquée à $Sh_{X_{t-9}}$

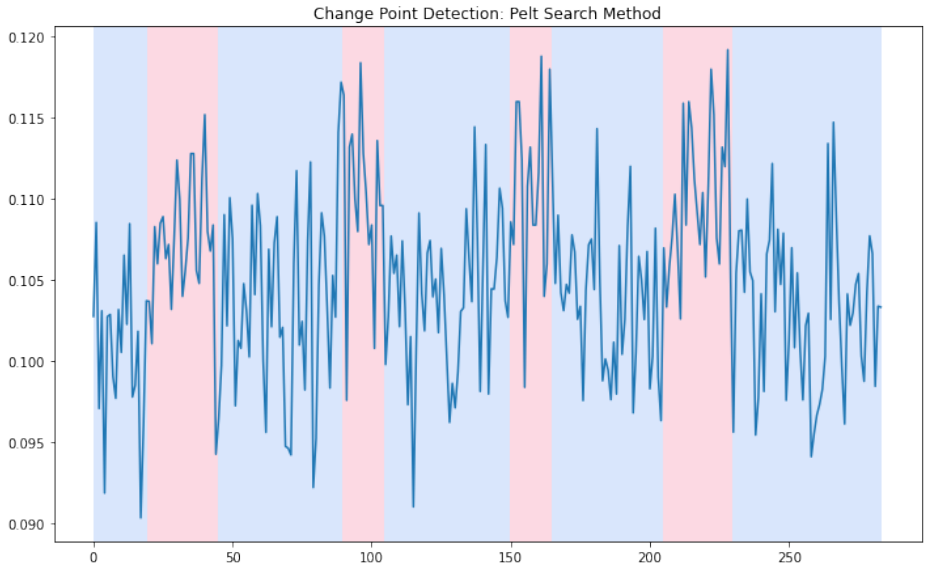

Figure 4.20: Méthode $PELT$ avec $\beta = 2$ appliquée à $Sh_{X_{t-5}}$

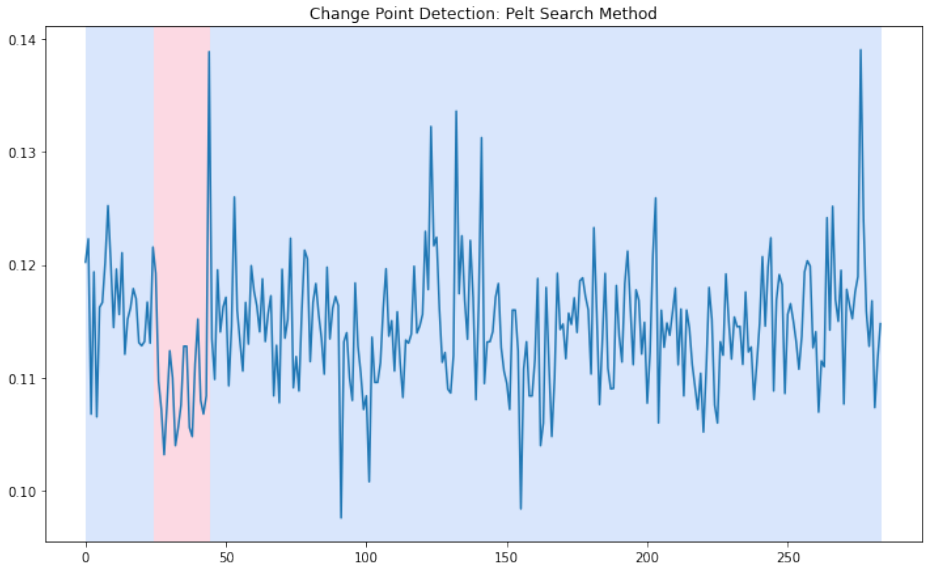

Figure 4.21: Méthode $PELT$ avec $\beta = 2$ appliquée à $Sh_{X_{t-1}}$

Dans les figures (4.19), (4.20),(4.21), la transition de couleur représente un point de rupture et donc le début d'une nouvelle tendance de la série.
Dans les figures (4.19), (4.20), les points de rupture ont lieu le **03/07/2020**, **06/07/2020**, **11/07/2020**, **13/07/2020**, **17/07/2020**, **20/07/2020**, **24/07/2020**, **27/07/2020**, **30/07/2020** et le **02/08/2020**. On en déduit que les bandes rouges caractérisent les week-end. Ainsi pour $\beta = 2$, on peut affirmer que c'est au niveau des week-ends que $Sh_{X_{t-9}}$ et $Sh_{X_{t-5}}$ changent brutalement de tendance et de comportement global. Cela peut s'expliquer par le fait que dans la base de donnée, le prix de l'action reste constant tout au long du week-end et est égal au dernier prix à la fermeture. Concernant $Sh_{X_{t-1}}$ dans la figure (4.21), les points de ruptures sont situés au niveau du **03/07/2020** et du **06/07/2020**: l'unique bande rouge concerne le premier week-end de la base donnée.
$Sh_{X_{t-1}}$ est beaucoup moins sensible aux week-ends (absence de ruptures) et la transition d'une semaine à une autre se fait de façon naturelle sans anomalie visible. En procédant à une recherche d'anomalie pour d'autre valeurs du paramètre de pénalisation β, on remarque que les points de ruptures ont rarement lieu au début (fermeture du marché financier) ou à la fin du week-end (ouverture du marché financier). De façon générale, plus on s'éloigne de l'instant de prédiction t, plus la valeur de Shapley d'un instant passé présente des points de ruptures lors des week-ends; en conséquence son importance pour la prédiction est sensible aux transitions d'une semaine à une autre.

Après avoir analyser les points de rupture de l'évolution des Shapley sur une durée d'un mois, il est intéressant d'analyser une durée plus courte d'une semaine. On s'intéresse à la semaine allant du **06/07/2020** au **10/07/2020** et on fixe $\beta = 0.1$.

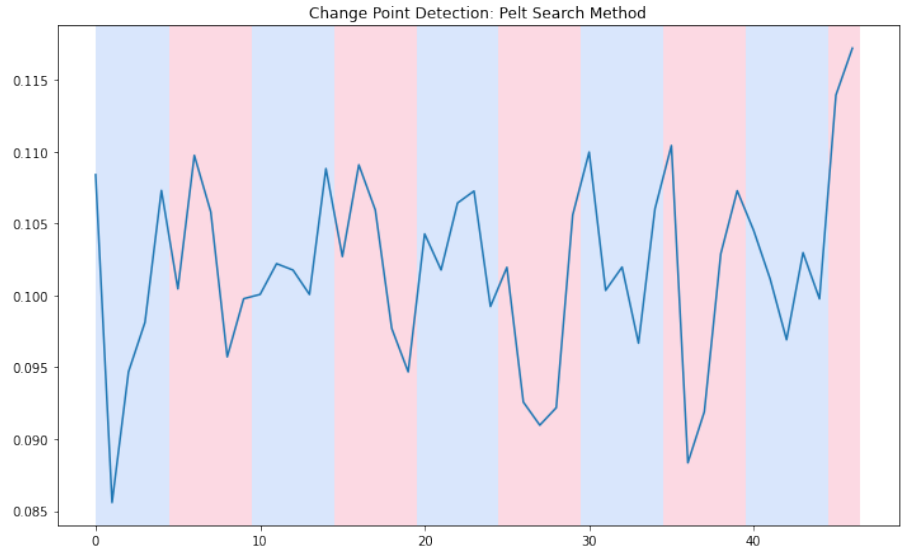

Figure 4.22: Méthode $PELT$ avec $\beta = 0.1$ appliquée à $Sh_{X_{t-1}}$

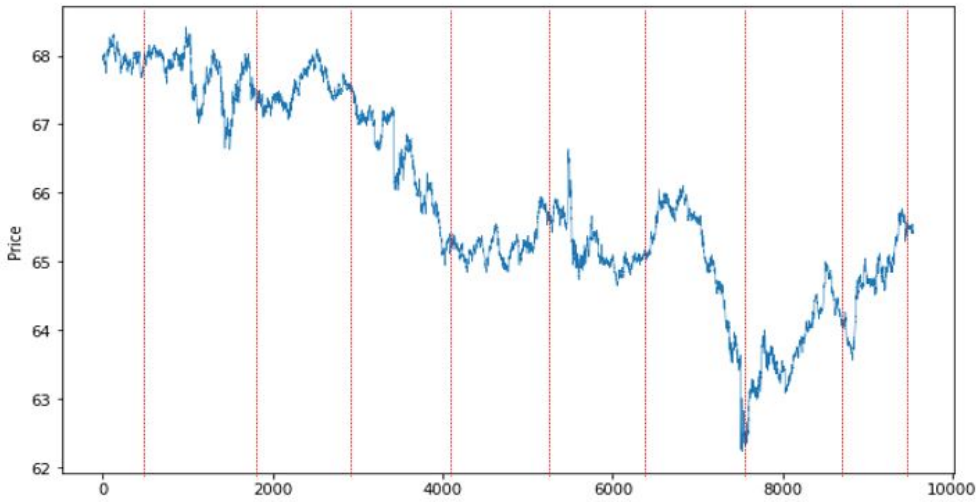

Figure 4.23: Prix de l'action du **06/07/2020** au **10/07/2020**. Les instants de rupture relatifs à la figure (4.22) sont représentés par les segments rouges.

En comparant les figures (4.22) et (4.24), on remarque que plusieurs points de rupture propres à l'évolution de la valeur de Shapley présentés dans (4.22) sont aussi détectés lors de l'analyse des points de rupture sur la série financière en question (4.24).

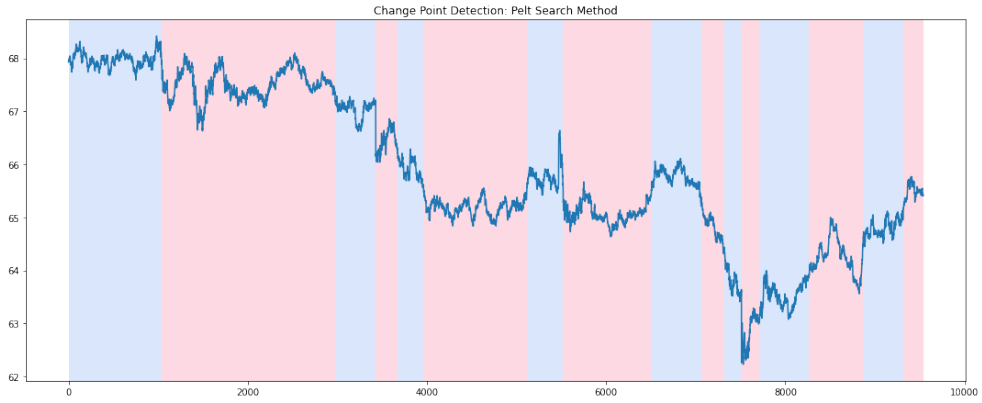

Figure 4.24: Méthode $PELT$ avec $\beta = 20$ appliquée à la série financière relative au prix de l'actif du **06/07/2020** au **10/07/2020**.

Remark 8 Dans cette dernière partie, l'analyse a porté sur l'évolution de $Sh_{X_{t-1}}$. L'analyse des autres instants (par exemple X_{t-2}) donne des résultats très semblables.

Chapter 5

Conclusion

5.1 Conclusion

Dans ce travail, nous nous sommes intéressés à la notion de causalité et fourni une approche de la causalité en s'inspirant des techniques d'analyse de sensibilité et en se basant sur les **valeurs de Shapley**. Nous avons aussi traité la question de l'estimation et du calcul des valeurs de Shapley, en particulier l'estimation par méthode de Monte Carlo appliquée à l'extension multilinéaire de ces valeurs.
L'approche mise en place a été appliquée à des séries financières relatives aux cotations de CAC40 du 01/07/2020 à 09 : 00 : 00 au 30/10/2020 à 17 : 30 : 00 avec un pas de 15 secondes. Nous avons pu donc analyser le comportement ainsi que la répartition des valeurs de Shapley sur un intervalle de temps donné en optant pour la régression kNN comme méthode de prédiction sur les séries.
L'évolution temporelle des valeurs de Shapley calculées a permis de mener une étude des points de rupture de ces valeurs grâce à l'algorithme $PELT$ et de les comparer avec les ruptures relatives aux séries étudiées.

5.2 Perspectives

Dans le dernier chapitre de ce rapport, on s'est intéressé aux relations "causales" au sein d'une série prise individuellement. Or l'idée principale est de quantifier les relations de causalités entre des séries temporelles et de déterminer non seulement l'influence des valeurs passées d'une série sur les prédictions futures de cette dernière, mais aussi sur les prédictions futures d'autres séries (5.1).

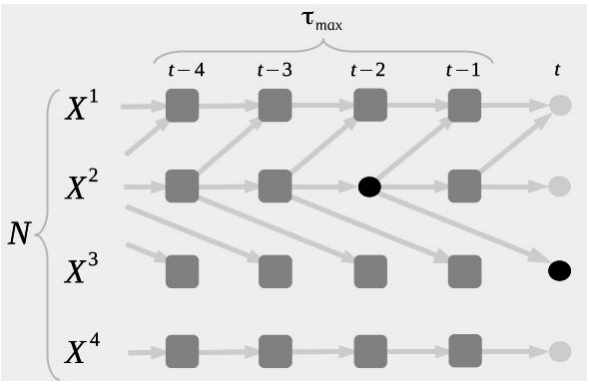

Figure 5.1

Une approche en cours de développement consiste à recourir à des LSTM comme modèles de prédictions. Plus précisément, en recourant à un Multivariate Multi-step Forecasting Stacked LSTM sequence to sequence Autoencoder (5.2) pour les prédictions futures sur les séries temporelles et en utilisant de la même manière les valeurs de Shapley, on peut quantifier l'influence de chaque instant passé sur les instants suivants.

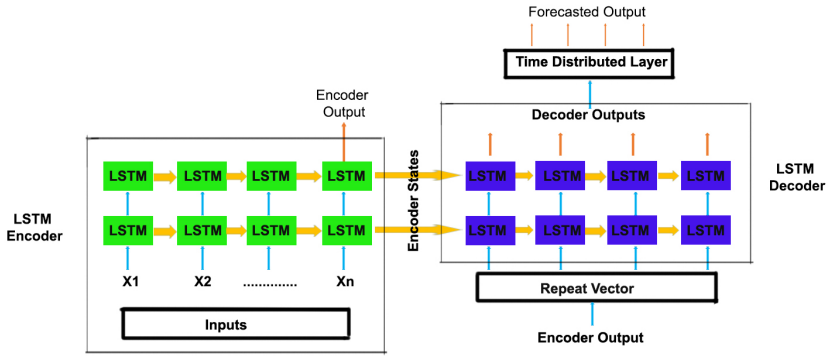

Figure 5.2

Or, le temps de calcul très important qu'implique cette approche fait qu'il est difficile d'avoir des résultats en entrainant le modèle de prédiction avec des moyens matériels peu sophistiqués.

Une autre piste à explorer consiste à analyser plus en détails l'évolution temporelle des valeurs de Shapley ainsi que leurs comportements en fonction de la structure de la série temporelle étudiée et donc de déterminer des tendances ou caractère saisonnier de ces valeurs qui aurons une interprétation et donc représenterons une source d'information.

Références

[1] Granger, C. W. J. Investigating Causal Relations by EconometricModels and Cross-spectral Methods . Econometrica 37, n 3 (1969):424-38. https://doi.org/10.2307/1912791.

[2] Shanmugam, Ramalingam. « Elements of Causal Inference: Foundations and Learning Algorithms ». Journal of Statistical Computation and Simulation 88, n 16 (2 novembre 2018): 3248-3248.

[3] Peters, Jonas, Dominik Janzing, et Bernhard Schölkopf. « Causal Inference on Time Series Using Restricted Structural Equation Models », s. d., 9.

[4] Amblard, Pierre-Olivier, et Olivier Michel. « The Relation between Granger Causality and Directed Information Theory: A Review ». Entropy 15, n 1 (28 décembre 2012): 113-43.

[5] Huang, Yuxiao, et Samantha Kleinberg. « Fast and Accurate Causal Inference from Time Series Data », s. d., 6.

[6] Marinazzo, D., M. Pellicoro, et S. Stramaglia. « Kernel-Granger Causality and the Analysis of Dynamical Networks ». Physical Review E 77, n 5 (27 mai 2008): 056215.

[7] Iooss, Bertrand, et Clementine Prieur. « Analyse de sensibilité avec entrées dépendantes: estimation par échantillonnage et par métamodèles des indices de Shapley », s. d., 6.

[8] Broto, Baptiste, François Bachoc, et Marine Depecker. « Variance Reduction for Estimation of Shapley Effects and Adaptation to Unknown Input Distribution ». SIAM/ASA Journal on Uncertainty Quantification 8, n 2 (janvier 2020): 693-716.

[9] Castellan, Gwenaëlle, Anthony Cousien, et Viet Chi Tran. « Non-Parametric Adaptive Estimation of Order 1 Sobol Indices in Stochastic Models, with an Application to Epidemiology ». Electronic Journal of Statistics 14, n 1 (2020): 50-81.

[10] Goffart, Jeanne, et Monika Woloszyn. « RBD-FAST: une méthode d'analyse de sensibilité rapide et rigoureuse pour la garantie de performance énergétique », 2018, 9.

[11] Benoumechiara, Nazih, et Kevin Elie-Dit-Cosaque. « Shapley Effects for Sensitivity Analysis with Dependent Inputs: Bootstrap and Kriging-Based Algorithms ». Édité par B. Bouchard, J.-F. Chassagneux, F. Delarue, E. Gobet, et J. Lelong. ESAIM: Proceedings and Surveys 65 (2019): 266-93.

[12] Jacques, Julien. « Pratique de l'analyse de sensibilité: comment évaluer l'impact des entrées aléatoires sur la sortie d'un modèle mathématique », s. d., 14.

[13] Grandjacques, Mathilde, Alexandre Janon, Benoit Delinchant, et Olivier Adrot. « Pick-Freeze Estimation of Projection on the Past Sensitivity Indices for Models with Dependent Causal Processes Inputs », s. d., 17.

[14] Da Veiga, Sébastien, Jean-Michel Loubes, et Maikol Solís. « Efficient Estimation of Conditional Covariance Matrices for Dimension Reduction ». Communications in Statistics - Theory and Methods 46, n 9 (3 mai 2017): 4403-24.

[15] A. Janon, T. Klein, A. Lagnoux-Renaudie, M. Nodet, and C. Prieur. Asymptotic normality and efficiency of two sobol index estimators. Disponible à http://hal.inria.fr/hal-

00665048, 2012.

[16] Cynthia FAURE. Détection ruptures et identification des causes ou des symptômes dans le fonctionnement des turboréacteurs durant les vols et les essais [Thèse]. Paris (250) : Université Paris 1; 2017.

[17] C. Truong, L. Oudre, N. Vayatis. Selective review of offline change point detection methods. Signal Processing, 167:107299, 2020.

[18] Okhrati, Ramin, et Aldo Lipani. « A Multilinear Sampling Algorithm to Estimate Shapley Values ». ArXiv:2010.12082 [Cs, Stat], 22 octobre 2020.

[19] Lütkepohl Helmut, « New introduction to multiple time series analysis », Springer, 2005.

[20] Campen, Tjeerd van, Herbert Hamers, Bart Husslage, et Roy Lindelauf. « A New Approximation Method for the Shapley Value Applied to the WTC 9/11 Terrorist Attack ». Social Network Analysis and Mining 8, n 1 (décembre 2018): 3. https://doi.org/10.1007/s13278-017-0480-z.

[21] Hausken, Kjell. « The Shapley Value of Coalitions to Other Coalitions ». Humanities and Social Sciences Communications 7, n 1 (décembre 2020): 104. https://doi.org/10.1057/s41599-020-00586-9.