



# Financial Data Analysis via Non-Uniform Fast Fourier Transform

01/06/2021

Rapport final

EQUIPE N° 10 :  
Imane EL BOUZID  
Anas HAIMOUD  
Richard LIN

RÉFÉRENT :  
Ioane MUNI TOKE

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>État de l'art</b>	<b>4</b>
<b>3</b>	<b>Étude de l'estimateur de Malliavin Mancino</b>	<b>6</b>
3.1	Préliminaires théoriques . . . . .	6
3.2	Principe . . . . .	7
3.3	Propriétés Théoriques . . . . .	9
3.3.1	Covariance Intégrée . . . . .	9
3.3.2	Covariance Instantanée . . . . .	10
<b>4</b>	<b>Calcul efficace des coefficients de Fourier</b>	<b>11</b>
4.1	Calcul proposé par Malliavin et Mancino . . . . .	11
4.2	Calcul par transformée de Fourier rapide . . . . .	11
4.2.1	Préliminaire théorique . . . . .	11
4.2.2	Calcul des coefficients de Fourier des log-prix . . . . .	13
4.2.3	Transformée de Fourier rapide avec zero padding . . . . .	13
4.3	Calcul par transformée de Fourier non uniforme . . . . .	14
4.3.1	Préliminaire théorique . . . . .	14
4.3.2	Paramétrage des noyaux utilisés . . . . .	16
<b>5</b>	<b>Performance de l'estimation de la covariance intégrée</b>	<b>18</b>
5.1	Estimateur de Hayashi Yoshida . . . . .	18
5.2	Simulation de données . . . . .	19
5.2.1	Mouvement brownien géométrique . . . . .	19
5.2.2	Génération de données asynchrones . . . . .	20
5.3	Distribution asymptotique de l'erreur . . . . .	20
5.3.1	Cas synchrone . . . . .	20
5.3.2	Cas asynchrone . . . . .	21
5.4	Précision de l'estimation de la corrélation . . . . .	22
5.4.1	Cas synchrone . . . . .	23
5.4.2	Cas asynchrone . . . . .	23
5.5	Influence des paramètres sur la précision de l'estimateur . . . . .	25
5.5.1	Effet de $N$ . . . . .	25
5.5.2	Effet de $n$ . . . . .	26
5.5.3	Effet de $\rho$ . . . . .	29
5.5.4	Effet de l'échantillonnage . . . . .	29
5.6	Optimisation de l'estimateur via transformée de Fourier non uniforme	31
5.6.1	Rapidité de l'estimation . . . . .	31
5.6.2	Précision de l'estimation . . . . .	33
5.6.3	Synthèse des résultats . . . . .	37

<b>6 Performance de l'estimation de la covariance instantanée</b>	<b>38</b>
6.1 Modèle de Heston généralisé . . . . .	38
6.2 Cas synchrone . . . . .	39
6.3 Cas asynchrone . . . . .	40
6.4 Synthèse . . . . .	43
<b>7 Application sur des données réelles</b>	<b>44</b>
7.1 Description des données . . . . .	44
7.2 Corrélation et effet Epps . . . . .	44
7.3 Etude des données sur le mois . . . . .	46
7.4 Etude des données quotidiennes . . . . .	48
<b>8 Conclusion et ouverture</b>	<b>54</b>
<b>9 Annexes</b>	<b>55</b>
9.1 Implémentation numérique du mouvement brownien géométrique . . . . .	55
9.2 Implémentation numérique de la NUFFT . . . . .	55
9.3 Implémentation numérique de l'estimateur de Hayashi Yoshida . . . . .	57
9.4 Figures supplémentaires . . . . .	58
9.4.1 Corrections des estimations de corrélations pour l'inversion de Dirichlet . . . . .	58
9.4.2 Matrices de corrélations pour $N = 0.075 \min(n_1, n_2)^{0.7}$ . . . . .	58
9.4.3 Matrices de corrélations pour l'inversion de Dirichlet . . . . .	59

---

# 1 Introduction

Avec l'essor du numérique, les transactions sur les marchés financiers se sont considérablement accélérées. L'échelle de temps caractéristique du trading haute fréquence se situe désormais à quelques microsecondes. Un calcul efficace de la covariance entre deux actifs requiert donc de construire des estimateurs précis de la covolatilité et d'optimiser leur temps de calcul, ce qui généralement impose de faire un compromis au niveau de la précision.

Au fil de ce projet, nous nous intéressons plus particulièrement à l'estimation de la covariance et de la corrélation de données financières asynchrones. Les estimateurs usuels reposent sur une hypothèse de synchronicité des temps d'observation des actifs et imposent généralement de faire des approximations ou des transformations de données pour les rendre synchrones, chose qui finit presque toujours par engendrer des biais d'estimation. Une autre faiblesse de ces estimateurs est généralement leur vulnérabilité aux bruits de micro-structure, qui sont encore plus présents dans le cas de données hautes fréquences. En outre, les techniques d'estimation de la covariance instantanée reposent presque toujours sur la dérivation numérique, qui est assez peu stable et coûteuse en termes de calcul.

Nous nous focalisons donc sur l'estimateur de Malliavin-Mancino qui permet d'apporter une solution à ces limitations grâce aux calculs des coefficients de Fourier de la matrice de covariance. Cela lui permet de mettre à profit la grande quantité de données disponibles sans nécessiter de traitement préalable pour synchroniser les données, tout en contrôlant sa précision grâce à une fréquence de coupure adéquate, contrairement à d'autres méthodes qui nécessiteraient de réduire le nombre de données disponibles [1]. Cet estimateur est alors plus robuste face aux bruits de micro-structure. Enfin, l'estimation de la variance instantanée se fait par intégration, qui est numériquement plus stable que la dérivation.

Si l'estimateur de Malliavin Mancino répond aux limitations précédemment posées, ses auteurs n'ont pas cherché durant la proposition de l'estimateur à optimiser son temps de calcul, ce qui ne permet donc pas de répondre à la problématique de la rapidité. L'article [7] a justement pour but d'optimiser le temps de calcul de l'estimateur de Malliavin Mancino via différentes méthodes dont la transformée de Fourier non uniforme. Nous nous baserons dans ce projet sur ce même article pour implémenter les méthodes citées, recréer et vérifier ses résultats avant de les appliquer à des données financières réelles.

---

## 2 État de l'art

L'estimateur de Malliavin Mancino, également connu sous le nom d'estimateur de Fourier, a été introduit pour la première fois par Paul Malliavin et Maria Elvira Mancino en 2002 et dans [15]. Alors que la plupart des estimateurs de la covariance et de la volatilité des prix d'actifs classiques sont construits sur l'hypothèse d'observations uniformément échantillonnées, l'estimateur de Malliavin et Mancino se fonde sur la construction des coefficients des séries de Fourier de la matrice de covariance à travers les coefficients de Fourier des log-prix. Cela permet de surmonter les difficultés que pose la mesure de la volatilité dans le cadre de données hautes fréquences. En effet, il ne nécessite pas d'interpolation pour transformer des séries temporelles non uniformément échantillonnées en séries uniformément échantillonnées, ce qui évite d'introduire des biais.

Malliavin et Mancino ont en effet pu établir un théorème dans [15] prouvant l'existence d'une relation générale entre les coefficients de Fourier de la covariance d'actifs financiers et les coefficients de Fourier de la variation du processus de prix. Cet estimateur offre l'avantage d'une convergence presque-sûre sous certaines hypothèses qui seront vues plus loin et une théorie de distribution limite de l'erreur étudiée dans [16]. L'estimateur de Malliavin Mancino permet également de calculer l'estimation instantanée de la matrice de covariance, estimation qui a été étudiée par [5] et par Cuchiero et Teichmann. Ces derniers ont travaillé sur l'amélioration de sa précision.

La version initiale présentée par [15] repose sur une approximation pour calculer les coefficients de Fourier qui résulte sur un calcul en complexité quadratique. Dans le cas où l'on dispose d'un grand nombre d'observations, ce calcul est difficile à mettre en pratique étant donné le temps d'exécution nécessaire. [7] ont proposé un calcul reposant sur la transformée de Fourier non uniforme permettant de passer à une complexité log-linéaire. Bien que ce soit généralement le cas, cette diminution remarque du temps de calcul ne s'est pas accompagnée d'une baisse de la précision.

Néanmoins, malgré les propriétés théoriques de convergences qui semblent excellentes de l'estimateur, en pratique ses performances dépendent fortement des fréquences de coupures choisies lors de l'estimation, et les conditions théoriques pour la convergence sont rarement vérifiées et ce particulièrement dans le cas asynchrone. [8] ont signalé cette dépendance et remarqué la détérioration des performances dans le cas asynchrone. Leur étude a également démontré que l'estimateur reste robuste et consistant pour l'estimation de la variance d'un même prix, quelque soient les conditions de cette estimation mais que l'asynchronicité engendre un biais d'estimation. Les auteurs ont proposé un facteur multiplicatif correctif dans certains cas précis, notamment quand les observations de deux prix asynchrones sont décalées d'un temps constant de la forme  $s\frac{\pi}{n}$ . [6] a également fait les mêmes observations sur l'estimateur en précisant que la performance dépend également de la nature de l'asynchronicité sous-jacente des données. Si l'estimateur reste robuste dans le cas de données manquantes, des observations à des temps distribués selon une loi exponentielle posent une plus grande difficulté.

L'estimateur de Malliavin Mancino a permis d'entreprendre plusieurs applica-

tions. En l'occurrence, [18] s'est appuyé sur la possibilité d'ajuster les fréquences de coupures choisies pour observer et appréhender les causes de l'effet Epps. Cet effet désigne une diminution de la valeur de la corrélation à mesure que le pas d'échantillonnage diminue. Cet effet a été découvert en 1975 par T. W. Epps dans [10]. Ses causes font aujourd'hui débat au niveau de la littérature. Certains estiment qu'il est le résultat de l'asynchronicité entre les données qui elle-même est la conséquence du caractère aléatoire des temps d'arrivées des transactions dans les marchés, d'autres soutiennent que c'est une conséquence des relations de lead-lag entre une paire donnée d'actifs, c-à-d, une paire composée d'un actif meneur et d'un actif suiveur. Des changements au niveau du meneur engendre des changement au niveau du suiveur. L'un est donc toujours statistiquement en avance sur l'autre. Enfin, la dernière cause potentielle soulevée pour expliquer l'effet Epps est la discrétisation des prix. A échelles de temps réduites, les valeurs des prix sont généralement concentrées autour de valeurs typiques et une modélisation continue peut ne pas être pertinente.

Interval	Pairs of Stocks					
	AMC-Chrysler	AMC-Ford	AMC-GM	Chrysler-Ford	Chrysler-GM	Ford-GM
10 minutes	.001	.009	-.009	-.014	.007	.055
20 minutes	.009	.018	.011	.017	.026	.118
40 minutes	.006	.012	.014	.041	.040	.197
One hour	-.043	.057	.064	.023	.065	.294
Two hours	.029	.060	.094	.112	.129	.383
Three hours	.031	.158	.111	.361	.518	.519
One day	-.067	.170	.078	.342	.442	.571
Two days	-.020	.223	.186	.336	.449	.572
Three days	-.098	.203	.100	.334	.542	.645

FIGURE 1 – Figure tirée de [10] illustrant l'effet Epps

Enfin, il existe également d'autres estimateurs de la covariance qui s'attaquent au problème de l'asynchronicité des données financières. Un des plus célèbres est celui de Hayashi-Yoshida, proposé par Takaki Hayashi et Nakahiro Yoshida en 2005 dans [14]. Cet estimateur, comme celui de Malliavin et Mancino, et contrairement aux estimateurs de covariance plus classiques, n'impose pas de synchroniser les données par interpolation, ce qui introduirait un biais dépendant à la fois du mode d'interpolation et des paramètres choisis. En revanche, cet estimateur ne fait pas intervenir la transformée de Fourier. Il est fondé sur une modification de l'estimateur de la covariance réalisée qui supprime les conditions de synchronicité. Son utilisation est valide sous des hypothèses très similaires à celles de Malliavin et Mancino. Toutefois, il ne présente pas l'avantage de pouvoir estimer la covariance à un instant  $t$  à l'image de la version instantanée de l'estimateur de Malliavin et Mancino et s'applique exclusivement à l'estimation de la covariance intégrée. De plus, il ne se prête pas à des applications telle que celle détaillée dans le paragraphe précédent, étant donné qu'il ne fait pas intervenir l'usage de fréquences de coupure pouvant simuler un échantillonnage pour analyser l'effet Epps.

---

### 3 Étude de l'estimateur de Malliavin Mancino

#### 3.1 Préliminaires théoriques

##### Séries de Fourier

On considère  $f : \mathbb{R} \rightarrow \mathbb{R}$  une fonction continue par morceaux et  $2\pi$ -périodique. On appelle coefficients de Fourier exponentiels de  $f$  la suite  $(c_n(f))_{n \in \mathbb{Z}}$  définie par

$$c_n(f) = \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-int} dt.$$

La série de Fourier de  $f$  est la série de fonctions :

$$S(t) = \sum_{n=-\infty}^{+\infty} c_n(f) e^{int}.$$

Les sommes partielles de cette série sont notées :

$$S_n(f)(t) = \sum_{k=-n}^n c_k(f) e^{ikt}.$$

##### Transformée de Fourier discrète

La transformée de Fourier discrète transforme une séquence de  $N$  nombres complexes  $x_n := x_0, x_1, \dots, x_{N-1}$  en une nouvelle séquence de nombres complexes,  $X_n := X_0, X_1, \dots, X_{N-1}$  définie par :

$$X(k) = \sum_{n=0}^{N-1} x_n e^{-j2\pi \frac{nk}{N}}.$$

En pratique, les  $N$  termes  $x_n$  peuvent être  $N$  échantillons d'un signal analogique échantillonné :  $x_n = x(nT_e)$ , et les  $X_k$  correspondent alors à une approximation (à un facteur multiplicatif  $T_e$  près) de la transformée de Fourier de ce signal aux  $N$  points de fréquence  $f_k = k \frac{f_e}{N}$ , où  $k$  varie entre 0 et  $N - 1$ , et donc  $f_k$  varie entre 0 et  $f_e$ .

Cette définition peut être généralisée dans le cas d'un espace multidimensionnel :

$$X_{\mathbf{k}} = \sum_{n=0}^{N-1} x_n e^{-i2\pi \mathbf{k} \cdot \frac{\mathbf{n}}{N}}.$$

Dans le cas où on dispose d'une fonction  $f$  périodique, la transformée de Fourier discrète permet d'obtenir une approximation des coefficients de Fourier de  $f$ .

### Théorème d'inversion de Dirichlet

**Théorème 1.** Soit  $f$  une fonction  $\mathcal{C}^1$  par morceaux et  $2\pi$  périodique. Alors, pour tout  $x \in \mathbb{R}$ ,  $S_n(f)(x)$  converge vers

$$\frac{f(x+) + f(x-)}{2},$$

où  $f(x+)$  (resp.  $f(x-)$ ) désigne la limite à droite (à gauche) de  $f$  en  $x$ .

### Théorème d'inversion de Fejér

Les moyennes de Césaro des séries de Fourier de  $f$  sont définies par :

$$\sigma_N(f) = \frac{S_0(f) + S_1(f) + \cdots + S_N(f)}{N+1}.$$

**Théorème 2.** Soit  $f$  une fonction continue et  $2\pi$  périodique. Alors les moyennes de Césaro de la série de Fourier de  $f$  convergent uniformément vers  $f$ .

**Remarque 1.** On peut montrer que :

$$\sigma_n(f)(t) = \sum_{k=-n}^{k=+n} \left(1 - \frac{|k|}{n}\right) c_k e^{ikt}.$$

## 3.2 Principe

Supposons que nous disposons d'un vecteur  $S(t) = (S_1(t), \dots, S_d(t))$  de prix d'actifs financiers observés sur un intervalle de temps  $[0, T]$ . Posons pour tout  $i \in \{1, \dots, d\}$  :  $p_i(t) = \log(S_i(t))$ , le log-prix de l'actif  $S_i$ .

Supposons que les  $p_i$  sont décris par l'équation suivante (dite équation stochastique d'Itô) :

$$dp_i(t) = \sigma_i(t)dW_i(t) + b_i(t)dt, \quad i = 1, \dots, d, \quad (1)$$

où :

- $W$  est un mouvement brownien sur l'espace probabilisé filtré  $(\Omega, \mathcal{F}, \mathbb{P})$  adapté à la filtration  $(\mathcal{F}_t)_{t \geq 0}$
- $b$  et  $\sigma$  sont des processus stochastiques adaptés à la filtration  $(\mathcal{F}_t)_{t \geq 0}$  tels que  $E \left[ \int_0^T b^2(t)dt \right] < \infty$  et  $E \left[ \int_0^T \sigma^4(t)dt \right] < \infty$ .

L'estimateur de Malliavin Mancino se base sur le calcul de la transformée de Fourier des dérivées des log-prix pour déduire les coefficients de Fourier de la matrice de covariance du vecteur des prix  $S(t)$  à partir des coefficients de Fourier des log-prix. La matrice de covariance est ensuite déduite grâce au théorèmes d'inversion de Dirichlet ou de Fejér.

En effet, en se plaçant sans perte de généralité dans l'intervalle  $[0, 2\pi]$ , notons :

$$\mathcal{F}(dp_j)(k) := \frac{1}{2\pi} \int_{[0, 2\pi]} \exp(-ikt) dp_j(t),$$

### 3.2 Principe

---

le coefficient de Fourier d'ordre  $k$  de la dérivée du log-prix  $dp_j$ . De même, on note :

$$\mathcal{F}(\Sigma^{l,j})(k) := \frac{1}{2\pi} \int_0^{2\pi} e^{-ikt} \Sigma^{l,j}(t) dt,$$

le coefficient de Fourier d'ordre  $k$  de la matrice de covariance  $\Sigma^{l,j}$  des actifs  $S_l$  et  $S_j$ . Le théorème suivant, établi par Malliavin et Mancino dans [15] permet d'établir le lien entre les coefficients de Fourier des log-prix et ceux de la matrice de covariance :

**Théorème 3.** *Si  $p(t)$  est un processus satisfaisant les hypothèses de (1), alors pour tout  $l, j = 1, \dots, d$  :*

$$\frac{1}{2\pi} \mathcal{F}(\Sigma^{l,j}) = \mathcal{F}(dp^l) * \mathcal{F}(dp^j),$$

où le produit de convolution utilisé est défini pour tout  $k$  par :

$$(\mathcal{F}(dp^l) * \mathcal{F}(dp^j))(k) := \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{|s| \leq N} \mathcal{F}(dp_l)(s) \mathcal{F}(dp_j)(k-s),$$

la convergence étant atteinte en probabilité.

En appliquant le théorème d'inversion de Dirichlet 1, on en déduit que la covariance instantanée vérifie l'équation :

$$\boxed{\Sigma^{l,j}(t) = \sum_{n=-\infty}^{n=+\infty} \left( \lim_{N \rightarrow \infty} \frac{2\pi}{2N+1} \sum_{s=-N}^{s=N} \mathcal{F}(dp_l)(s) \mathcal{F}(dp_j)(n-s) \exp(int) \right)}.$$

L'estimateur de la covariance instantanée associé à Dirichlet est donc :

$$\boxed{\Sigma_{M,N}^{l,j}(t) = \sum_{k=-M}^{k=M} \frac{2\pi}{2N+1} \sum_{s=-N}^{s=N} \mathcal{F}(dp_l)(s) \mathcal{F}(dp_j)(k-s) \exp(ikt).} \quad (2)$$

De même, en appliquant cette fois le théorème d'inversion de Fejér 2 et la remarque 1, on obtient :

$$\boxed{\Sigma^{l,j}(t) = \lim_{n \rightarrow \infty} \sum_{k=-n}^{k=n} \left( 1 - \frac{|k|}{n} \right) \lim_{N \rightarrow \infty} \sum_{s=-N}^{s=N} \frac{2\pi}{N+1} \mathcal{F}(dp_l)(s) \mathcal{F}(dp_j)(k-s) \exp(ikt).}$$

L'estimateur de la covariance instantanée associé à Fejér est donc :

$$\boxed{\Sigma_{M,N}^{l,j}(t) = \sum_{k=-M}^{k=M} \left( 1 - \frac{|k|}{M} \right) \frac{2\pi}{N+1} \sum_{s=-N}^{s=N} \mathcal{F}(dp_l)(s) \mathcal{F}(dp_j)(k-s) \exp(ikt).} \quad (3)$$

On peut également déduire des estimateurs de la **covariance intégrée** s'appuyant sur les coefficients de Fourier :

$$\text{Dirichlet : } \Sigma_{M,N}^{l,j} = \frac{(2\pi)^2}{2N+1} \sum_{s=-M}^{s=M} \mathcal{F}(dp_l)(s) \mathcal{F}(dp_j)(-s). \quad (4)$$

$$\text{Fejér : } \Sigma_{M,N}^{l,j} = \frac{(2\pi)^2}{N+1} \sum_{s=-M}^{s=M} \left(1 - \frac{|s|}{M}\right) \mathcal{F}(dp_l)(s) \mathcal{F}(dp_j)(-s). \quad (5)$$

**Remarque 2.** L'inversion de Fejér permet de filtrer les hautes fréquences en donnant plus de poids aux basses fréquences. Cela rend cet estimateur plus stable en présence de bruit haute fréquence [7].

**Remarque 3.** Les équations (2) et (3) font intervenir le choix de deux entiers  $n$  et  $N$ .  $M$  détermine le nombre de modes utilisés dans la reconstruction de la covariance instantanée tandis que  $N$  détermine le nombre de coefficients utilisés pour estimer les coefficients de Fourier de la covariance [5]. L'évaluation de la covariance instantanée demande donc l'évaluation des coefficients de Fourier des dérivées des log-prix d'ordre  $-(N+M)$  à  $N+M$ , ce qui peut rapidement devenir coûteux en terme de calcul [5].

**Remarque 4.** Malliavin et Mancino ont ainsi fourni un estimateur non paramétrique de la matrice de covariance instantanée et intégrée d'un vecteur prix vérifiant les hypothèses formulées en (1) résistant à l'asynchronicité des données. Il s'agit maintenant de déterminer des méthodes de calcul efficaces des coefficients de Fourier des dérivées des log-prix pour pouvoir utiliser les équations (2), (3), (4) et (5). Malliavin et Mancino ont directement calculé ces coefficients de manière approchée avec une complexité quadratique en le nombre d'observations disponibles.

### 3.3 Propriétés Théoriques

Les théorèmes de cette partie permettent d'étudier les propriétés asymptotiques de l'estimateur de Malliavin Mancino au niveau de l'estimation de la covariance instantanée et intégrée.

#### 3.3.1 Covariance Intégrée

**Théorème 4.** *L'estimateur de Malliavin Mancino pour la volatilité intégrée est consistant.*

Dans la suite de cette partie, on s'intéresse au comportement de l'erreur asymptotique dans les cas univarié et multivarié.

**Cas univarié :**

**Théorème 5.** *Soit  $S_i$  un des actifs de l'univers dont on dispose de  $n_i$  observations. Pour tout  $M, N$  on note :  $\widehat{\sigma}_{n_i, M, N}^2 = \widehat{\Sigma}_{M, N}^{i,i}$ . On pose  $\rho(n_i) = \max_{0 \leq h \leq n-1} |t_{h+1}^i - t_h^i|$ . Si  $\lim_{n_i \rightarrow \infty} \rho(n_i) = 0$ , alors :*

$$\rho(n_i)^{-1/2} \left( \widehat{\sigma}_{n_i, M, N}^2 - \int_0^{2\pi} \sigma^2(t) dt \right) \xrightarrow{n_i, N, M \rightarrow \infty} \mathcal{N} \left( 0, 2 \int_0^{2\pi} \sigma^4(t) dt \right),$$

où la convergence se fait en loi.

Les preuves de ces deux théorèmes se trouvent dans [15].

**Théorème 6.** Avec les mêmes notations que celles du théorème précédent, si  $\frac{N}{n_i} \rightarrow 0$ , alors :

$$\rho(n_i)^{-1/2\gamma} \left( \widehat{\sigma}_{n_i, M, N}^2 - \int_0^{2\pi} \sigma^2(t) dt \right) \xrightarrow[n_i, N, M \rightarrow \infty]{} \mathcal{N} \left( 0, 2 \int_0^{2\pi} \sigma^4(t) dt \right),$$

où  $\gamma > 1$  est tel que  $N^\gamma = O(n_i)$ , la convergence se faisant en probabilité.

La preuve de ce théorème se trouve dans [8]. Ce théorème démontre que la possibilité de régler le rapport  $\frac{N}{n_i}$  est une caractéristique importante de l'estimateur, et permet d'influer sur sa précision. Nous développerons cet aspect par la suite dans le cas de données asynchrones [16].

**Cas multivarié :**

**Théorème 7.** Soient  $S_i, S_j$  des actifs de l'univers dont on dispose respectivement de  $n_i$  et  $n_j$  observations. On note  $n = \min(n_i, n_j)$  et  $\rho(n) = \max(\rho(n_i), \rho(n_j))$ . Si  $\mathbf{N}\rho(\mathbf{n}) \rightarrow \mathbf{0}$  quand  $\mathbf{N}, \mathbf{n} \rightarrow +\infty$ , alors :

$$\rho(n)^{-\frac{1}{2\gamma}} \left( \widehat{\Sigma}_{n, N}^{i,j} - \int_0^{2\pi} \Sigma^{i,j}(t) dt \right) \xrightarrow[n, N \rightarrow \infty]{} \mathcal{N} \left( 0, \int_0^{2\pi} \Sigma^{i,i}(t) \Sigma^{j,j}(t) + (\Sigma^{i,j}(t))^2 dt \right).$$

### 3.3.2 Covariance Instantanée

**Cas univarié :**

**Théorème 8.** En utilisant les mêmes notations que précédemment :

$$\lim_{n_i, N, M \rightarrow \infty} \sup_{t \in [0, 2\pi]} |\widehat{\sigma}_{n_i, N, M}^2(t) - \sigma^2(t)| = 0.$$

De plus, sous les conditions  $\frac{N}{n} \rightarrow \mathbf{0}$  et  $\frac{M}{n} \rightarrow \mathbf{0}$  quand  $\mathbf{n}, \mathbf{N}, \mathbf{M} \rightarrow \mathbf{0}$ , la convergence en loi suivante est vérifiée :

$$\sqrt{\frac{n}{M}} (\widehat{\sigma}_{n, N, M}^2(t) - \sigma^2(t)) \xrightarrow[n, N, M \rightarrow \infty]{} \mathcal{N} \left( 0, \frac{4}{3} \sigma^4(t) \right).$$

**Cas multivarié :**

**Théorème 9.** Encore une fois, sous les mêmes notations que la partie précédente, si  $\mathbf{N}\rho(\mathbf{n}) \rightarrow \mathbf{0}$  et  $\frac{M}{N} \rightarrow \mathbf{0}$  quand  $\mathbf{n}, \mathbf{N}, \mathbf{M} \rightarrow \mathbf{0}$ , alors :

$$\lim_{n, N, M \rightarrow \infty} \sup_{t \in [0, 2\pi]} \left| \widehat{\Sigma}_{n_1, n_2, N, M}^{1,2}(t) - \Sigma^{1,2}(t) \right| = 0.$$

Pour l'instant, il n'y a pas encore de résultat théorique général donnant la distribution asymptotique de l'erreur de cet estimateur [16].

Les théorèmes mentionnés dans cette partie peuvent être retrouvés dans [16].

---

## 4 Calcul efficace des coefficients de Fourier

Comme vu précédemment, l'estimateur de Malliavin Mancin s'appuie sur le calcul des coefficients de Fourier des log-prix pour en déduire les coefficients de Fourier de la matrice de covariance. La matrice de covariance est reconstruite à partir de ces coefficients. L'étape de calcul des coefficients de Fourier est la plus coûteuse, le calcul proposé par Malliavin et Mancino dans [15] a une complexité quadratique en le nombre d'observations. Cette partie a pour but de présenter une vue d'ensemble de leur méthode de calcul ainsi que d'explorer des techniques d'amélioration basées sur l'article [7].

### 4.1 Calcul proposé par Malliavin et Mancino

Dans [15], Malliavin et Mancino approximent les coefficients de Fourier des log-prix de la manière suivante :

On note par  $\{0 = t_0^j < t_1^j < \dots < t_{n_j}^j = 2\pi\}$  les temps d'observation du prix  $S_j$  et faisons l'approximation :

$$\exp(-ikt) \approx \exp(-ikt_l^j), \quad t \in [t_l^j, t_l^{j+1}[.$$

Alors :

$$\begin{aligned} \mathcal{F}(dp_j)(k) &= \frac{1}{2\pi} \sum_{l=0}^{n_j-1} \exp(-ikt_l) \int_{[t_l, t_{l+1}]} dp_j(t) \\ &= \frac{1}{2\pi} \sum_{l=0}^{n_j-1} e^{-ikt_l^j} \delta_l(p_j), \end{aligned}$$

où  $\delta_l(p_j) = p_j(t_{l+1}) - p_j(t_l)$  pour tout  $l \in \{0, \dots, n_j - 1\}$ .

Malliavin et Mancino se sont exclusivement intéressés au calcul de leur estimateur de cette façon et on démontré que leurs coefficients approchés convergent bien vers les coefficients réels quand  $n_j \rightarrow +\infty$ . La complexité de ce calcul est en  $O(n_j^2)$ . Cette complexité quadratique le rend peu efficace dans le cas de l'étude de grands échantillons d'observations.

Une variante de ce calcul, proposé par [7] est de l'utilisation sous formulation matricielle afin d'exploiter certaines bibliothèques optimisées pour la multiplication matricielle de langages tels que Python, Julia ou Matlab.

### 4.2 Calcul par transformée de Fourier rapide

#### 4.2.1 Préliminaire théorique

La Transformée de Fourier rapide est simplement le calcul de la Transformée de Fourier discrète à l'aide d'un algorithme plus efficace permettant de réduire le nombre d'opérations nécessaires et donc la complexité du calcul.

## 4.2 Calcul par transformée de Fourier rapide

---

Le calcul naïf de la transformée de Fourier discrète d'un échantillon de taille  $N$  nécessite d'effectuer :

$$\begin{cases} N^2 & \text{multiplications complexes ,} \\ N(N - 1) & \text{additions complexes .} \end{cases}$$

La complexité de l'algorithme naïf est  $O(N^2)$ . L'algorithme de calcul de la Transformée de Fourier Rapide permet de réduire cette complexité à un  $O(N \log(N))$ .

L'algorithme de calcul de FFT le plus connu est celui de **Cooley-Tukey**. Cet algorithme nécessite que la taille  $N$  de l'échantillon soit une puissance, usuellement de 4 ou de 8. L'algorithme repose sur un réarrangement des sommes permettant d'obtenir les coefficients de la TFD en deux parties. En base 2, nous obtenons une somme sur les indices pairs et une somme sur les indices impairs comme ce qui suit :

$$\begin{aligned} X(k) &= \sum_{n=0}^{N-1} x(n)e^{-j2\pi \frac{nk}{N}} \\ &= \sum_{i=0}^{N/2-1} x(2i)e^{-j2\pi \frac{2ik}{N}} + \sum_{i=0}^{N/2-1} x(2i+1)e^{-j2\pi \frac{2(i+1)k}{N}} \\ &= \sum_{i=0}^{N/2-1} x(2i)e^{-j2\pi \frac{ik}{N/2}} + e^{-j2\pi \frac{k}{N}} \sum_{i=0}^{N/2-1} x(2i+1)e^{-j2\pi \frac{ik}{N/2}} \\ &= \sum_{i=0}^{N/2-1} y(i)e^{-j2\pi \frac{ik}{N/2}} + w^k \sum_{i=0}^{N/2-1} z(i)e^{-j2\pi \frac{ik}{N/2}}. \end{aligned}$$

Les deux termes composant  $X_k$  se déduisent donc directement des TFD de taille  $\frac{N}{2}$  des signaux  $y(i) = x(2i)$  et  $z(i) = x(2i+1)$ . Ainsi, on a obtenu que :

$$\forall k \in \left[0, \frac{N}{2} - 1\right], \quad X(k) = Y(k) + w^k Z(k).$$

En outre, on peut monter que :

$$\forall k \in \left[\frac{N}{2}, N - 1\right], \quad X(k) = Y(k - N/2) + w^k Z(k - N/2).$$

On en déduit que :

$$\forall k \in \left[0, \frac{N}{2} - 1\right], \quad \begin{cases} X(k) = Y(k) + w^k Z(k) \\ X(k + N/2) = Y_k + w^{k+N/2} Z(k) = Y(k) - w^k Z(k). \end{cases}$$

La TFD de l'échantillon  $(x_n)$  peut donc être calculée grâce à la TFD de deux signaux de taille  $\frac{N}{2}$  ainsi que  $\frac{N}{2}$  multiplications et  $N$  additions. Ce même calcul peut être réitéré pour réduire à chaque fois de moitié le temps la taille de la TFD à calculer puis calculer directement les transformées de Fourier discrètes intermédiaires ce qui permet une réduction du temps de calcul total à  $O(N \log(N))$ .

### 4.2.2 Calcul des coefficients de Fourier des log-prix

Les coefficients de Fourier des log-prix sont obtenus par transformée de Fourier discrète en utilisant bien sûr la FFT. Si  $n$  est le nombre d'observations de l'actif étudié, la complexité du calcul est donc en  $O(n \log(n))$  ce qui représente une amélioration sensible par rapport au  $O(n^2)$  initial.

Toutefois, La FFT ne peut être utilisée que dans le cas d'échantillonnages uniformes (c-à-d, les observations du signal doivent être équidistantes), ce qui réduit le champ d'utilisation de cette technique.

### 4.2.3 Transformée de Fourier rapide avec zero padding

La Zero Padded FFT permet d'utiliser la FFT sur des échantillonnages non uniformes d'un signal. Supposons que l'on dispose d'observations d'un signal sur des instants  $0 = t_0 < t_1 < \dots < t_n$  non nécessairement équidistants. On note  $\Delta t$  la distance minimale entre deux points de cette grille.

Le Zero Padding consiste à constituer une nouvelle grille de points à partir de la grille initiale, cette fois-ci uniforme et de période d'échantillonnage  $\Delta t$ . Les observations initiales sont placées sur les temps d'observations les plus proches de la nouvelle grille tandis que le reste de la grille est complété par des 0. Un schéma explicatif de ce principe peut être retrouvé en Figure 4.2.3. La FFT est ensuite appliquée sur les observations de la grille finale.

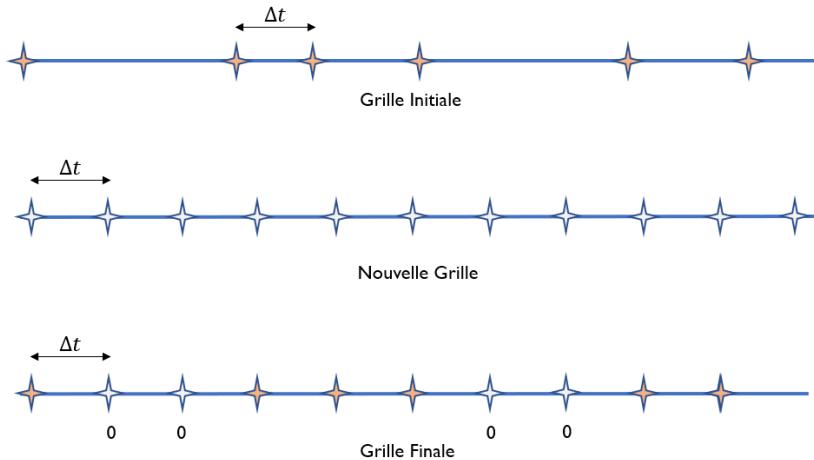


FIGURE 2 – Figure explicative du Zero Padding

**Remarque 5.** Il est important de noter que si la distance entre les points d'observation n'est pas toujours un multiple de  $\Delta t$ , il y a un décalage des observations entre la grille initiale et la grille finale. Si toutefois la distance entre tous les points d'observation est toujours un multiple (quelconque) de  $\Delta t$ , alors les points d'observations finaux et initiaux se superposent aux mêmes instants. Cette remarque montre qu'utiliser la Zero Padded FFT est légitime en cas de génération d'asynchronicité à partir de la suppression d'instants d'observation d'une grille initialement synchrone [7].

## 4.3 Calcul par transformée de Fourier non uniforme

Utiliser des techniques comme le zero-padding introduit des biais comme l'effet Epps. L'avantage de l'utilisation de la transformée de Fourier non uniforme est sa complexité en  $O(N \log(N) + M \log(\frac{1}{\epsilon}))$  où  $\epsilon$  est la précision souhaitée dans le calcul de la transformée de Fourier,  $N$  le nombre d'observations et  $M$  le nombre de modes de Fourier à calculer [9].

### 4.3.1 Préliminaire théorique

On peut toujours se ramener au cas d'une fonction  $f$   $2\pi$ -périodiques en jouant sur les échelles. On considère dans  $\mathbb{R}^d$  des points quelconques  $\mathbf{x}_j$  pour  $j \in \llbracket 0, N-1 \rrbracket$  et on note  $f_j := f(\mathbf{x}_j)$ . On pose  $K := K_{N_1} \times \cdots \times K_{N_d}$  où :

$$K_{N_i} := \begin{cases} \left\{-\frac{N_i}{2}, \dots, \frac{N_i}{2}-1\right\} & \text{si } N_i \text{ impair ,} \\ \left\{-\frac{N_i-1}{2}, \dots, \frac{N_i-1}{2}-1\right\} & \text{sinon.} \end{cases}$$

avec  $N_i$  le nombre de coefficients de Fourier à calculer de l'actif  $i$ . On distingue différents types de transformées de Fourier non uniforme (à  $2\pi$  près en fonction de la convention choisie) [13] :

- **Type I (dit *adjoint*) :**

$$F(\mathbf{k}) := \sum_{j=0}^{N-1} f_j e^{-i\mathbf{k} \cdot \mathbf{x}_j} \quad \text{pour } \mathbf{k} \in K.$$

La transformée de Fourier de  $f$  est calculée sur des fréquences régulières, les  $k \in K$  en des points disposés sur une grille quelconque de  $\mathbb{R}^d$ .

- **Type II (dit *forward*) :**

$$f_j := \sum_{\mathbf{k} \in K} F(\mathbf{k}) e^{i\mathbf{k} \cdot \mathbf{x}_j} \quad \text{pour } j \in \llbracket 0, N-1 \rrbracket.$$

La transformée de Fourier de  $F$  est calculée en des points disposés sur une grille régulière et les fréquences, ici les  $x_j$  sont quelconques.

- **Type III :**

Soit  $\mathbf{s}_k$  des points quelconques de  $\mathbb{R}^d$  :

$$F(k) = \sum_{j=1}^{N-1} f_j e^{-i\mathbf{s}_k \cdot \mathbf{x}_j}.$$

Il n'y a ici aucune restriction de régularité ni sur les fréquences où se calcule la transformée de Fourier ni les points d'échantillonnages de la fonction  $f$ .

On s'intéresse, ici, à la **transformée de Fourier non uniforme de type I**. Sans perte de généralité, on peut supposer que les  $x_j$  sont dans l'intervalle  $[0, 2\pi]$ . En dimension 1, il s'agit d'évaluer :

$$F(k) = \frac{1}{N} \sum_{j=0}^{N-1} f_j e^{-ikx_j} \quad \text{pour } k \in \{0, \dots, M-1\}, \quad M \in \mathbb{N}^*.$$

### 4.3 Calcul par transformée de Fourier non uniforme

---

Ce calcul coûte  $O(MN)$  opérations, ce qui est beaucoup plus que la complexité de la FFT. On va essayer de se ramener à un cas d'application de la FFT.

On peut remarquer que :

$$\frac{1}{2\pi} \sum_{n=0}^{N-1} f_n e^{-ikx_n} = \frac{1}{2\pi} \int \left( \sum_{n=0}^{N-1} f_n \delta(t - x_n) \right) e^{-ikt} dt := G(k).$$

C'est donc la transformée de Fourier évaluée en  $k$  de la fonction suivante :

$$g(x) := \sum_{j=0}^{N-1} f_j \delta(x - x_j),$$

où  $\delta$  est la distribution de Dirac.

On ne peut pas se contenter de faire un nouvel échantillonnage régulier en  $x$  puis d'appliquer une FFT à la fonction  $g$  en raison de la présence de la distribution de Dirac. C'est pourquoi on la régularise par convolution avec un noyau (*kernel* en anglais) judicieusement choisi, qu'on notera  $\varphi$ .

En notant  $p$  la période de ce noyau, on pose :

$$\tilde{\varphi}(x) := \sum_{r \in \mathbb{Z}} \varphi(x - rp)$$

La fonction  $\tilde{g} := g * \tilde{\varphi}$  est alors périodique,  $C^\infty$ . On peut maintenant créer une grille plus fine régulière à partir des observations initiales  $x_j$  et appliquer la FFT à la fonction  $\tilde{g}$  sur cette nouvelle grille. Soit  $M_r$  la taille de la grille ainsi définie. Le paramètre  $\sigma = \frac{M_r}{M}$  est le **paramètre de suréchantillonnage (over-sampling)** où  $M$  est le **nombre de modes de Fourier que l'on souhaite calculer**. En général, il est pris empiriquement égal à 2 [7, 9, 13].

On a alors pour tout  $\xi_\ell = \frac{2\pi}{M_r} \ell$  dans la nouvelle grille suréchantillonnée :

$$\tilde{g}(\xi_\ell) = \sum_{j=1}^{N-1} f_j \tilde{\varphi}(\xi_\ell - x_j), \quad \ell = 0, \dots, M_r - 1. \quad (6)$$

On peut donc maintenant évaluer la transformée de  $\tilde{g}$  en utilisant la FFT sur la grille régulière suréchantillonnée. On ne garde que les  $M$  premiers mode de la transformée de Fourier, étant donné que ce sont eux qui nous intéressent. Par ailleurs, sous l'hypothèse du choix judicieux d'un noyau et d'une période, on a l'approximation suivante :

$$\begin{aligned} \tilde{\Phi}(k) &= \frac{1}{p} \int_{-\frac{p}{2}}^{\frac{p}{2}} \sum_{r \in \mathbb{Z}} \varphi(x - rp) e^{-ik\frac{2\pi}{p}x} dx \\ &\approx \Phi(k). \end{aligned}$$

En effet, si le support des noyaux est suffisamment restreint et la périodisation  $p$  assez grande, le calcul précédent revient à prendre l'intégrale du terme  $r = 0$  dans la somme. Cela montre bien l'importance du choix des paramètres !

Enfin, avec les propriétés de la transformée de Fourier, on retrouve facilement les coefficients cherchés :  $F(k) = \Phi(k)^{-1}G(k)$ .

L'étape la plus coûteuse est le calcul des  $\tilde{g}(\xi_\ell)$ . En effet, les coûts aussi bien de stockage que de calculs augmentent exponentiellement en fonction de la dimension à cause du maillage et de l'interpolation. La solution pour réduire le temps de calcul est de **paramétriser le support du noyau**  $\varphi$  pour que  $\varphi$  soit non nulle sur un intervalle  $[-\alpha, \alpha]$  permettant de réduire le nombre de termes non nuls de la somme (6) ou encore de définir un nombre de points  $M_{sp}$  à prendre en compte dans le calcul de  $\tilde{f}(\xi_\ell)$ . On se restreint alors aux  $M_{sp}$   $x_j$  les plus proches de  $\xi_\ell$  de chaque côté. Le nombre de points pris dans chaque calcul est donc  $\omega = 2M_{sp} + 1$ . Cette approximation est valide étant donné que les noyaux choisis ont des pics centrés autour des  $x_j$  [13].

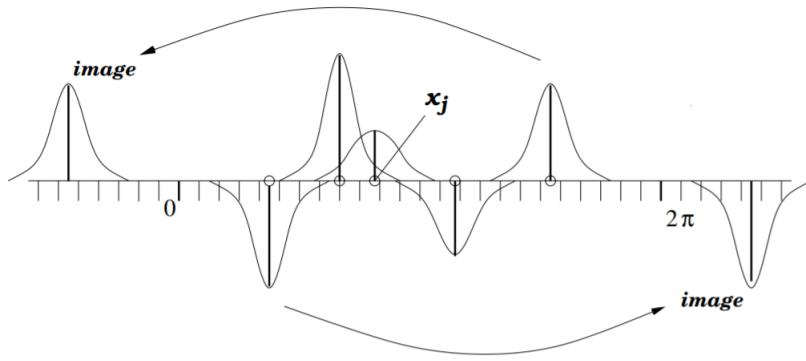


FIGURE 3 – Figure tirée de [13] illustrant la validité de l'approximation de la somme (6)

Empiriquement, l'erreur  $\epsilon$  décroît exponentiellement avec  $\omega$ . On peut donc faire l'approximation  $\omega = c|\log \epsilon|$  [2]. La complexité du calcul (6) est donc en  $O(M\omega) = O(M \log(\frac{1}{\epsilon}))$  tandis que la complexité du calcul des FFT de la fonction convoluée est en  $O(N \log(N))$ . On retrouve bien la complexité énoncée au début de cette partie.

Il est également important de noter que la précision dépendant du paramètre  $M_{sp}$ , il faudra configurer ce paramètre de façon à obtenir la précision souhaitée.

#### 4.3.2 Paramétrage des noyaux utilisés

Le choix des noyaux à utiliser pour le calcul de la NUFFT est un choix délicat pour assurer la convergence de l'algorithme. Ces noyaux ont fait l'objet de plusieurs études [2, 13, 9]. Comme dans l'article [7], nous utiliserons les noyaux suivants :

##### Noyau Gaussien :

Le noyau gaussien est défini par :

$$\varphi_G(x) = e^{-x^2/4\tau} \text{ pour } x \in \mathbb{R}, \quad \tau = \frac{1}{M^2} \frac{\pi}{\sigma(\sigma - 0.5)} M_{sp}.$$

### 4.3 Calcul par transformée de Fourier non uniforme

---

Il est périodique de période  $2\pi$ . Sa transformée de Fourier est donnée par :

$$\hat{\varphi}_G(k) = 2\sqrt{\tau\pi}e^{-k^2\tau}$$

Selon [7],  $M_{sp} = \left\lfloor \frac{-\ln(\epsilon)(\sigma-1/2)}{(\pi(\sigma-1))} + \frac{1}{2} \right\rfloor$  est un paramétrage empirique de  $M_{sp}$  permettant d'obtenir des résultats satisfaisants.

**Noyau de Kaiser-Bessel :**

Le noyau de Kaiser-Bessel est défini par :

$$\varphi_{KB}(x) = \frac{1}{\pi} \begin{cases} \frac{\sinh(b\sqrt{M_{sp}^2 - M_r^2 x^2})}{\sqrt{M_{sp}^2 - M_r^2 x^2}} & \text{si } |x| \leq \frac{M_{sp}}{M_r}, \\ \frac{\sin(b\sqrt{M_r^2 x^2 - M_{sp}^2})}{\sqrt{M_r^2 x^2 - M_{sp}^2}} & \text{sinon.} \end{cases}$$

Elle est 1-périodique. Sa transformée de Fourier est donnée par :

$$\hat{\varphi}_{KB}(k) = \frac{1}{M_r} I_0 \left( m \sqrt{b^2 - (2\pi k/M_r)^2} \right),$$

où  $b = \pi \left( 2 - \frac{1}{\sigma} \right)$  et  $I_0(x) = \sum_{k=0}^{\infty} \frac{\left( \frac{1}{4} x^2 \right)^k}{(k!)^2}$  pour tout  $x \in \mathbb{R}$ .

Encore une fois selon [7],  $M_{sp} = \left\lfloor \frac{1}{2} (\lceil \log_{10} \left( \frac{1}{\epsilon} \right) \rceil + 2) \right\rfloor$  donne des résultats empiriquement satisfaisants.

**Exponentielle semi-cercle :**

L'exponentielle de semi-cercle est définie par :

$$\phi_{ES}(x) = \begin{cases} e^{\beta\sqrt{1-x^2}-1} & \text{si } |x| \leq 1, \\ 0 & \text{sinon.} \end{cases}$$

La transformée de Fourier de l'exponentielle de semi-cercle n'est pas analytiquement connue, mais une approximation efficace de sa transformée peut être retrouvée en [2]. Dans notre cas, nous nous contenterons de l'estimer numériquement.

Empiriquement une valeur satisfaisante de  $\beta$  est  $\beta = 2.3\omega$ .

En pratique, pour paramétriser le support de l'exponentielle de semi-cercle sur  $[-\alpha, \alpha]$  avec  $\alpha = \frac{\pi\omega}{M_r}$ , on utilise le noyau translaté :

$$\varphi_{ES}(x) = \phi_{ES}\left(\frac{x}{\alpha}\right).$$

Un paramétrage satisfaisant de  $M_{sp}$  est  $M_{sp} = \left\lfloor \frac{1}{2} (\lceil \log_{10} \left( \frac{1}{\epsilon} \right) \rceil + 2) \right\rfloor + 2$ . On déduit  $\omega$  à partir de  $\omega = 2M_{sp} + 1$  [7]. Nous faisons également le choix de prendre  $N = M$  pour plus de simplicité d'implémentation.

---

## 5 Performance de l'estimation de la covariance intégrée

### 5.1 Estimateur de Hayashi Yoshida

Afin d'avoir une base de comparaison pour étudier les performances de l'estimateur de Malliavin Mancino, nous nous sommes intéressés à l'estimateur de Hayashi Yoshida. Introduit dans [14], l'un de ses principaux avantages est sa capacité à utiliser pleinement le caractère asynchrone des données financières, donc sans effectuer de rééchantillonnage. C'est pour cela qu'il est souvent comparé à l'estimateur de Malliavin-Mancino qui possède également cette propriété. Cette sous-partie a pour but de rapidement présenter son principe et ses propriétés.

Le raisonnement de l'article [14] se base sur une amélioration de l'**estimateur de la covariance réalisée**, qui s'utilise uniquement pour des données synchrones. Cet estimateur est défini dans le cas de deux log-prix  $P^1$  et  $P^2$  par :

$$V := \sum_{i=1}^{n-1} (P_{t_{i+1}}^1 - P_{t_i}^1)(P_{t_{i+1}}^2 - P_{t_i}^2).$$

et a la propriété de converger en probabilité vers la covariance intégrée, c-à-d,  $V \rightarrow \int_0^T \sigma_{12}(t) dt$ , lorsque le pas temporel tend vers 0.

Les données financières étant par nature asynchrones, les prix ne sont pas forcément définis aux mêmes instants, beaucoup de données seront manquantes. Il serait possible de d'abord choisir un intervalle commun, puis de réaliser une interpolation permettant de récupérer les données manquantes mais cela introduirait un biais. C'est pourquoi Hayashi et Yoshida ont proposé un estimateur de la covariance intégrée, valable dans le cas asynchrone, qui s'exprime sous la forme suivante :

$$V := \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} (P_{t_{i+1}^1}^1 - P_{t_i^1}^1)(P_{t_{j+1}^2}^2 - P_{t_j^2}^2) \mathbb{1}_{[t_i^1, t_{i+1}^1] \cap [t_j^2, t_{j+1}^2] \neq \emptyset}.$$

Dans leur article original [14], Hayashi et Yoshida montrent que sous certaines conditions, leur estimateur est consistant. Ils s'intéressent particulièrement à des mouvements browniens corrélés :  $dP_t^l = \mu_t^l dt + \sigma_t^l dW_t^l$ ,  $l = 1, 2$ , avec  $d < W^1, W^2 >_t = \rho_t dt$ ,  $\rho$  étant une fonction inconnue et déterministe,  $\mu^l$  une fonction progressivement mesurable et  $\sigma^l > 0$  une fonction déterministe et bornée. Les observations sont effectuées jusqu'à temps arbitraire  $T$ . Soit  $(I^i)_{i=1,2,\dots}$  et  $(J^i)_{i=1,2,\dots}$  des partitions de l'intervalle  $[0, T]$ , dont le cardinal combiné vaut  $n$ . On notera  $|.|$  le cardinal d'un intervalle.

**Théorème 10.** *Supposons que :*

- (a)  $(I^i)$  (resp.  $(J^i)$ ) est indépendante de  $P^1$  (resp.  $P^2$ )
- (b)  $\mathbb{E}[\max_i |I^i| \vee \max_j |J^j|] \xrightarrow{n \rightarrow +\infty} 0$ .

*Alors :*

1. Si  $\forall k \in \llbracket 1, 2 \rrbracket, \sup_{t \leq T} |\mu_t^k| \in L^4$ , alors l'estimateur converge vers la covariance intégrée en  $L^2$  lorsque  $n$  devient grand.
2. Si  $\forall k \in \llbracket 1, 2 \rrbracket, \sup_{t \leq T} |\mu_t^k| < \infty$  p.s., alors l'estimateur est consistant pour la covariance intégrée, soit converge en probabilité vers la covariance intégrée lorsque  $n$  devient grand.

**Remarque 6.** Si  $\forall t \in \llbracket 0, T \rrbracket, \mu_t^l = 0$  alors l'estimateur est non biaisé.

**Remarque 7.** Sous les mêmes hypothèses que celles du théorème précédent, il est possible de montrer que l'estimateur de la corrélation est consistant lorsque  $n \rightarrow \infty$ .

**Remarque 8.** L'estimateur de Hayashi-Yoshida ne vérifie pas l'inégalité de Cauchy-Schwarz. Par conséquent, il est théoriquement possible d'obtenir des estimations de corrélations supérieures à 1 dans le cas de séries singulièrement auto corrélées [3]. En pratique, les séries de rendements ont une faible auto corrélation ce qui exclut ce cas [4].

**Remarque 9.** En dépit des apparences, cet estimateur est en réalité proche de celui de Malliavin Mancino. Dans l'article [17], les auteurs montrent l'approximation suivante :

$$2\pi\mathcal{F}(\Sigma_{ij})(0) \approx \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} (P_{t_{k+1}}^i - P_{t_k}^i)(P_{t_{l+1}}^j - P_{t_l}^j) \left( \sum_{q=1}^N \frac{\cos(q(t_k - t_l))}{N} \right).$$

En comparant avec la formulation de l'estimateur de Hayashi-Yoshida, on constate que la forme est identique, mais que les pondérations diffèrent.

## 5.2 Simulation de données

### 5.2.1 Mouvement brownien géométrique

Nous conduisons dans cette partie des simulations de Monte Carlo en simulant des actifs de prix suivant un mouvement brownien géométrique.

Le processus  $B = (B_t)_{t \in [0, T]}$  est un mouvement brownien sur  $(\Omega, \mathcal{F}, \mathbb{P})$  si et seulement si :

1.  $B$  est issu de 0, c'est-à-dire que  $B_0 = 0$   $\mathbb{P}$ -presque sûrement
2.  $B$  est à trajectoires continues
3.  $B$  est à accroissement indépendants c'est-à-dire que pour tous  $0 = t_0 < t_1 < \dots < t_n$ , la famille de variables aléatoires  $(B_{t_n} - B_{t_{n-1}}, \dots, B_{t_1} - B_{t_0})$  est indépendante
4. pour tous  $0 \leq s < t \leq T, B_t - B_s$  suit une loi gaussienne centrée de variance  $t - s$ .

Un mouvement brownien géométrique de paramètres  $(\mu, \sigma)$  est un processus  $S = (S_t)_{t \in [0, T]}$  tel que pour tout  $t \in [0, T]$  :

$$S_t = S_0 e^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma B_t}.$$

avec  $S_0$  une variable aléatoire  $\mathcal{F}_0$ -mesurable,  $B$  un mouvement brownien,  $\mu \in \mathbb{R}$  et  $\sigma > 0$ .

Le mouvement brownien géométrique  $S$  est également défini par l'équation aux dérivées partielles :

$$\frac{dS(t)}{S(t)} = \mu dt + \sigma dB(t). \quad (7)$$

La particularité du mouvement brownien géométrique est d'être un processus stochastique à valeurs positives. Il peut donc être utilisé pour modéliser l'évolution de cours, de taux d'intérêts... Il est courant, en finance, de modéliser les prix par l'équation (1) en prenant  $S(t)$  comme prix de l'actif sous-jacent.  $\sigma$  est alors dite **la volatilité du prix de l'action** et  $\mu$  est le **taux de dérive de l'action**.

Nous avons utilisé le schéma de discrétisation de [12] pour implémenter numériquement ce processus.

### 5.2.2 Génération de données asynchrones

En premier lieu, afin de générer des données asynchrones, nous avons commencé par générer des échantillons de prix synchrones puis par supprimer aléatoirement de manière uniforme un pourcentage donné des observations de chaque vecteur prix. Bien que cette technique soit très simple, elle correspond plus à une représentation en données manquantes qu'à l'obtention de données réellement asynchrones. En outre, [6], a signalé que les performances de l'estimateur de Malliavin et Mancino dépendent également de la manière de générer ces données asynchrones.

C'est pourquoi, nous nous sommes également intéressés à la génération de données asynchrones en échantillonnant les temps d'observation avec une loi exponentielle. Cette expérience est d'autant plus intéressante vu que les temps d'attente entre instants d'observations sont généralement modélisés par une loi exponentielle. Concrètement, on commence par générer un échantillon de donnée synchrones aux dates d'observations  $\{t_1, \dots, t_n\}$  et on fixe un paramètre  $\lambda$ . Les temps d'observation qui sont retenus sont les plus proches de  $\sum_{i=0}^{i=k} X_i$ , où  $X_i$  est un tirage de la loi  $\mathcal{E}(\lambda)$ . On procède ainsi jusqu'à ce que  $\sum_{i=0}^{i=N} X_i > t_n$ .

## 5.3 Distribution asymptotique de l'erreur

### 5.3.1 Cas synchrone

Les théorèmes 6 et 7 nous fournissant la loi asymptotique de la distribution de l'erreur ainsi que sa variance asymptotique, nous avons conduit des simulations pour les vérifier. L'expérience a consisté à simuler deux prix de variance 0.2 et 0.1 avec une corrélation de 0.35 suivant un mouvement brownien géométrique comportant

### 5.3 Distribution asymptotique de l'erreur

chacun 512 observations avec un pas  $\rho(n) = \frac{1}{86400}$ . Cette distribution a été obtenue à partir de 1000 répétitions.

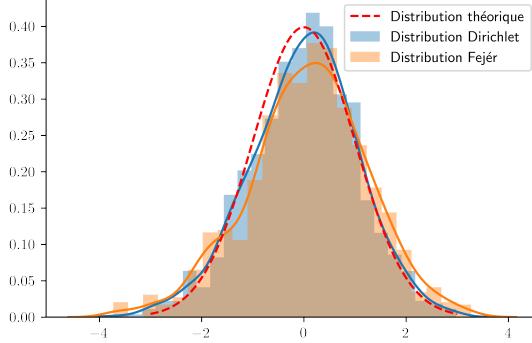


FIGURE 4 – Distribution de  
 $\rho(n)^{-1/2} \frac{\hat{\sigma}_{n,N}^2 - \int_0^{2\pi} \sigma^2(t) dt}{(2 \int_0^{2\pi} \hat{\sigma}^4)^{1/2}}$

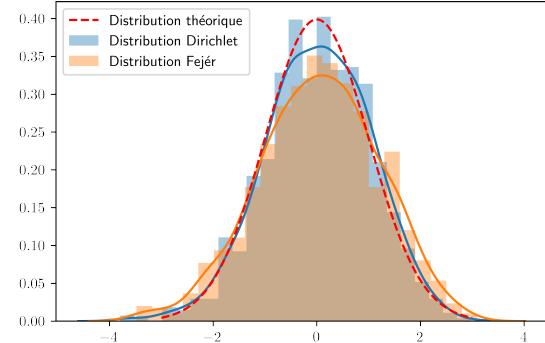


FIGURE 5 – Distribution de  
 $\rho(n)^{-1/2} \frac{(\hat{\Sigma}_{n,N}^{i,j} - \int_0^{2\pi} \Sigma^{i,j}(t) dt)}{(\int_0^{2\pi} \Sigma^{i,i}(t) \Sigma^{j,j}(t) + (\Sigma^{i,j}(t))^2 dt)^{1/2}}$

Les distributions ont bien l'allure d'une loi normale, toutefois les p-valeurs du test d'adéquation de Shapiro-Wilk sont de l'ordre de 0.014 pour l'erreur sur la covariance et de 0.023 pour la variance. Pour un niveau de risque  $\alpha = 0.01$ , on peut conclure que l'hypothèse de normalité de ces deux distributions est acceptée.

#### 5.3.2 Cas asynchrone

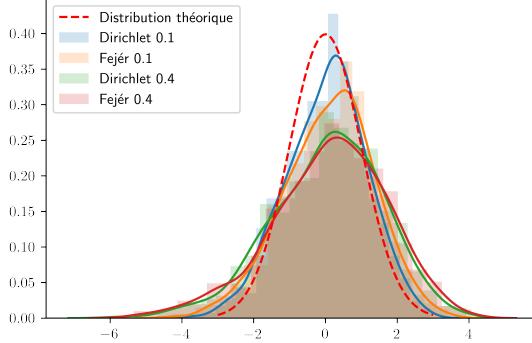


FIGURE 6 – Distribution de  
 $\rho(n)^{-1/2} \frac{\hat{\sigma}_{n,N}^2 - \int_0^{2\pi} \sigma^2(t) dt}{(2 \int_0^{2\pi} \hat{\sigma}^4)^{1/2}}$

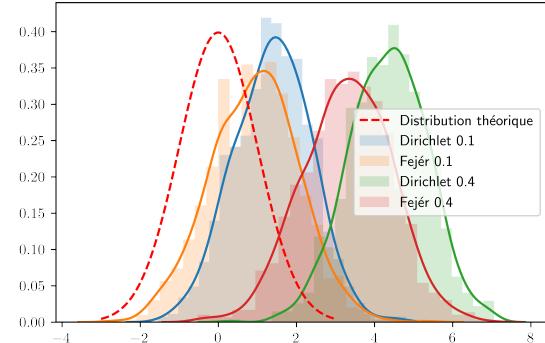


FIGURE 7 – Distribution de  
 $\rho(n)^{-1/2} \frac{(\hat{\Sigma}_{n,N}^{i,j} - \int_0^{2\pi} \Sigma^{i,j}(t) dt)}{(\int_0^{2\pi} \Sigma^{i,i}(t) \Sigma^{j,j}(t) + (\Sigma^{i,j}(t))^2 dt)^{1/2}}$

On remarque tout de suite que les résultats sont bien moins conformes au théorèmes de convergence donnés par Malliavin et Mancino. En particulier, dans le cas asynchrone, la distribution de l'erreur exhibe un biais non négligeable et est significativement différente de la distribution théorique attendue. Les résultats s'éloignent d'autant plus de ce cadre théorique que le pourcentage de données supprimées est grand. Ces résultats semblent assez étonnantes.

C'est dans [8] que nous avons trouvé une explication à cela. En effet, les théorèmes de convergence asymptotique donnés par Malliavin et Mancino supposent en général que  $\rho(n) \rightarrow 0$  et  $\frac{N}{n} \rightarrow 0$ . Ce qui en général, n'est pas vérifié. Malgré le pas relativement réduit pris ici ( $\rho(n) = dt = \frac{1}{86400}$ ), les résultats de la simulation divergent quand même. Il apparaît qu'il est également crucial de contrôler le rapport  $\frac{N}{n}$ , problématique à laquelle nous nous intéresserons un peu plus loin.

On peut aussi remarquer que l'estimation de la variance se comporte significativement mieux que l'estimation de la covariance entre deux prix différents et ce, malgré le fait que les conditions théoriques de convergence ne soient pas vérifiées. Dans [16] on peut trouver la preuve que l'espérance de l'erreur d'estimation, dans le cas où les coefficients de Fourier sont calculés selon la manière proposée par Malliavin et Mancino, prend la forme :

$$\mathbb{E} \left[ \widehat{\Sigma}_{n_1, n_2, N}^{1,2} - \int_0^{2\pi} \Sigma^{1,2}(t) dt \right] = \sum_{l=1}^{n_1-1} \sum_{r=1}^{n_2-1} (D_N(t_l^1 - t_r^2) - 1) \mathbb{E} \left[ \int_{I_l^1 \cap I_r^2} \Sigma^{1,2}(t) dt \right] \quad (8)$$

où  $D_N$  est le noyau de Dirichlet d'ordre  $N$  et  $I_l^1$  et  $I_r^2$  sont des intervalles vérifiant  $I_l^1 \cap I_r^2 = \emptyset$  si  $l \neq r$ . Ainsi dans le cas univarié, cette somme est toujours nulle étant donné que  $D_N(0) = 1$ . [8] présente une preuve démontrant que l'estimateur de Malliavin Mancino est consistant dans le cas univarié asynchrone quelque soit la manière de calculer les coefficients de Fourier.

## 5.4 Précision de l'estimation de la corrélation

L'estimation de la volatilité d'un même actif semblant assez robuste, on s'intéresse particulièrement à la précision de l'estimation de la corrélation entre deux actifs différents. Nous utilisons l'estimateur de Hayashi Yoshida pour benchmarker les résultats. Les expériences suivantes reposent sur une simulation de deux actifs suivant un mouvement brownien géométrique simulé de 10000 points avant sous-échantillonnage uniforme et 30000 points avant sous-échantillonnage exponentiel. La corrélation induite varie de  $-1$  à  $1$ . Les courbes sont obtenues par une moyenne sur 20 itérations. La fréquence de coupure des modes de Fourier est la fréquence de Nyquist.

### 5.4.1 Cas synchrone

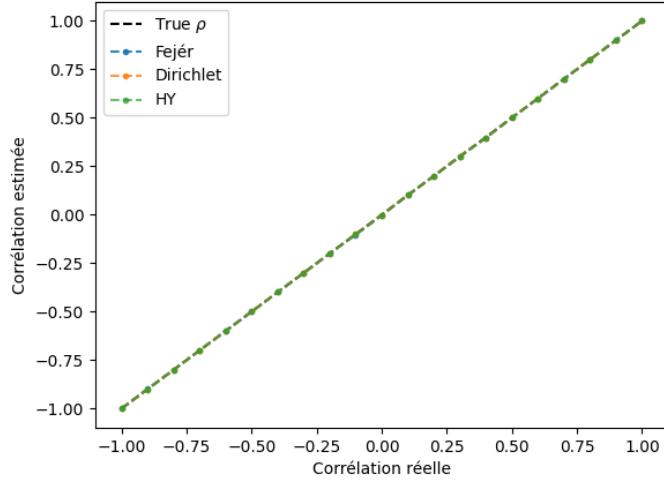


FIGURE 8 – Estimation de la corrélation - Cas synchrone

Il semble que les deux estimateurs arrivent quasi parfaitement à recouvrir la corrélation. L'estimateur de Malliavin Mancino semble marginalement plus performant que celui de Hayashi Yoshida.

### 5.4.2 Cas asynchrone

Dans un premier temps, on génère des données asynchrones par downsample.

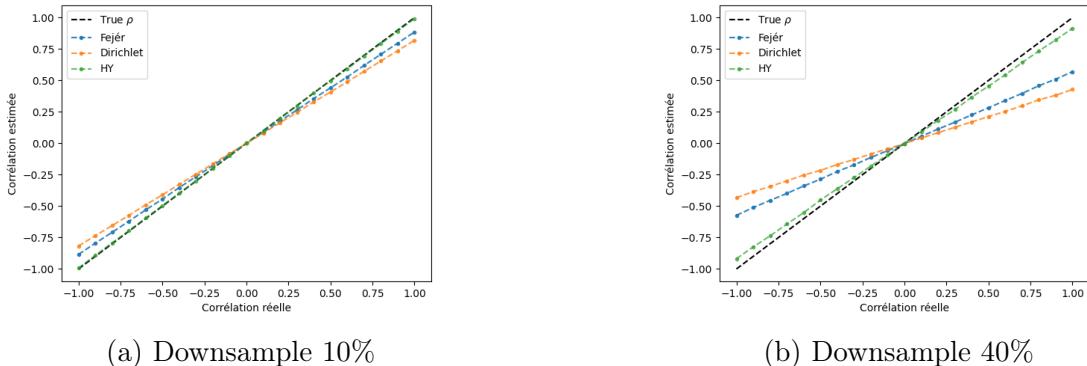


FIGURE 9 – Estimation de la corrélation sous différents pourcentages de sous-échantillonnage uniforme

On remarque que plus le pourcentage de données manquantes augmente, plus la qualité de l'estimation se dégrade. L'estimateur de Hayashi Yoshida semble être plus robuste que celui de Malliavin Mancino et produire une estimation plus précise. On

remarque aussi que l'erreur est plus forte au niveau des valeurs de corrélations les plus hautes en valeur absolue.

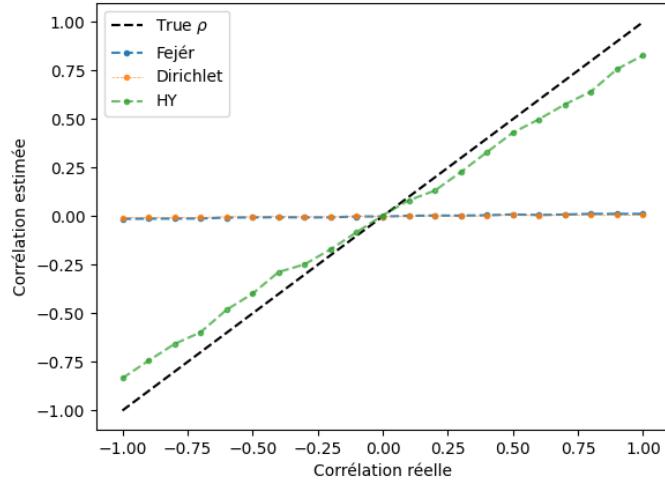


FIGURE 10 – Estimation de la corrélation sous échantillonage exponentiel  $\lambda_1 = 45$ ,  $\lambda_2 = 35$

L'erreur au niveau de l'estimateur de Hayashi Yoshida est tout de suite plus forte avec ce type d'échantillonnage. Toutefois, l'estimateur de Hayashi Yoshida arrive toujours à approcher raisonnablement la courbe réelle tandis que l'estimateur de Malliavin Mancino échoue complètement à reconstruire les corrélations différentes de 0. L'erreur de l'estimation est conséquente au point de rendre l'estimateur inutilisable dans ce cas. [8] attribue ce mauvais fonctionnement au non respect de la condition de convergence  $\frac{N}{n} \xrightarrow[n, N \rightarrow \infty]{} 0$ , avec  $n = \min(n_1, n_2)$ . Prenant compte cela, [16] a conduit des simulations de Monte Carlo pour déterminer pour différents  $n$ , le  $N = f(n)$  correspondant qui minimise l'erreur d'estimation dans le cas particulier de deux actifs dont les observations sont uniformes mais décalés de  $\frac{\pi}{n}$ . Empiriquement,  $N$  peut donc être représenté, dans ce cas précis d'asynchronicité, par la courbe  $f(n) = 0.85n^{3/4}$ . On réitère donc l'expérience précédente en imposant que  $N = 0.85n^{3/4}$  pour voir à quelle mesure cela peut améliorer le résultat.

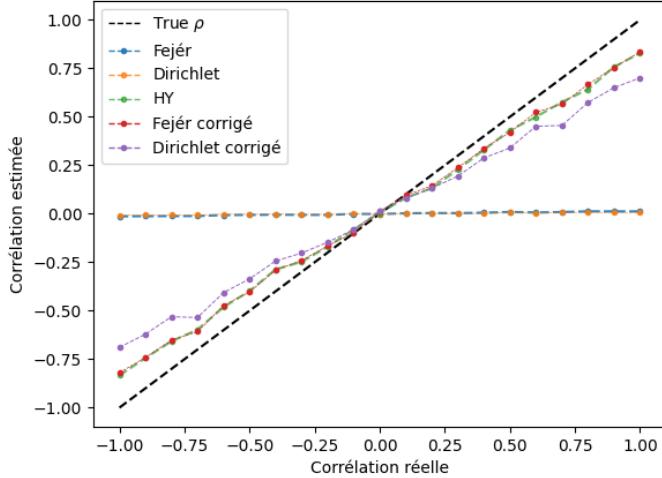


FIGURE 11 – Estimation de la corrélation sous échantillonnage exponentiel à  $N = 0.85n^{3/4}$

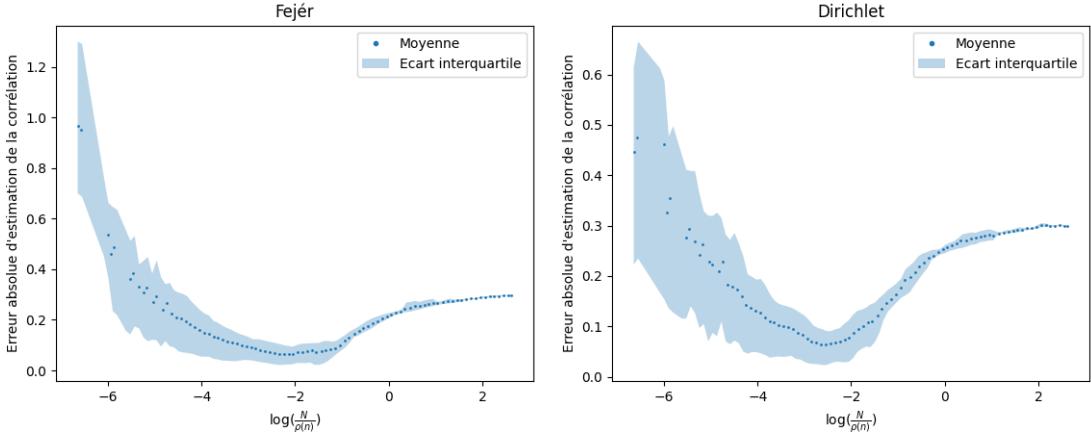
L'amélioration est drastique, l'estimateur de Malliavin Mancino reconstruit la corrélation avec une précision beaucoup plus fine et surperforme l'estimateur de Hayashi Yoshida. Maîtriser le rapport  $\frac{N}{n}$  est donc absolument crucial pour obtenir une estimation précise de la corrélation dans le cas asynchrone. Cette forte dépendance entre performances de l'estimateur de Malliavin Mancino et paramètres est ce qui va nous motiver à entreprendre la partie suivante.

## 5.5 Influence des paramètres sur la précision de l'estimateur

Comme mentionné précédemment, la précision de l'estimateur dépend fortement de la valeur du rapport  $\frac{N}{\rho(n)}$  choisi, avec pour cette section uniquement,  $\rho(n) := \min(n_1, n_2)$ ,  $n_1$  et  $n_2$  représentant la taille de chacun des signaux. Mais elle dépend aussi de la covariance entre les deux actifs. Dans cette partie, nous analysons l'influence des différents paramètres de simulation sur la précision.

### 5.5.1 Effet de $N$

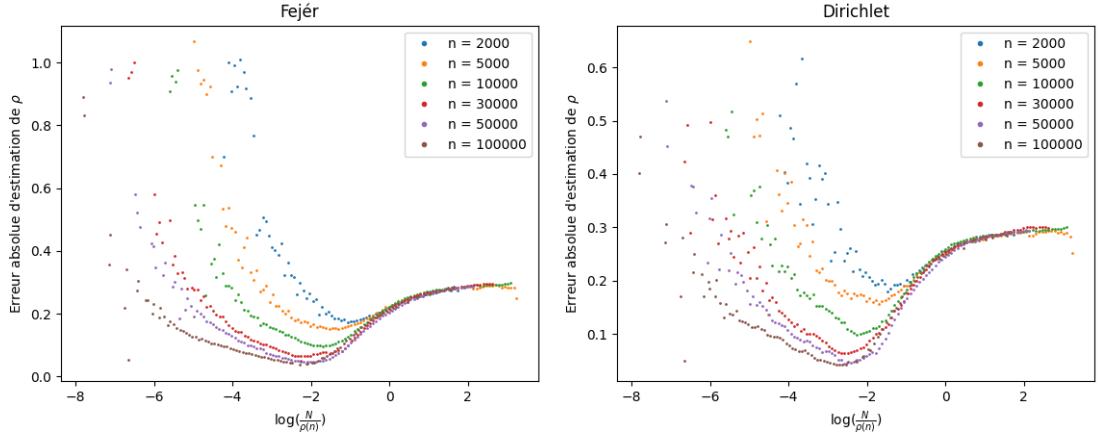
Pour rappel,  $N$  correspond au plus grand mode de fréquence calculé pour construire l'estimateur. Le nombre de modes de Fourier calculés au total pour un  $N$  fixé est  $2N + 1$ . Pour observer l'effet de la variance de  $N$  nous menons l'expérience suivante : deux actifs comptant  $n = 30000$  observations, de corrélation  $\rho = 0.3$ , suivant un mouvement brownien géométrique sont générés puis échantillonnes avec des exponentielles de paramètres  $\lambda_1 = 40$  et  $\lambda_2 = 30$ . Le nombre d'échantillons effectifs après échantillonnage par exponentielle est variable, on a donc en général  $n_1 \neq n_2$ , ce qui justifie notre choix d'étudier les erreurs selon  $\frac{N}{\rho(n)}$ . Cette simulation est répétée 100 fois, ce qui nous permet d'observer plus ou moins la distribution de l'effet et sa moyenne.


 FIGURE 12 – Erreurs obtenues en faisant varier  $\frac{N}{\rho(n)}$ 

L'intervalle des  $\log(\frac{N}{\rho(n)})$  générés est uniformément subdivisés en intervalles de mêmes pas. La moyenne des simulations est faite sur chacun de ces intervalles. On peut remarquer que l'inversion de Fejér est plus stable que celle de Dirichlet, l'erreur pour chaque trajectoire semble avoir une variance moindre pour la première méthode. En outre, on observe que l'erreur n'est pas monotone mais présente plutôt un minimum aux alentours de  $\log(\frac{N}{\rho(n)}) \approx -2.5$ . L'erreur est plus importante quand le logarithme est faible, ce qui correspond aux valeurs de  $N$  inférieures, ce qui enfreint donc la condition  $N \rightarrow \infty$  du théorème 6. De même, l'erreur est plus élevée pour des valeurs supérieures de  $N$ ; la condition  $\frac{N}{\min(n_1, n_2)} \rightarrow 0$  n'est en effet plus respectée. Il est donc important de trouver un compromis permettant de diminuer l'erreur.

### 5.5.2 Effet de $n$

On s'intéresse maintenant à une expérience plus générale que la précédente, où le nombre initial d'observations  $n \in \{2000, 5000, 10000, 30000, 50000, 100000\}$ . Les actifs sont générés dans les mêmes conditions que précédemment et échantillonnés avec les mêmes paramètres de lois exponentielles. Les courbes représentées sont toutes des moyennes.


 FIGURE 13 – Erreurs obtenues suivant  $\log\left(\frac{N}{\rho(n)}\right)$  à différents  $n$ 

On remarque qu'un nombre de points plus important permet en général une estimation plus précise, ce qui est en accord avec l'intuition et le fait que l'estimateur de Malliavin Mancino utilise toute l'information disponible. Théoriquement plus  $n$  augmente, plus on se situe dans les conditions d'application du théorème 6. Le rapport  $\frac{N}{\rho(n)}$  est également lui aussi diminué quand  $n$  augmente. Ceci justifie pourquoi les courbes correspondant aux  $n$  les plus grands se situent en général en-dessous des autres. On peut aussi remarquer que le nombre de points observés joue un rôle assez moindre au niveau des  $\log\left(\frac{N}{\rho(n)}\right)$  les plus élevés, ce qui souligne encore plus l'importance du paramétrage de ce rapport.

A partir de cette figure, nous nous sommes demandés s'il n'était pas possible d'émettre des recommandations sur le choix du  $N$  optimal. Nous avons alors pris les positions des minima pour chacune de ces courbes que nous avons tracés selon  $\rho(n)$ . La courbe obtenue est visible sur la figure 14. Il semblerait que sur nos tests effectués, qu'il y ait une relation linéaire entre  $\log\left(\frac{N_{opt}}{\rho(n)}\right)$  et  $\log(\rho(n))$  soit qu'il existe une relation de la forme  $N_{opt} = \alpha\rho(n)^\beta$ , autrement dit de la même forme que la recommandation donnée par Malliavin et Mancino. A partir de cette figure, nous avons  $N_{opt} = 0.094\rho(n)^{0.69}$  pour Fejér et  $N_{opt} = 0.058\rho(n)^{0.71}$  pour Dirichlet. Bien que les constantes multiplicatives ne soient pas du tout les mêmes, l'exposant de la puissance est presque identique : 0.7 avec nos simulations contre 0.75 pour leur recommandation. Enfin, gardons tout de même à l'esprit que la corrélation elle-même a une influence non négligeable sur la précision de l'estimation, point qui est développé dans la partie suivante.

## 5.5 Influence des paramètres sur la précision de l'estimateur

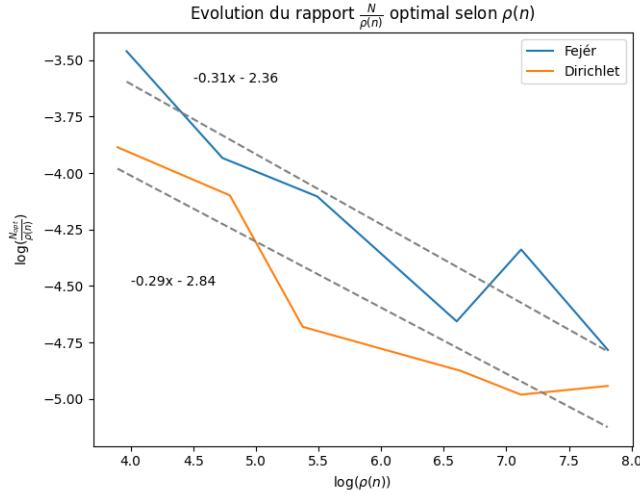


FIGURE 14 – Evolution du rapport optimal  $\frac{N}{p(n)}$  en fonction de  $p(n)$

Pour tester la validité de cette démarche, nous avons repris les courbes 27 et 10 avec  $N = 0.075\rho(n)^{0.7}$ . Pour ne pas surcharger les graphes, nous traçons uniquement les estimations par inversion de Fejér ; celle avec Dirichlet peuvent être trouvées en annexes. Les estimations sont meilleures qu'avec la correction proposée par Malliavin et Mancino. Cependant, les estimations avec cette nouvelle correction semblent de variance plus forte : il est nécessaire de faire un compromis entre la précision et la variance des estimations. Cela pouvait déjà s'observer sur la figure 12. La variance des erreurs est élevée pour de petites valeurs de  $N$  et diminue au fur et à mesure que  $N$  augmente, jusqu'à estimer que des 0.

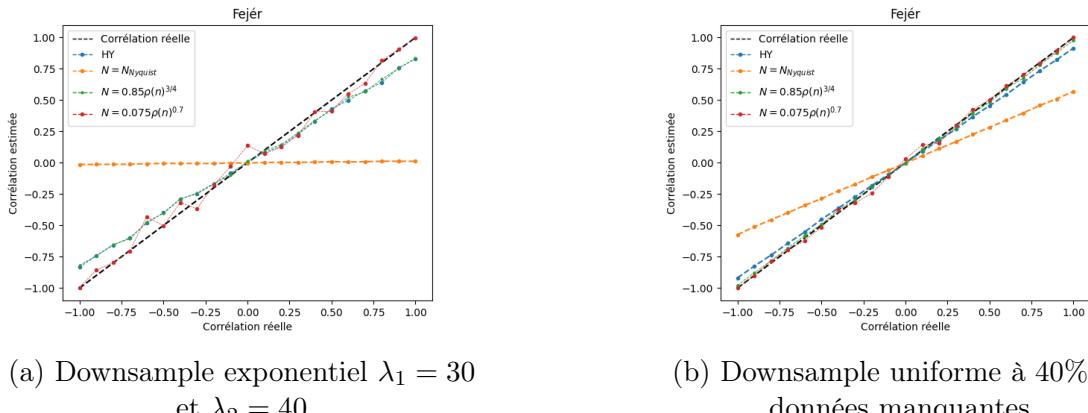


FIGURE 15 – Comparaison d'estimations de corrélations pour plusieurs fréquences de coupures

### 5.5.3 Effet de $\rho$

L'équation 8 fait également intervenir la covariance, ce qui nous a poussé à nous interroger sur l'effet de la corrélation sur l'estimation.

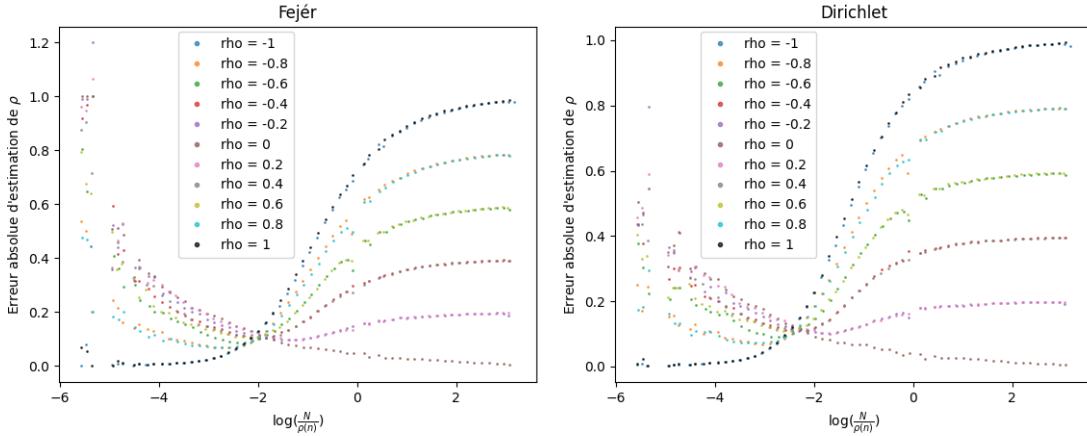


FIGURE 16 – Erreurs obtenues suivant  $\log(\frac{N}{\rho(n)})$  à différents  $\rho$

On peut constater que le rapport  $\frac{N}{n}$  n'est pas minimisé aux mêmes positions selon  $\rho$ , ce qui rend plus difficile la tâche de trouver une relation empirique pour optimiser l'estimation.

On peut remarquer que les différentes courbes tendent à se regrouper deux à deux vers les abscisses les plus élevées. En effet, il apparaît de l'équation 8 que l'erreur, quand la covariance est constante, dépend proportionnellement de la corrélation. La valeur absolue de l'erreur selon deux corrélations opposées doit donc être la même, ce qui est bien ce que l'on observe.

Constatons enfin que lorsque  $N$  devient grand, l'estimation n'est plus tout fiable : la valeur estimée est nulle quelle que soit la corrélation réelle.

### 5.5.4 Effet de l'échantillonnage

Nous nous sommes aussi intéressés à l'influence des paramètres de l'échantillonnage sur la précision de l'estimation. Les observations générées ont été échantillonées avec des lois exponentielles et uniformes de différents paramètres, indiqués en légende, tout en essayant d'avoir un nombre de points similaires pour ne pas trop influencer les résultats, comme observé précédemment.

## 5.5 Influence des paramètres sur la précision de l'estimateur

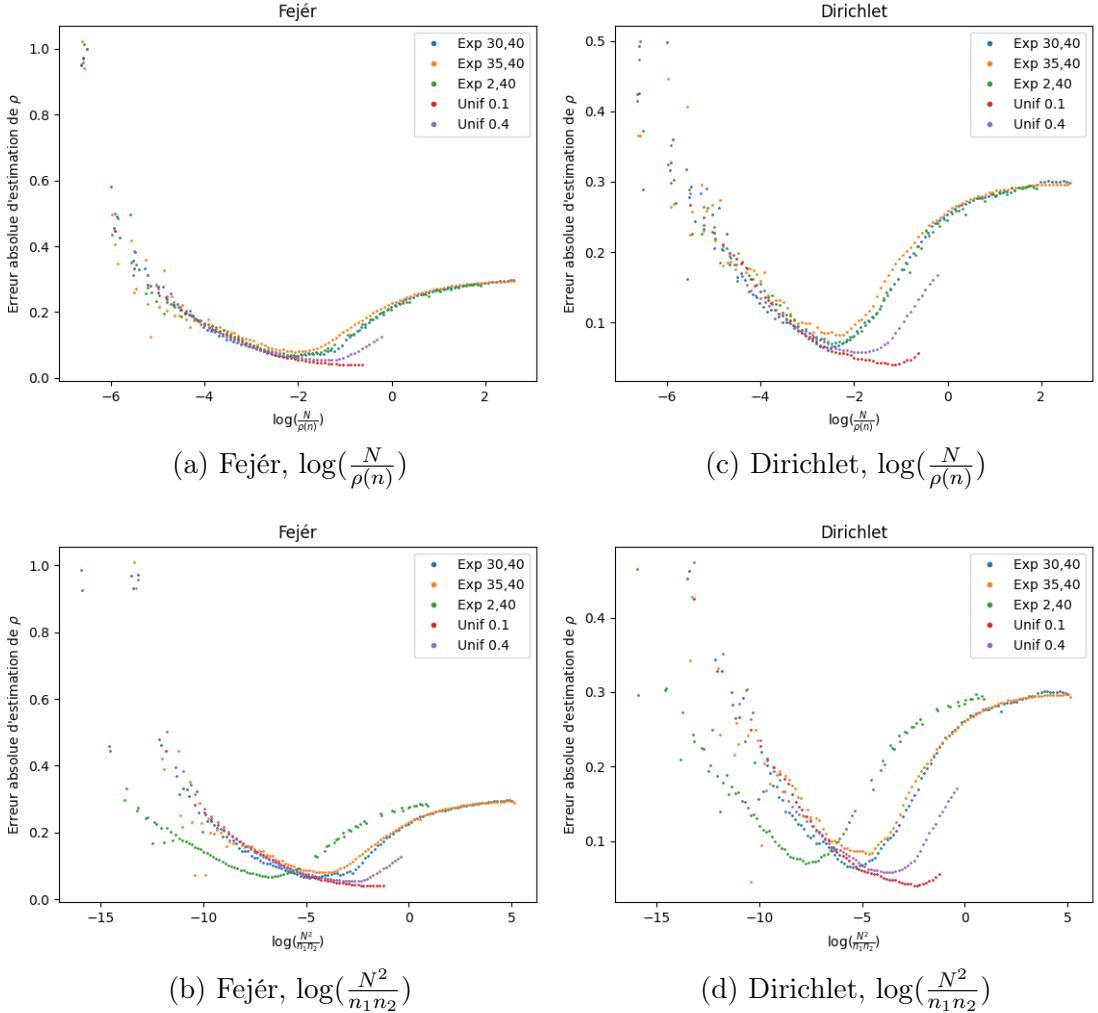


FIGURE 17 – Erreurs obtenues selon différents échantillonnages

Profitons de l'occasion pour justifier le choix de l'étude en fonction du rapport  $\frac{N}{\rho(n)}$ . Comme les quantités  $n_1$  et  $n_2$  étaient variables, nous pensions qu'il aurait aussi été pertinent d'étudier la précision selon  $\frac{N^2}{n_1 n_2}$ , mais comme nous pouvons le voir sur la figure 17, l'étude en fonction du rapport  $\frac{N}{\rho(n)}$  est plus judicieux ; les minima semblent être moins dispersés.

Remarquons également que la façon de choisir les échantillons a une influence non-négligeable : alors que pour les lois de Poisson ont des positions de minima assez proche, ceux obtenus par sous-échantillonage uniforme à 10% ou 40% sont plus éloignés, quel que soit le rapport étudié. Pour pouvoir conclure réellement, il faudrait faire plus de tests : il faudrait sans doute changer les modèles de simulations de prix et éventuellement simuler des bruits de microstructure ; il faudrait également prendre la façon la plus réaliste de choisir des échantillons.

## 5.6 Optimisation de l'estimateur via transformée de Fourier non uniforme

Après avoir étudié les propriétés de convergence de l'estimateur et après nous être intéressés à l'influence de divers paramètres pouvant intervenir dans les simulations, nous nous intéressons maintenant à optimiser la vitesse et la précision de l'estimateur en appliquant les techniques de calcul reposant sur la transformée de Fourier non uniforme détaillées dans la partie précédente.

### 5.6.1 Rapidité de l'estimation

On commence par vérifier le gain en rapidité introduit par l'utilisation de la transformée de Fourier non uniforme. Nous implémentons les noyaux précédemment définis, GS désigne le noyau gaussien, KB le noyau de Keiser-Bessel et ES le noyau de l'exponentielle de semi-cercle. On s'intéresse à la rapidité de ces méthodes suivant la tolérance dans un premier temps puis suivant le nombre de points d'observation dans un deuxième temps.

#### Rapidité sous différents niveaux de tolérance :

La simulation consiste à générer deux mouvements brownien géométriques synchrones de 10000 observations et de calculer leur matrice de covariance avec ces trois implantations puis de comparer leurs temps d'exécution au temps d'exécution de l'implantation sous forme matricielle du calcul initial proposé par Malliavin et Mancino, désigné par Vectorized au niveau des légendes des différentes figures. Les corrélations entre ces mouvements browniens sont générées aléatoirement.

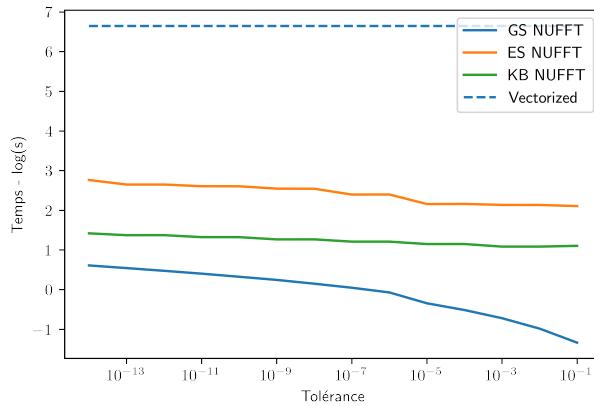


FIGURE 18 – Temps d'estimation en fonction de la tolérance

Observons tout d'abord que la courbe de Vectorized est cohérente avec celle attendue ; en effet son calcul est exact et ne dépend pas d'une tolérance. Quant à l'interprétation des résultats, on remarque tout de suite que l'utilisation de la transformée de Fourier non uniforme avec le noyau gaussien et le noyau de Keiser-Bessel

permet un gain considérable au niveau du temps d'exécution. Ces deux implémentations sont plus rapides que celle proposée par Malliavin et Mancino quelque soit la tolérance imposée. L'utilisation de l'exponentielle de semi-cercle retourne des résultats un peu moins satisfaisants. Nous pensons que cela est dû à un calcul pas assez efficace de la transformée de Fourier de ce noyau reposant sur l'utilisation de la fonction **quad** de **scipy.integrate**, qui est assez lente.

### Rapidité sous différents nombres d'observations :

La simulation consiste cette fois à varier le nombre d'observations de deux vecteurs prix synchrones suivant un mouvement brownien géométrique afin d'observer le gain en rapidité engendré, la tolérance pour les méthodes reposant sur la transformée de Fourier non uniforme est prise égale à  $10^{-6}$ . La corrélation entre les actifs est générée aléatoirement. Les courbes suivantes représentent une moyenne sur 10 simulations des résultats.

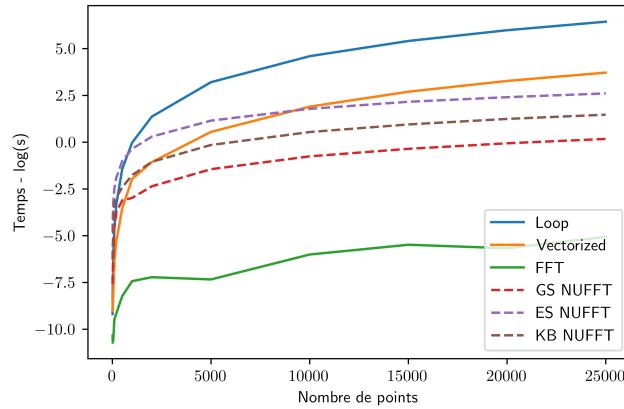


FIGURE 19 – Evolution du temps de calcul en fonction du nombre d'observations

Le gain en vitesse est d'autant plus grand que le nombre d'observations est grand. Les courbes des méthodes Vectorized et Loop ont bien une allure en  $2\log(N)$ . La FFT reste la méthode de calcul la plus rapide, en effet elle ne comporte pas les étapes de convolution et de déconvolution que les NUFFT comportent. Mais celle-ci ne peut être utilisée que dans un cas synchrone tel que celui-ci. L'utilisation de la transformée de Fourier non uniforme permet donc de généraliser l'utilisation de la FFT et d'exploiter sa rapidité mais cela se fait au prix de l'introduction de nouvelles opérations qui ralentissent de manière assez considérable le calcul.

### 5.6.2 Précision de l'estimation

#### Cas synchrone

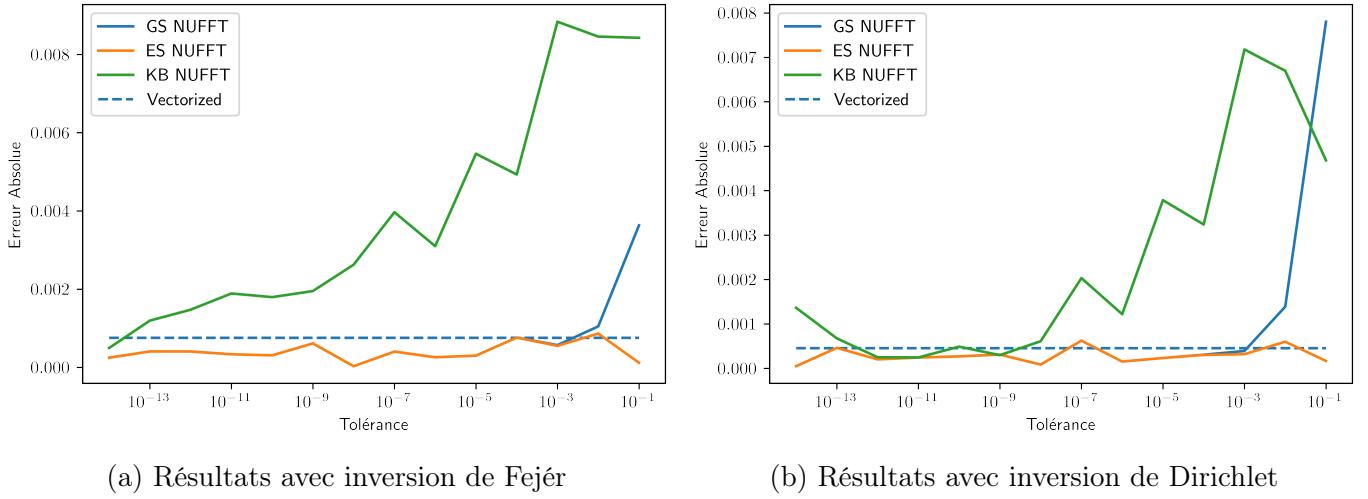


FIGURE 20 – Evolution de l'erreur absolue sur la variance en fonction de la tolérance

La précision de l'estimation de la variance dans le cas synchrone, quelque soit le type d'inversion utilisé est en général meilleure pour l'exponentielle de semi-cercle et le noyau gaussien jusqu'à un seuil de tolérance d'à peu près  $10^{-2}$ . Cette meilleure précision peut possiblement s'expliquer par le fait que plus la tolérance est réduite, meilleure est l'estimation des coefficients de Fourier, alors que le calcul initial proposé par Malliavin et Mancino repose sur de simples approximations qui ne permettent pas de régler leurs précisions suivant une tolérance donnée. Le noyau de Kaiser-Bessel cependant exhibe une précision beaucoup moins bonne.

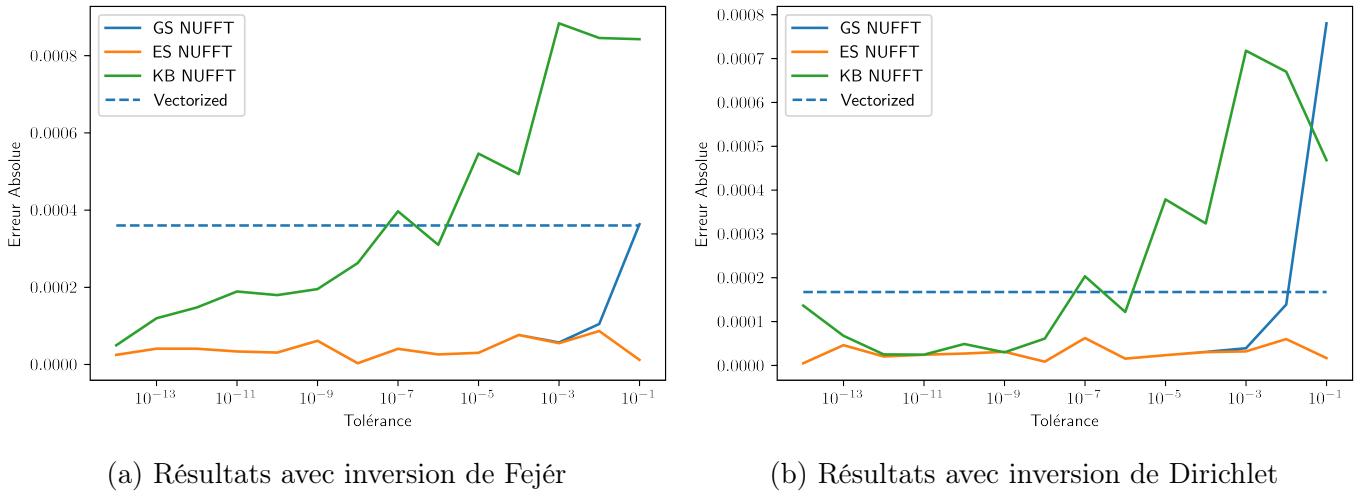


FIGURE 21 – Evolution de l'erreur absolue sur la covariance en fonction de la tolérance

## 5.6 Optimisation de l'estimateur via transformée de Fourier non uniforme

On peut observer les mêmes résultats en ce qui concerne la covariance, la transformée de Fourier non uniforme permet d'obtenir des résultats en général plus précis avec les noyaux gaussien et exponentielle de semi-cercle mais le noyau de Keiser Bessel a encore une fois une erreur absolue moyenne plus élevée que les autres.

### Cas asynchrone - Downsample

La précision du noyau gaussien et de l'exponentielle pour l'estimation de la variance de semi-cercle est encore une fois comparable à celle de l'implémentation initiale de Malliavin et Mancino pour l'inversion de Fejér ainsi que celle de Dirichlet, comme on peut l'observer sur les figures 22 et 23. La précision est même parfois meilleure à des niveaux de tolérances plus faibles. Encore une fois, la précision du noyau de Keiser-Bessel est beaucoup moins bonne. On peut observer que l'implémentation initiale de Malliavin et Mancino retourne des résultats légèrement meilleurs quand le pourcentage de données manquantes est plus faible.

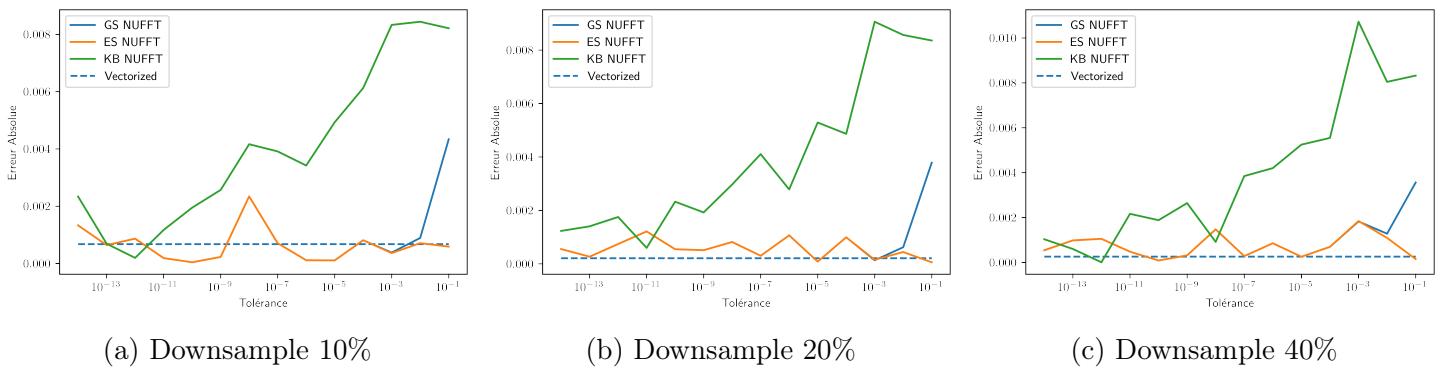


FIGURE 22 – Evolution de l'erreur absolue sur la variance selon la tolérance - Inversion de Fejér

On peut observer des résultats similaires pour l'inversion de Dirichlet, avec une erreur amplifiée aux niveaux des tolérances les plus grandes pour le noyau gaussien.

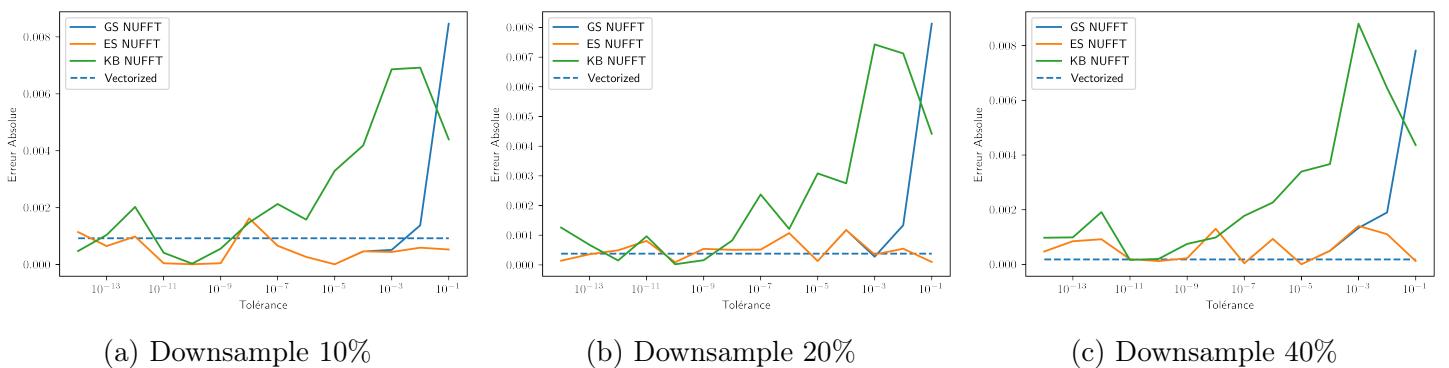


FIGURE 23 – Evolution de l'erreur absolue sur la variance selon la tolérance - Inversion de Dirichlet

## 5.6 Optimisation de l'estimateur via transformée de Fourier non uniforme

Au niveau de l'erreur absolue pour la covariance, on remarque que, que ce soit pour l'inversion de Fourier ou celle de Dirichlet, l'utilisation de la transformée de Fourier non uniforme permet d'obtenir une meilleure précision. Une piste pour expliquer cela est que l'estimation initiale de Malliavin et Mancino repose sur une approximation de l'intégrale calculant les coefficients de Fourier en considérant ses termes comme constants. Cette approximation ne peut que perdre en qualité quand les écarts entre les temps d'observations deviennent plus grands lorsqu'un downsample est conduit. La transformée de Fourier non uniforme, quant à elle, vise à calculer les coefficients de Fourier réels sous un seuil de tolérance donné, la précision de calcul des coefficients n'est donc pas censée être affectée par des temps plus longs entre prix successifs.

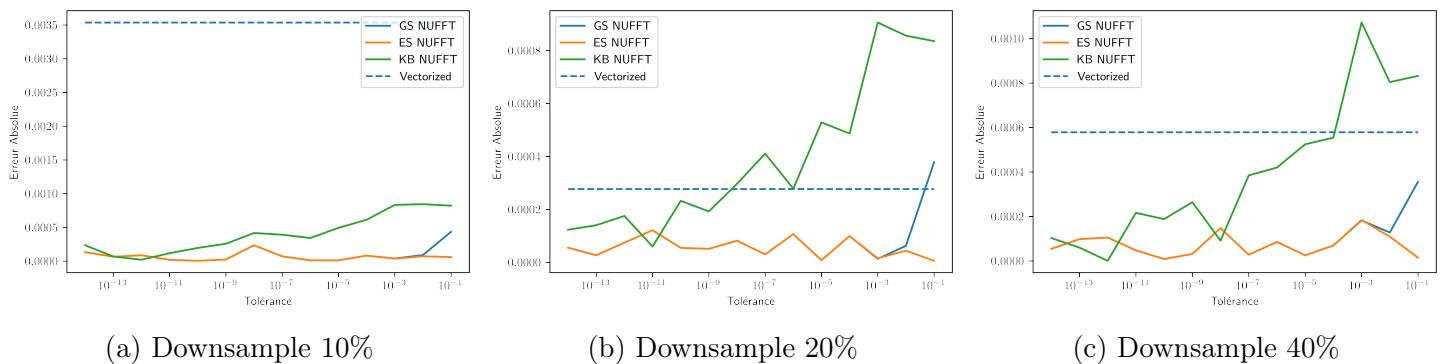


FIGURE 24 – Evolution de l'erreur absolue sur la covariance selon la tolérance - Inversion de Fejér

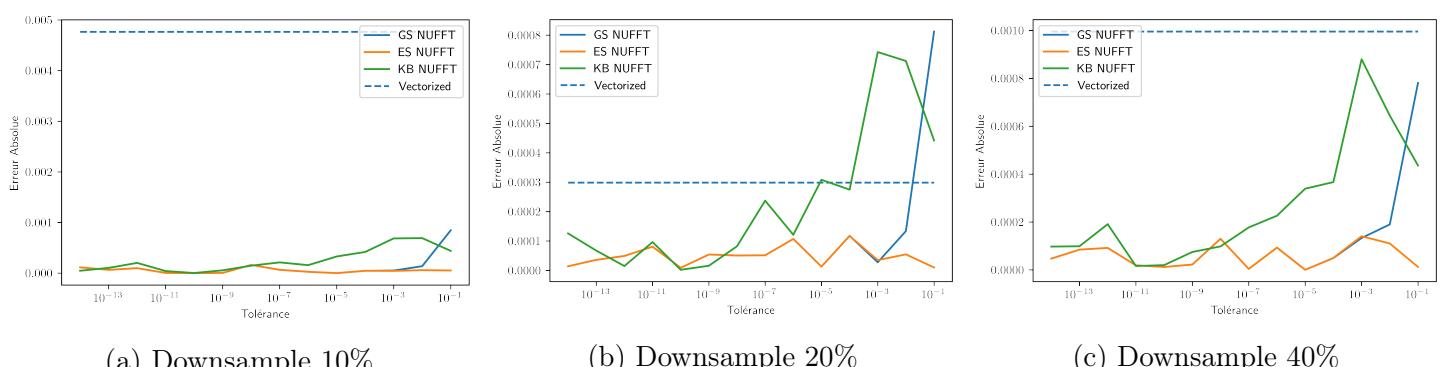


FIGURE 25 – Evolution de l'erreur absolue sur la covariance selon la tolérance - Inversion de Dirichlet

### Cas asynchrone - Échantillonnage exponentiel

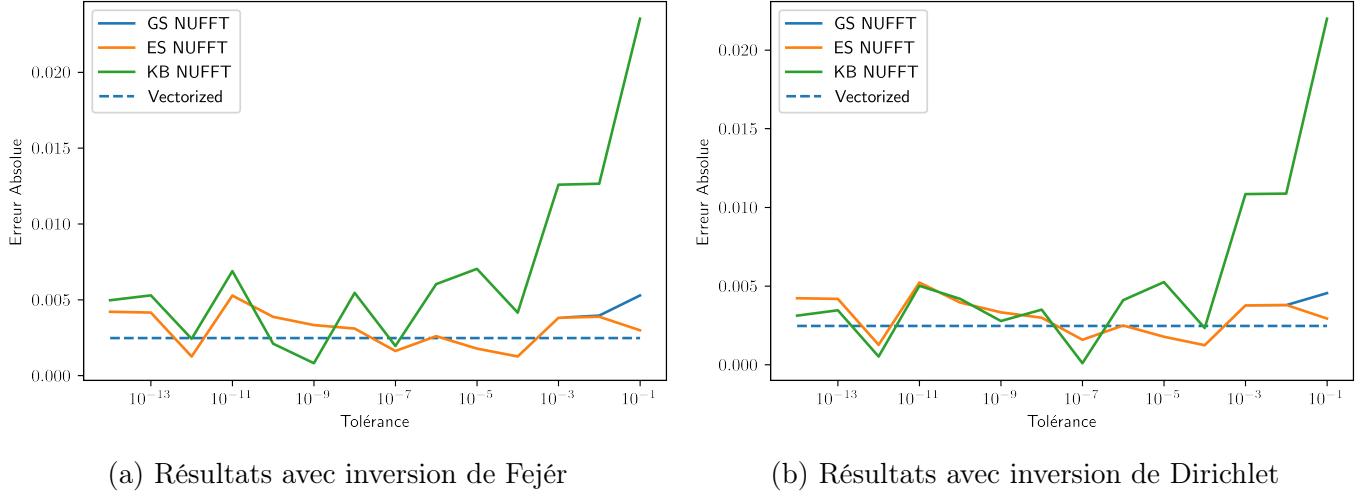


FIGURE 26 – Evolution de l'erreur absolue sur la variance en fonction de la tolérance

On remarque, que comme pour le Downsample, la précision pour l'inversion de Fejér comme pour l'inversion de Dirichlet est généralement comparable à celle de Vectorized, exception faite du noyau de Keiser-Bessel dont la précision reste encore une fois moins performante que celle des autres noyaux. En ce qui concerne la covariance, on remarque que les trois noyaux présentent de meilleurs résultats que pour Vectorized. Une piste d'explication à laquelle nous avons pensé est que dans le cas de l'échantillonnage exponentiel, l'approximation faite par Malliavin et Mancino pour leurs calculs d'intégrales se détériorent fortement.

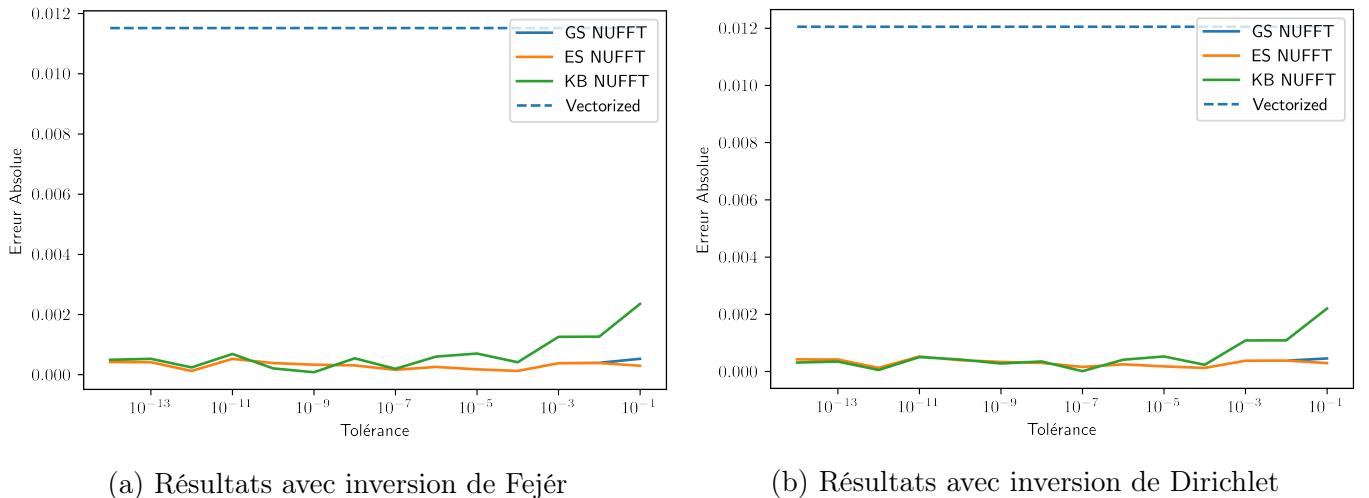


FIGURE 27 – Evolution de l'erreur absolue sur la covariance en fonction de la tolérance

### 5.6.3 Synthèse des résultats

Cette partie nous a permis de voir que l'utilisation de la transformée de Fourier non uniforme permet effectivement d'obtenir une réduction très intéressante du temps de calcul, avec une complexité en  $O(n \log(n))$  plutôt qu'une complexité quadratique tout en gardant une précision comparable voire meilleure dans certains cas, particulièrement celui de l'échantillonnage exponentiel. En particulier, la transformée de Fourier non uniforme faisant intervenir **le noyau gaussien exhibe les meilleurs résultats en termes de rapidité et de précision** du moment que la tolérance est maintenue à un niveau inférieur à  $10^{-2}$ . **Nous concluons donc au terme de ces simulations que c'est le meilleur noyau à utiliser.** Ceci nous pousse donc à le retenir plutôt que les autres lors de l'analyse des données réelles.

---

## 6 Performance de l'estimation de la covariance instantanée

### 6.1 Modèle de Heston généralisé

Le mouvement brownien géométrique ne permet de générer qu'un vecteur de prix d'actifs à matrice de covariance constante. Nous souhaitons également nous intéressons à un modèle où la volatilité et la covariance varient en fonction du temps. Nous avons donc décidé d'implémenter le modèle de Heston généralisé à plus d'un actif, tel que présenté dans [5] :

$$\begin{cases} X(t) = X(0) + \int_0^t -\frac{1}{2}\Sigma^{\text{diag}}(s)ds + \int_0^t \sqrt{\Sigma(s)}dZ \\ \Sigma(t) = \Sigma(0) + \int_0^t (b + M\Sigma(t) + \Sigma(t)M^\top) dt + \sqrt{\Sigma(t)}dBH + HdB^\top\sqrt{\Sigma(t)} \end{cases}$$

Où  $\Sigma(t)$  est la matrice de covariance à l'instant  $t$ ,  $X(t)$  est le vecteur des log-prix à l'instant  $t$ ,  $Z$  et  $B$  sont respectivement un vecteur bidimensionnel et une matrice  $2 \times 2$  de mouvements browniens et  $Z = \sqrt{1 - \rho^\top \rho}W + B\rho$  où  $\rho \in [-1, 1]^2$  tel que  $\rho^\top \rho \leq 1$  et  $W$  un vecteur 2D de mouvements browniens indépendant de  $B$ .  $M$  et  $H$  sont des matrices inversibles et  $b$  est une matrice 2D telle que  $b - H^2 \in S_2^+$ , cette condition étant nécessaire pour que  $\Sigma$  soit symétrique définie positive à tout instant. Devant la difficulté de régler les paramètres de ce modèle, nous empruntons pour nos simulations les paramètres de [5] :

Paramètre	Valeur
$(X_0^1, X_0^2)$	(4.6, 4.6)
$\begin{pmatrix} \Sigma^{11}(0) & \Sigma^{12}(0) \\ \Sigma^{12}(0) & \Sigma^{22}(0) \end{pmatrix}$	$\begin{pmatrix} 0.09 & -0.036 \\ -0.036 & 0.09 \end{pmatrix}$
$M$	$\begin{pmatrix} -1.6 & -0.2 \\ -0.4 & -1 \end{pmatrix}$
$\alpha = H^2$	$\begin{pmatrix} 0.0725 & 0.06 \\ 0.06 & 0.1325 \end{pmatrix}$
$b$	3.5 $\alpha$
$\rho$	(-0.3, -0.5)

Un exemple de matrice de covariance ainsi générée avec ces paramètres est :

## 6.2 Cas synchrone

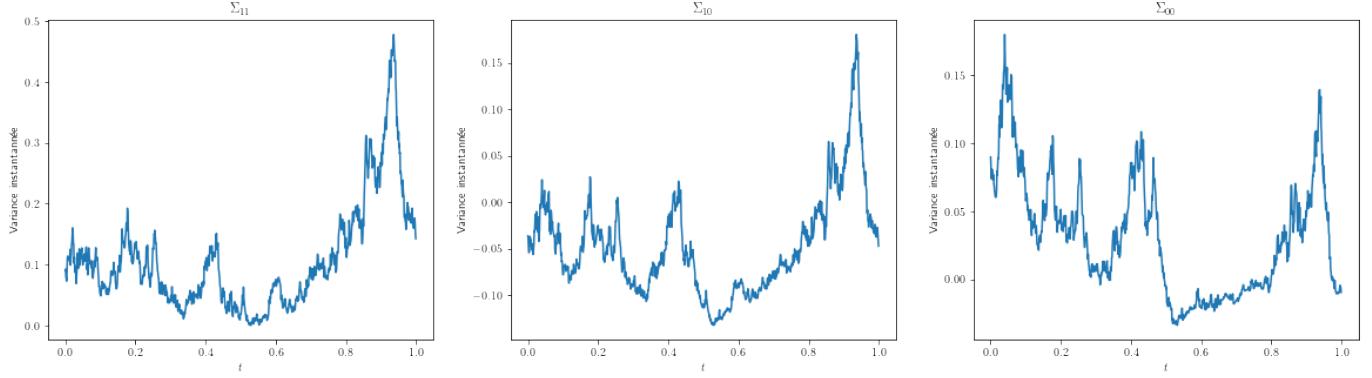


FIGURE 28 –  $\Sigma(t)$  - Modèle de Heston

## 6.2 Cas synchrone

On rappelle l'expression de l'estimateur de la covariance instantanée donné par Malliavin et Mancino, sous l'inversion de Fejér :

$$\Sigma_{M,N}^{l,j}(t) = \sum_{k=-M}^{k=M} \left(1 - \frac{|k|}{M}\right) \frac{2\pi}{N+1} \sum_{s=-N}^{s=N} \mathcal{F}(dp_l)(s) \mathcal{F}(dp_j)(k-s) \exp(ikt)$$

Dans le cas synchrone, Mancino conseille dans [16] d'adopter les paramètres  $N = \frac{n}{2}$  et  $M = \frac{1}{2\pi 8} \sqrt{n \log(n)}$ . Nous implémentons donc ainsi l'estimation. Un exemple de résultats, pour deux prix de 10000 observations suivant le modèle de Heston :

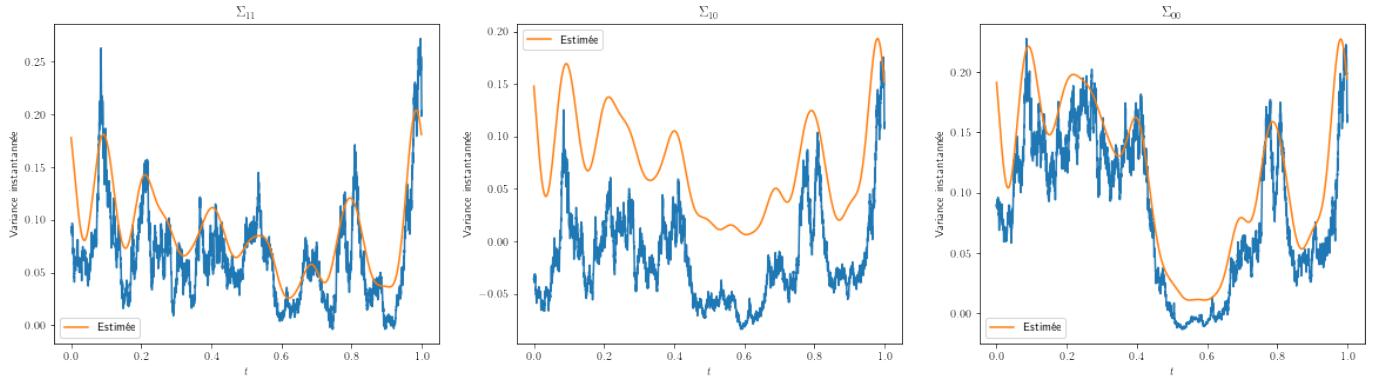


FIGURE 29 – Exemple d'estimation de la covariance instantanée - Modèle de Heston

De même, un exemple d'estimation pour deux prix synchrones de 10000 observations encore une fois suivant un mouvement brownien géométrique est :

### 6.3 Cas asynchrone

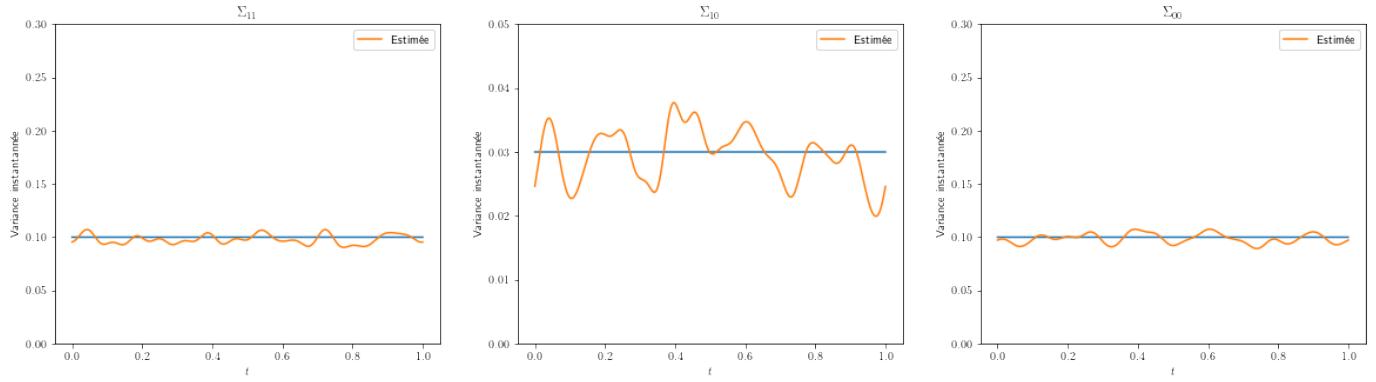


FIGURE 30 – Exemple d'estimation de la covariance instantanée - Modèle GBM

On remarque directement que l'estimateur semble suivre et capturer les tendances de la variance. L'estimation de la covariance présente plus de fluctuations et semblent même comporter un biais positif. Dans les deux cas, l'estimateur rend compte des oscillations majeures.

On s'intéresse à la distribution limite de l'erreur, étant donné que le théorème 8 fournit un résultat théorique intéressant dessus. On choisit donc deux instants aléatoires, en l'occurrence  $t = 0.098$  et  $t = 0.352$ , pour lesquels on étudie la distribution de l'erreur sur la variance sur 1000 expériences un actif de 1000 observations suivant un GBM de variance  $\sigma = \sqrt{0.1}$  :

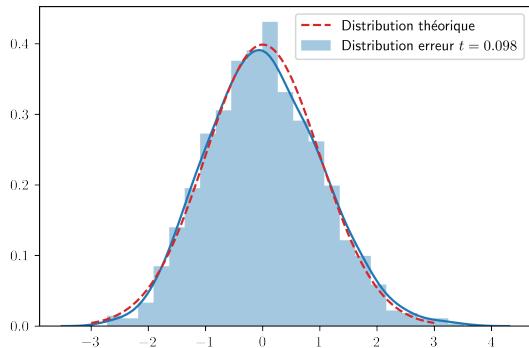


FIGURE 31 – Distribution de  $\sqrt{\frac{3n}{4M}} \frac{(\hat{\sigma}_{n,N,M}^2(t) - \sigma^2(t))}{\sigma^2(t)}$  à  $t = 0.098$

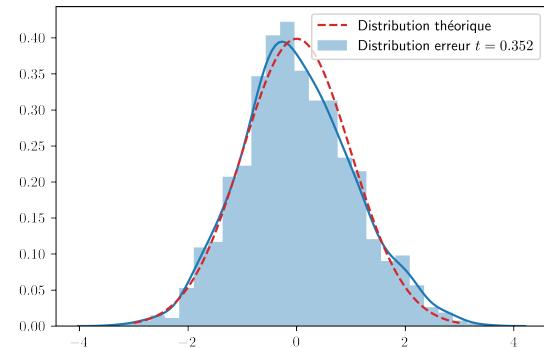


FIGURE 32 – Distribution de  $\sqrt{\frac{3n}{4M}} \frac{(\hat{\sigma}_{n,N,M}^2(t) - \sigma^2(t))}{\sigma^2(t)}$  à  $t = 0.352$

On peut donc bien observer empiriquement le théorème de convergence 8. Les tests de normalité de Shapiro Wilk retournent des p-valeurs de 0.03 et de 0.01 pour respectivement  $t = 0.098$  et  $t = 0.352$ .

### 6.3 Cas asynchrone

On s'intéresse maintenant au cas asynchrone, on peut visualiser sur les figures suivantes des exemples d'estimation de la matrice de covariance instantanée dans

### 6.3 Cas asynchrone

différentes situations de prix asynchrones suivant un modèle de Heston avec les paramètres explicités précédemment :

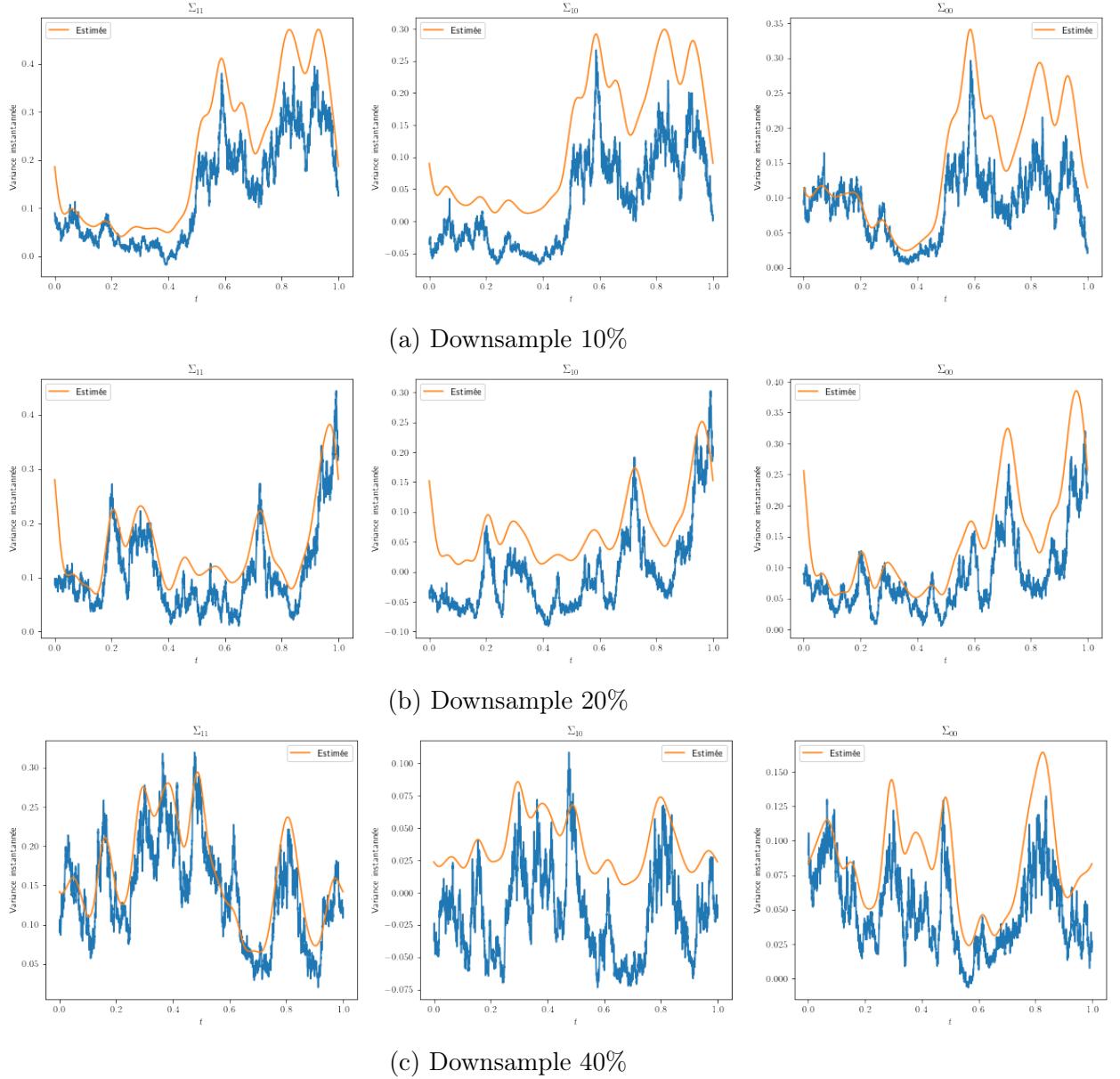


FIGURE 33 – Estimation dans le cas où les données asynchrones sont générées par Downsample

L'estimation de la variance d'un même actif reste assez précise et recouvre toujours les tendances majeures de la volatilité. L'estimation de la covariance semble également retracer la tendance des variations de la courbe réelle mais présente un biais positif. Encore une fois, l'explication donnée à ceci par [5] et [16] est le non respect en pratique des conditions de convergence du théorème 9. La solution proposée est donc de régler les paramètres  $N$  et  $M$  de manière empirique de façon à ce que l'erreur absolue ou relative soit minimisée. Ce réglage est d'autant plus im-

### 6.3 Cas asynchrone

portant pour obtenir une estimation satisfaisante de la covariance entre deux prix différents, étant donné que dans certains cas, à l'image du cas synchrone, choisir des paramètres inadéquat peut conduire à un échec total de l'estimation. Un exemple d'un tel cas est présenté dans la figure suivante, où l'asynchronicité est générée par échantillonnage exponentiel de paramètres  $\lambda_1 = 4$  et  $\lambda_2 = 6$  :

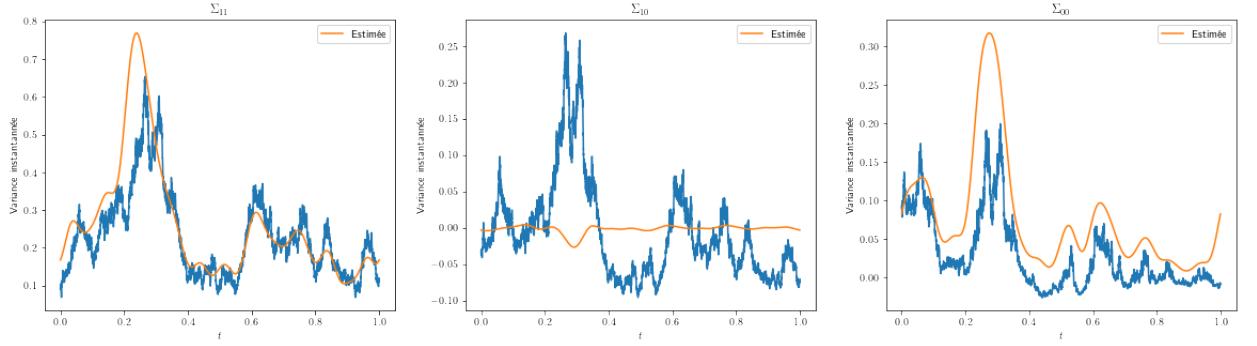


FIGURE 34 – Exemple d'estimation dans le cas où les données asynchrones sont générées par échantillonnage exponentiel

L'article [5] explore plus extensivement l'effet des fréquences de coupure  $M$  et  $N$  sur l'estimation mais à part déduire empiriquement une fréquence à fixer à partir de simulations, ne conclut pas de méthode précise permettant d'avoir une idée des bonnes fréquences de coupure à prendre. L'article précise également que si  $M$  est trop petit, l'estimateur ne parvient pas à recouvrir les détails de l'évolution de la corrélation mais quand il est trop large, l'estimation finit par présenter des motifs de zigzags rapides qui relèvent plus du bruit engendré par la présence d'harmoniques trop nombreuses. Nous avons essayé d'observer ces effets suite à une simulation de deux prix suivant le modèle de Heston de 10000 observation, dans le cas asynchrone par échantillonnage exponentiel, et nous obtenons les résultats suivants par moyenage sur 10 itérations :

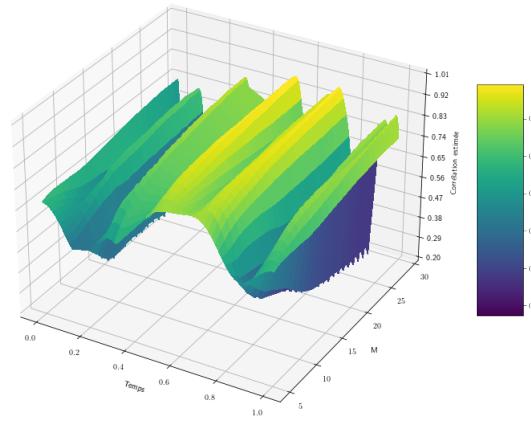


FIGURE 35 – Evolution de l'estimation de la corrélation instantanée en fonction de  $M$

## 6.4 Synthèse

L'estimateur instantané de la covariance semble très prometteur, et d'après les expériences que nous avons conduites, semble capable de capturer les tendances de la variance et de rendre compte de la covariance avec un biais assez moindre, du moins dans le cas synchrone et dans le cas d'asynchronicité générée par sous-échantillonage uniforme. Toutefois, il serait malvenu de généraliser aussi rapidement et le sujet mérite d'être étudié beaucoup plus en profondeur. Les performances sont aussi fortement dépendantes des fréquences de coupure choisies dans le cas asynchrone, il serait très intéressant de conduire des simulations visant à quantifier cette dépendance et déduire des expressions empiriques permettant d'avoir des résultats fiables. Nous n'avons pas eu le temps d'explorer cela mais nous sommes convaincus que ce serait une extension très intéressante de notre travail puisque que très peu d'estimateurs offrent la possibilité d'estimer directement la covariance de manière instantanée de manière stable.

---

## 7 Application sur des données réelles

### 7.1 Description des données

Les données que nous étudions sont les données de transaction ce la BNP, EDF, LVM, PRTP et SOGN sur le mois de juin 2015. Les transactions avant 10h et après 16h ont été supprimées, étant donné que ces périodes sont sujettes à une forte activité qui pourrait artificiellement modifier nos résultats. Nous effectuons un léger traitement des données : lorsqu'il y a plusieurs transactions en même temps, nous choisissons de garder le prix de la dernière enregistrée. Une autre approche serait de faire une moyenne pondérée par les volumes, de sorte que la valeur obtenue reflète l'influence sur le prix.

	Nombre de transactions	$\frac{1}{\lambda}$ (s)	Intervalle de confiance à 95%
BNPP	96765	5.72	[5.69, 5.76]
EDF	27048	20.44	[20.20, 20.69]
LVMH	51957	10.65	[10.56, 10.74]
PRTP	25144	21.98	[21.71, 22.25]
SOGN	108493	5.10	[5.08, 5.14]

TABLE 1 – Estimation des paramètres des exponentielles modélisant les temps d'attentes

La colonne  $\frac{1}{\lambda}$  indique la moyenne des temps en secondes entre deux transactions successives sur une même journée. Si on suppose que le temps écoulé entre les instants d'observation suit une loi exponentielle,  $\lambda$  est l'estimateur par maximum de vraisemblance du paramètre de cette loi. L'intervalle de confiance donné est un intervalle de confiance asymptotique, étant donné le nombre important de transactions, son utilisation est légitime.

Cette estimation nous a permis d'obtenir, par simulation de Monte Carlo, les fréquences de coupures  $N$  optimales à utiliser dans l'estimation de la corrélation intégrée grâce à la relation  $N = 0.094 \min(n_1, n_2)^{0.69}$  pour Fejér et  $N = 0.058 \min(n_1, n_2)^{0.71}$  pour Dirichlet. Naturellement, les  $N$  ainsi obtenus sont influencés par la modélisation sous-jacente des données.

### 7.2 Corrélation et effet Epps

Dans un premier temps, afin de pouvoir comparer les valeurs de corrélation obtenues avec les estimateurs de Malliavin Mancino ainsi que de Hayashi Yoshida, nous choisissons de procéder de manière classique, c'est-à-dire en ré-échantillonnant les séries temporelles avec des intervalles de temps réguliers. Cela nous permet alors d'appliquer des estimateurs synchrones, comme celui de Pearson. On peut ainsi observer l'effet Epps qui est plus ou moins présent selon le couple d'actifs observé sur les figures 36 et 37.

## 7.2 Corrélation et effet Epps

---

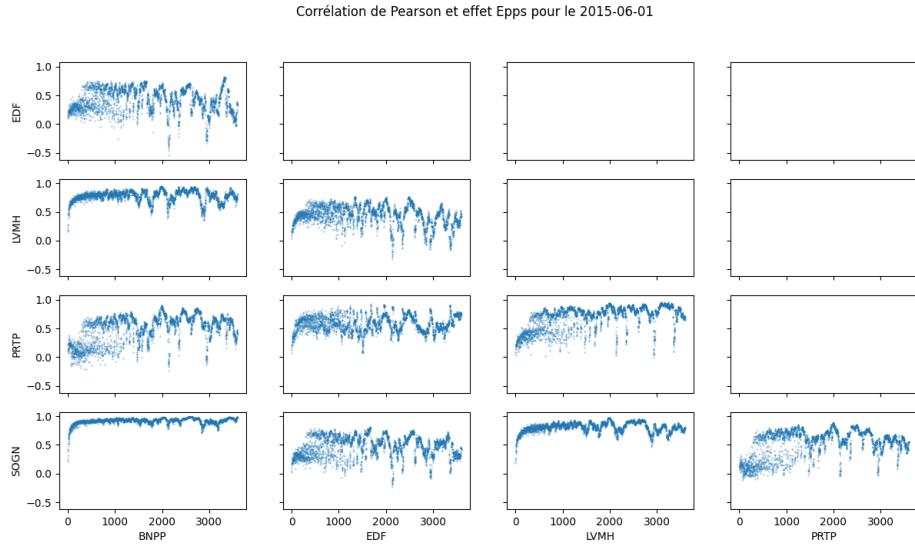


FIGURE 36 – Corrélations au 01/06/2015

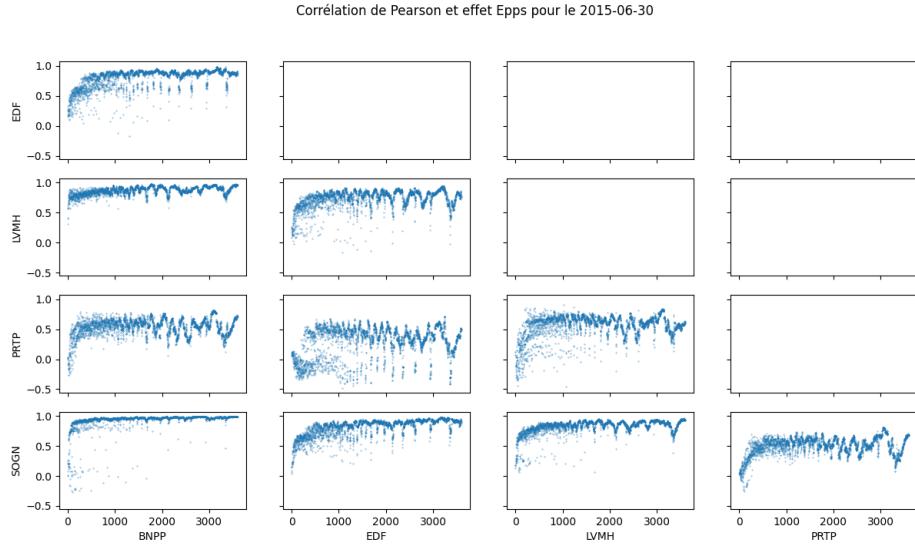


FIGURE 37 – Corrélations au 30/06/2015

Cette première étape préliminaire nous permet de mettre en évidence que les données présentent bien une dépendance entre le pas d'observation et la corrélation obtenue. Si l'estimateur de Hayashi-Yoshida ne demande pas de réglage particulier, il est nécessaire de spécifier une fréquence d'échantillonnage dans le domaine des fréquences à celui de Malliavin Mancino, ce qui revient à déterminer le nombre de coefficients de Fourier à calculer. Dans [18] et [7], c'est cette spécificité de l'estimateur de Malliavin Mancino qui est utilisée pour pouvoir analyser l'effet Epps. C'est ce que nous avons fait sur les figures 39 et 38. Nous observons le même type d'allure que dans [18]. Encore une fois, nous constatons que l'inversion de Fejér est plus robuste

### 7.3 Etude des données sur le mois

---

que celle de Dirichlet, mais cela ne veut pas dire qu'elle est forcément plus précise.

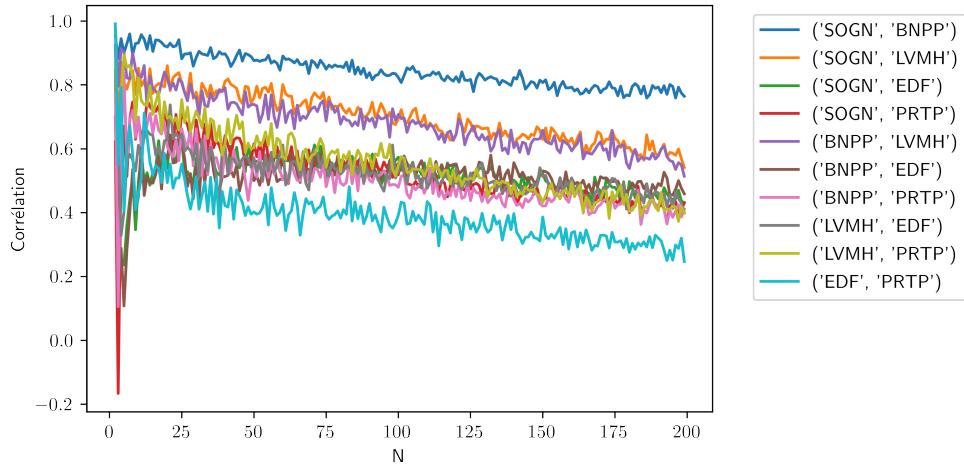


FIGURE 38 – Effet de la fréquence de coupure - Inversion de Fejér - 01/06/2015

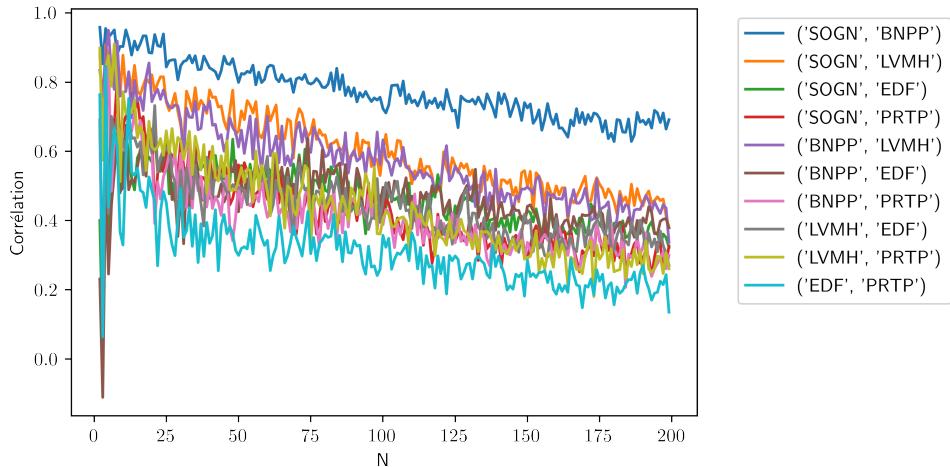


FIGURE 39 – Effet de la fréquence de coupure - Inversion de Dirichlet - 01/06/2015

On observe bien que la corrélation diminue avec la fréquence de coupure, ce qui correspond bien aux observations de Epps, qui concluent que les corrélations diminuent quand l'intervalle de temps considéré est diminué. La même remarque peut être refaite dans ce cas si on se remémore la relation donnée par Reno dans [18] et l'article [7] :

$$N = \left\lfloor \frac{1}{2} \left( \frac{T}{\Delta t} - 1 \right) \right\rfloor \quad (*)$$

### 7.3 Etude des données sur le mois

Nous avons tout d'abord appliqué les estimateurs pour chaque jour, sur les données entre 10h et 16h. Les résultats sont présentés sur la figure 40. Commençons par soulever un avantage inattendu de l'estimateur de Malliavin-Mancino : le  $N$  optimal est en général très petit devant la fréquence de Nyquist. Les temps de calculs s'en

### 7.3 Etude des données sur le mois

trouvent alors fortement diminués : alors que les estimations avec Hayashi-Yoshida ont pris 627s en moyenne par jour, celles avec Malliavin-Mancino n'en n'ont pris que 0.16s.

Pour les résultats, en termes de covariances, les estimations avec Hayashi-Yoshida semblent plus stables et donnent des résultats de covariances plus faibles par rapport à celles avec Malliavin-Mancino, et les allures sont simulaires. Pour la corrélation, la différence est plus marquée. En effet, la corrélation trouvée l'estimateur de Hayashi-Yoshida varie autour de 0.5 alors que celle trouvée avec Malliavin-Mancino est plutôt aux alentours de 0.8 pour leur recommandation et 0.9 avec la régression linéaire. Ces valeurs sont proches de la corrélation au plateau de la figure 36. L'estimation avec la régression semble être la plus stable : la chute de corrélation pour l'estimateur de Malliavin Mancino est nettement moins marquée.

Notons également la baisse de corrélation du 2 Juin pour Hayashi-Yoshida. Celle-ci provient sans doute d'une fréquence de trading trop élevée, ce qui fait que l'estimateur ne parvient pas à retrouver la vraie corrélation. La problématique est la même que celle de l'effet Epps : la corrélation mesurée chute lorsque le pas temporel devient petit. Une des explications possibles est que cet estimateur n'est pas robuste aux bruits de micro-structure.

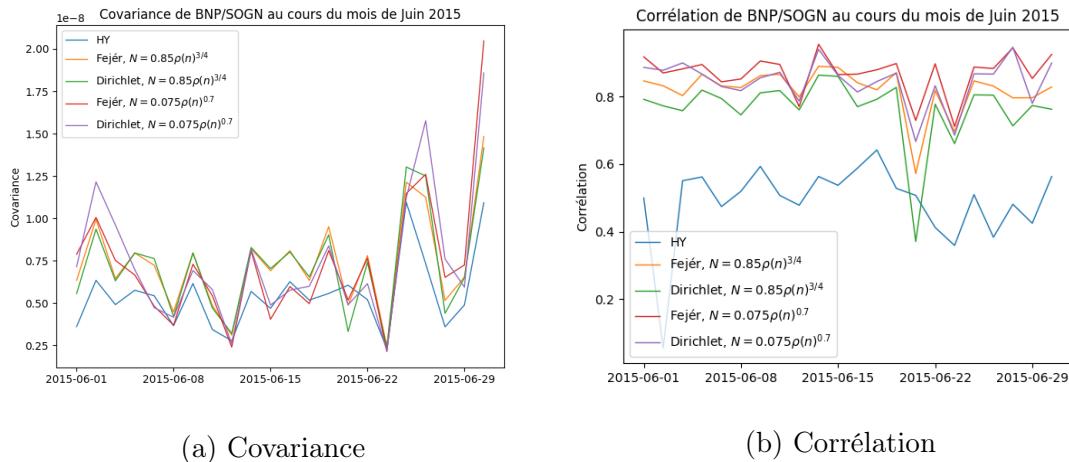


FIGURE 40 – Covariances et corrélations de la paire BNP/SOGN sur le mois de Juin 2015

Le temps de calcul pour Hayashi-Yoshida étant très long, nous avons décidé de tracer les matrices de corrélation uniquement avec l'estimateur de Malliavin-Mancino. Nous avons d'abord tracé une version en prenant la moyenne des corrélations sur le mois 41, puis une autre en concaténant nos données 42. La version avec  $N = 0.85 \min(n_1, n_2)^{3/4}$  est présentée ici, celle avec  $N = 0.075 \min(n_1, n_2)^{0.7}$  peut être retrouvée en annexe 9.4.2. De façon assez surprenante, certains résultats ne sont pas du tout les mêmes. En effet, les corrélations moyennes sont presque toutes inférieures aux corrélations obtenues avec concaténation des données. On pourrait peut être expliquer ce phénomène par le fait que les données concaténées présentent nécessairement des "sauts". Ceux-ci entraînent alors un biais positif de

## 7.4 Etude des données quotidiennes

corrélation. Une solution serait d'utiliser l'estimateur de Cuchiero-Teichmann, qui est une variante de celui de Malliavin-Mancino et est plus robuste aux sauts selon [5].

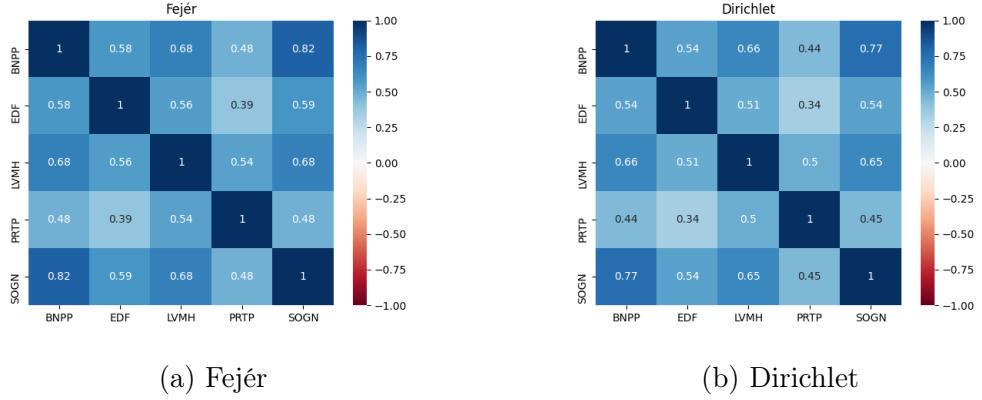


FIGURE 41 – Matrice de corrélations moyenne sur Juin 2015

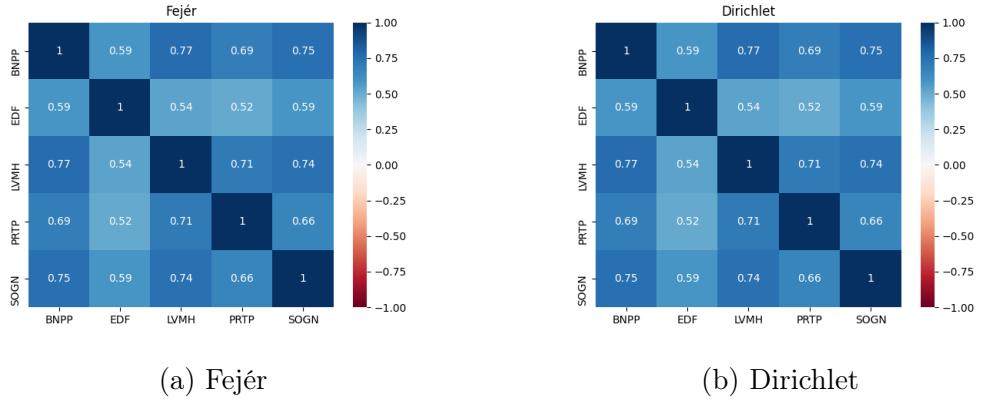


FIGURE 42 – Matrice de corrélations des données concaténées de Juin 2015

L'équation (\*) propose une relation entre un "pas" de de rééchantillonnage et  $N$ . Nous nous sommes alors intéressés à l'influence du temps  $\Delta t$  sur les estimations de corrélations, en le faisant varier dans  $\{60\text{s}, 100\text{s}, 150\text{s}, 200\text{s}\}$ . Nous choisissons de travailler avec des données concaténées, car il s'agit de la démarche adoptée dans [7]. Les résultats sont tracés sur la figure 43, ceux obtenus par inversion de Dirichlet sont en annexe 52. Sans grande surprise, les corrélations mesurées sont toutes conformes à l'effet Epps : à mesure que le pas temporel augmente, les corrélations mesurées augmentent.

## 7.4 Etude des données quotidiennes

Nous nous intéressons maintenant à l'étude des données financières en intraday. Nous reprenons la démarche de l'article [19]. Nous subdivisons nos données

## 7.4 Etude des données quotidiennes

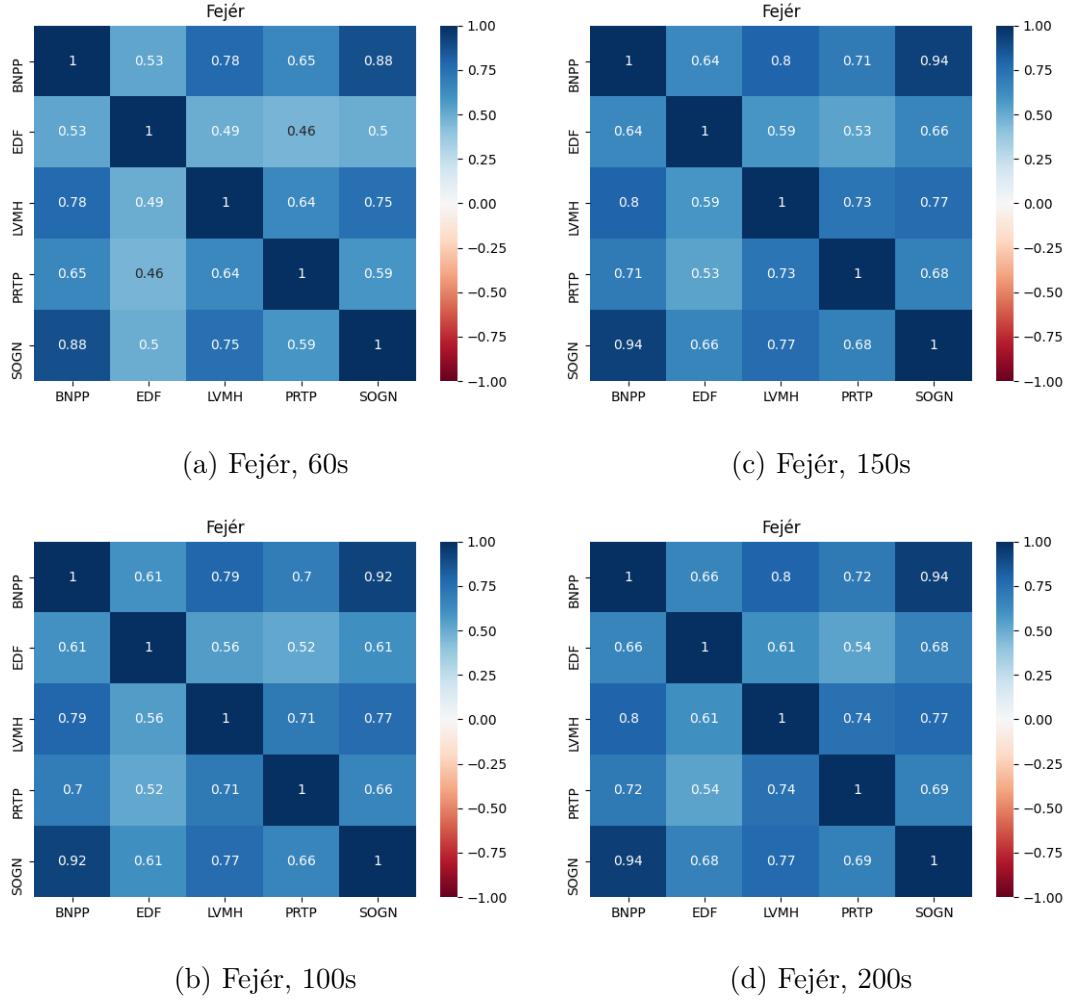


FIGURE 43 – Matrices de corrélations pour différents pas temporels

en intervalles de 5 min puis nous appliquons les différents estimateurs. Ce calcul est effectué sur tout le mois, puis nous prenons les différents quartiles pour chaque tranche de 5 min, notamment en raison de la distribution non gaussienne. Pour le calcul avec l'estimateur de Malliavin-Mancino, nous avons choisi de prendre  $N = 0.85 \min(n_1, n_2)^{3/4}$  avec leur recommandation ; le nombre de points était trop faible pour utiliser le  $N$  obtenu par régression. Nous observons alors sur la figure 44 que la covariance prend une forme en "U", phénomène déjà décrit dans [11] et retrouvé expérimentalement dans [19]. Cependant, la corrélation mesurée ne présente pas de tendances particulières. Dans les courbes obtenues, il apparaît clairement que la mesure obtenue avec inversion de Fejér conduit à des corrélations plus élevées. Cela peut s'expliquer par le fait que cette inversion, bien que plus robuste dans la mesure où elle conduit à des résultats moins dispersés n'est pas plus précise pour autant, comme nous l'avions remarqué dans les sections précédentes. Remarquons enfin que la mesure de la corrélation ici par Hayashi-Yoshida est restée cohérente avec celle sur toute la journée ; les valeurs sont toutes voisines de 0.5 alors que celles

## 7.4 Etude des données quotidiennes

---

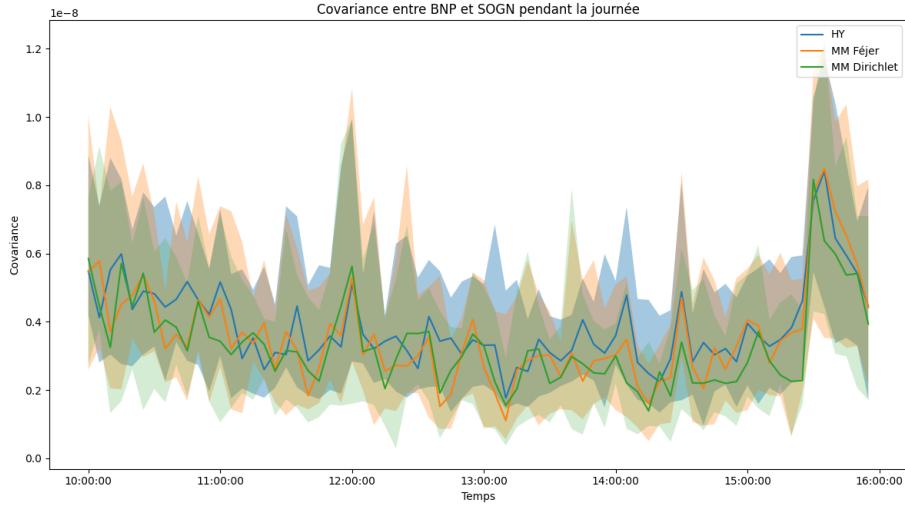


FIGURE 44 – Volatilité médiane et écart interquartile sur le mois de Juin 2015

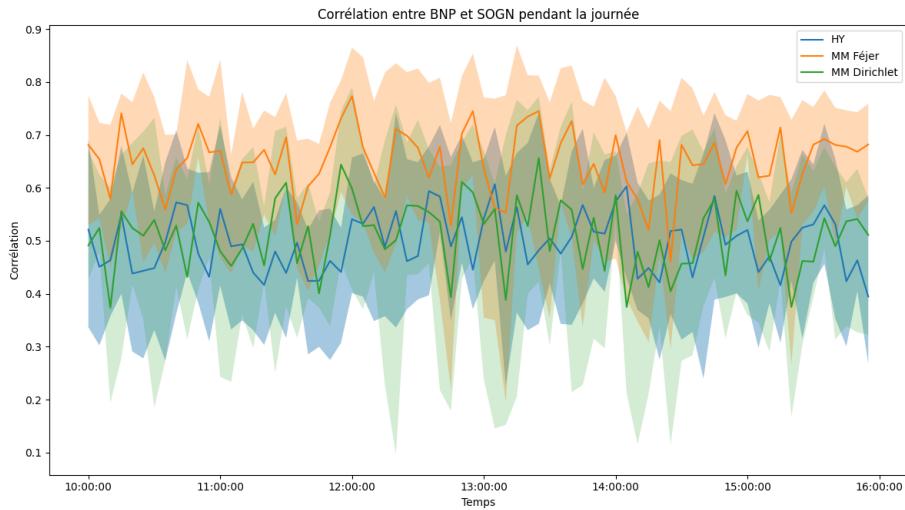


FIGURE 45 – Corrélation médiane et écart interquartile sur le mois de Juin 2015

par Malliavin-Mancino ont été fortement modifiées.

Nous souhaitons également confronter cela aux résultats que nous obtenons par estimation instantanée de la covariance et de la corrélation au niveau de chaque jour du mois. Nous traçons ensuite les valeurs médianes pour chaque instant. On commence par prendre pour les paramètres  $M$  et  $N$ , les paramètres mentionnés dans l'étude [5]. A savoir :

$$\begin{cases} N^* &= \lfloor \frac{n}{2} \rfloor \\ M^* &= \lfloor \frac{1}{16\pi} \sqrt{n} \log n \rfloor \end{cases}$$

## 7.4 Etude des données quotidiennes

---

Nous obtenons les résultats suivants :

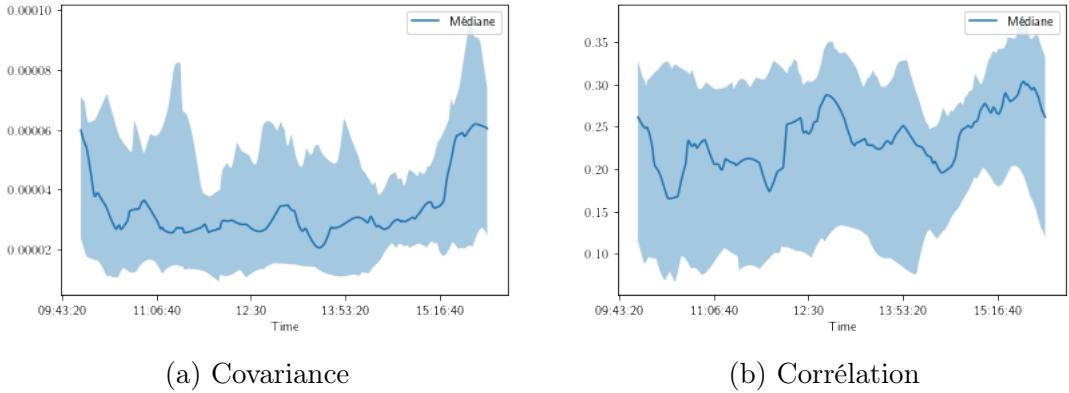


FIGURE 46 – Résultats d'estimation pour  $N^*$  et  $M^*$

La forme en 'U' de la covariance est encore une fois observable. Elle est possiblement attribuable à tendance décrite par [16] de l'estimateur à sur-estimer la covariance au niveau de l'extrémité des intervalles de temps. La corrélation est toutefois sensiblement différente de celle estimée précédemment. Cependant, étant donné la nature de la paire BNPP et SOGN, des fortes similitudes entre ces deux actions et des valeurs de corrélations retrouvées dans la littérature, nous doutons que la corrélation ici soit exacte et pensons qu'elle est sous-estimée.

Nous recommençons donc l'expérience avec de nouvelles valeurs pour les paramètres, cette fois-ci avec les valeurs conseillées par [16] dans le cas asynchrone :

$$\begin{cases} N^{**} = \lfloor 0.85n^{3/4} \rfloor \\ M^{**} = \lfloor \frac{1}{16\pi} \sqrt{n^{3/4}} \log n^{3/4} \rfloor \end{cases}$$

et nous obtenons les courbes suivantes :

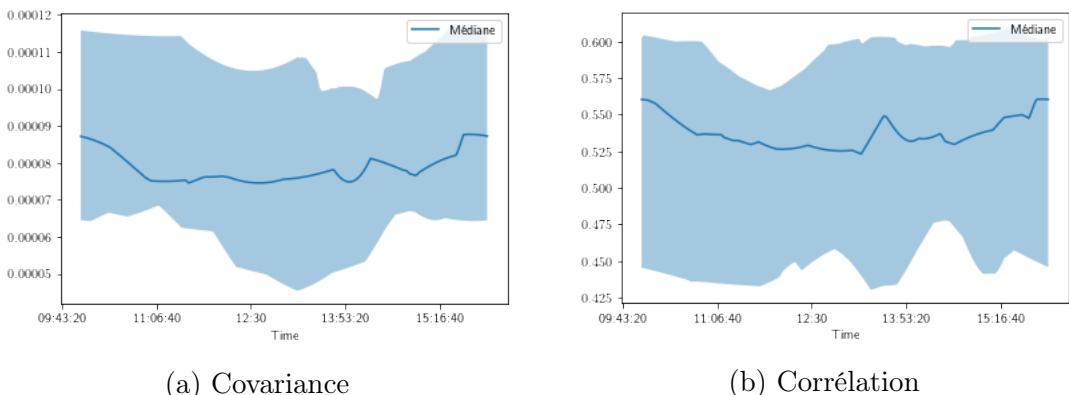
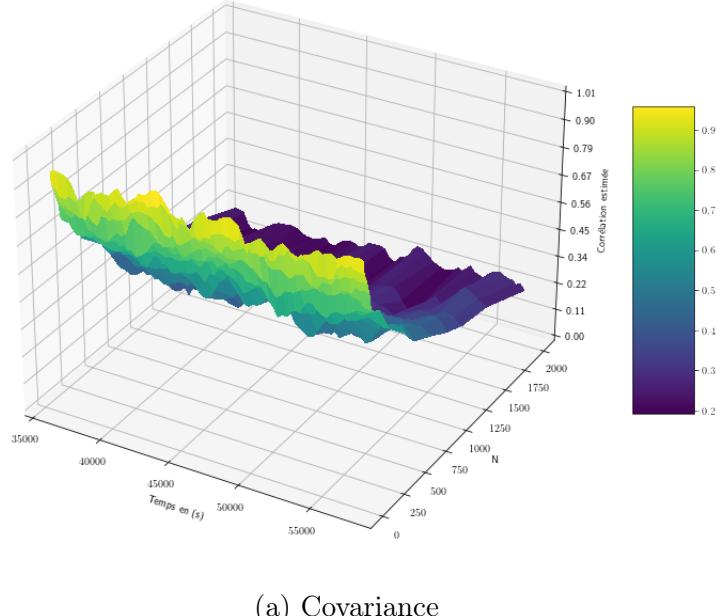


FIGURE 47 – Résultats d'estimation pour  $N^{**}$  et  $M^{**}$

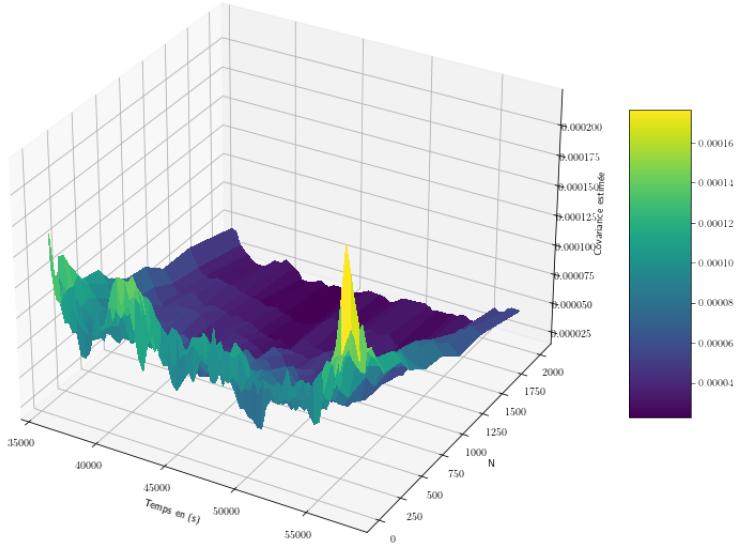
Les résultats sont encore une fois très différents, la corrélation ainsi que la variance ont une apparence quasi constante et la corrélation se rapproche beaucoup

## 7.4 Etude des données quotidiennes

plus en ordre de grandeur des estimations précédentes avec Malliavin-Mancino intégré et Hayashi-Yoshida. N'ayant pas de méthode précise pour déterminer à quel paramètres  $N$  et  $M$  s'en tenir pour avoir une estimation fiable, nous procérons à l'image de [5] : nous fixons  $M = 10$  arbitrairement et à faisons varier  $N$  entre 10 et 2500. Nous obtenons ainsi :



(a) Covariance



(b) Corrélation

FIGURE 48 – Evolution de la covariance et de la corrélation instantanées en fonction de  $N$  à  $M = 10$  fixé

On remarque immédiatement que l'estimation dépend fortement du niveau du paramètre  $N$ . En outre, on observe que plus  $N$  est grand, plus la corrélation et la covariance estimées diminuent. En se remémorant la relation  $*$ , on peut en déduire que plus le pas d'échantillonnage est grand, plus la corrélation et la covariance augmentent. On observe donc directement l'effet Epps aux niveau de l'estimation instantanée de la corrélation et la covariance de données réelles. L'article [5] observe également le même effet, et détermine les fréquences de coupures à retenir grâce à une comparaison avec l'estimateur de Cuchiero-Teichmann.

---

## 8 Conclusion et ouverture

Dans cette étude, nous avons mis en évidence quelques propriétés de l'estimateur de Malliavin-Mancino et mis en place une méthode de calcul rapide des coefficients de Fourier qui permet de réduire le temps d'exécution nécessaire, sans pour autant nuire à la qualité de l'estimation. Nous avons comparé ses performances avec l'estimateur de Hayashi-Yoshida théoriquement et sur des données réelles. Il apparaît que sous réserve d'un choix judicieux de  $N$ , l'estimateur de Malliavin-Mancino est meilleur que celui de Hayashi-Yoshida aussi bien en termes de temps de calcul qu'en précision, et résiste mieux aux bruits de micro-structure.

Cependant ces résultats sont à relativiser. Tout d'abord, les performances théoriques ont été faites avec des simulations rudimentaires (Mouvement brownien géométrique et modèle de Heston) non bruitées. Il serait intéressant de comparer les estimations dans des conditions plus variées et de diversifier les estimateurs (estimateur du maximum de vraisemblance, estimateur de Cuchiero-Teichmann...).

Enfin, la précision de l'estimateur de Malliavin Mancino est fortement conditionnée par le choix de  $N$ . A notre connaissance, hormis quelques cas particuliers bien spécifiques, il n'y a pas de formule théorique permettant de choisir  $N$ , ni de recommandation dans le cas général, la valeur proposée par Malliavin et Mancino était obtenue pour deux actifs désynchronisés d'un pas constant ; nous avons proposé  $N = 0.075 \min(n_1, n_2)^{0.7}$  par régression linéaire. Une des pistes de travail serait d'affiner cette recommandation à des cas plus généraux, et tester de manière plus extensive. Une autre serait d'essayer de trouver une méthode théorique pour fixer  $N$ .

Rappelons tout de même que le principal avantage de l'estimateur de Malliavin-Mancino est sa capacité à estimer directement la covariance instantanée de façon numériquement stable, ce que ne permet pas l'estimateur de Hayashi-Yoshida. Sur cet aspect, la précision est également conditionnée par deux paramètres  $N$  et  $M$  qu'il est nécessaire de choisir. Le travail est alors similaire à celui du paragraphe précédent, mais bien plus conséquent puisque nous avons deux paramètres.

---

## 9 Annexes

### 9.1 Implémentation numérique du mouvement brownien géométrique

Nous commençons par générer un échantillon de données tests de vecteurs prix  $S(t) = (S_1(t), \dots, S_d(t))$  les équations qui définissent un mouvement brownien géométrique :

$$\frac{dS_i(t)}{S_i(t)} = \mu_i dt + \sigma_i dW_i(t), \quad i = 1, \dots, d,$$

où  $W_i$  est un mouvement brownien standard et  $\text{Corr}(W_i, W_j) = \rho_{ij}$ .

Nous nous sommes basés sur cet algorithme donné en [12] pour l'implémentation numérique :

- On définit une matrice  $\Sigma$  telle que  $\Sigma_{ij} = \sigma_i \sigma_j \rho_{ij}$ , par abus de notation, on dit que le vecteur  $S$  suit un mouvement brownien géométrique. **BGM**( $\mu, \Sigma$ ) où  $\mu = (\mu_1, \dots, \mu_d)$
- Soit  $A$  une matrice telle que  $AA^\top = \Sigma$ .
- On obtient  $\frac{dS_i(t)}{S_i(t)} = \mu_i dt + \sum_{j=1}^d A_{ij} dW_j(t)$ .
- Cette représentation permet d'obtenir un algorithme récursif simple de calcul des prix :  $S_i(t_{k+1}) = S_i(t_k) e^{(\mu_i - \frac{1}{2}\sigma_i^2)(t_{k+1}-t_k) + \sqrt{t_{k+1}-t_k} \sum_{j=1}^d A_{ij} Z_{k+1,j}}$  où  $Z_k = (Z_{k1}, \dots, Z_{kd}) \sim N(0, I)$  et  $Z_1, \dots, Z_d$  sont indépendantes.

**Remarque 10.** Si  $\Sigma$  est symétrique définie positive, il est possible de prendre la décomposition de Cholesky de  $\Sigma$  comme matrice  $A$ , cela permet notamment de réduire le nombre d'additions et de multiplications requis dans les calculs. On se placera toujours dans ce cas en identifiant chaque matrice symétrique à sa matrice symétrique définie positive la plus proche au sens de la norme de Frobenius.

### 9.2 Implémentation numérique de la NUFFT

Après avoir rencontré des problèmes d'implémentation de la transformée de Fourier non uniforme dûs à une confusion par rapport aux conventions de définition utilisées, nous avons pensé qu'il serait utile d'inclure en annexe la démarche à suivre.

Par souci de cohérence avec la formule de calcul des coefficients de Fourier, on définit le produit de convolution de deux fonctions  $f$  et  $g$   $T$ -périodiques comme suit :

$$h(x) := f * g(x) = \frac{1}{T} \int_0^T f(y)g(x-y) dy.$$

Avec la convention usuelle suivante :

$$F(k) = \frac{1}{T} \int_0^T f(t)e^{-ik\frac{2\pi}{T}t} dt,$$

Les coefficients de Fourier vérifient alors la propriété :

$$H(k) = F(k)G(k).$$

On suppose que l'on dispose d'un échantillon de  $n$  observations aux positions  $0 := x_0 < \dots < x_{n-1} := T$  d'une fonction. Par homothétie, on se ramène à une fonction  $p$ -périodique mesurée aux points  $\tilde{x}_0 := p\frac{x_0}{T} < \dots < \tilde{x}_{n-1} := p\frac{x_{n-1}}{T}$ . On note  $p$  la période du noyau  $\tilde{\varphi}$  utilisé pour la transformée de Fourier non uniforme. On a alors les expressions suivantes :

$$\begin{aligned} f(x) &:= \sum_{j=0}^{N-1} f_j \delta(x - \tilde{x}_j) \\ \tilde{\varphi}(x) &:= \sum_{r \in \mathbb{Z}} \varphi(x - rp). \end{aligned}$$

Ainsi :

$$\begin{aligned} f_\tau(x) &:= f * \tilde{\varphi}(x) \\ &= \frac{1}{p} \int_0^p f(y) \tilde{\varphi}(x - y) dy \\ &= \frac{1}{p} \sum_{j=0}^{n-1} f_j \tilde{\varphi}(x - \tilde{x}_j) \\ &= \frac{1}{p} \sum_{j=0}^{n-1} \sum_{r \in \mathbb{Z}} f_j \varphi(x - \tilde{x}_j - rp) \end{aligned}$$

Nous souhaitons évaluer cette fonction aux  $\xi_i := \frac{pi}{M_r}$ . Les noyaux utilisés pour la transformée de Fourier non uniforme étant généralement concentrés autour d'un pic, quitte à translater de  $p$ , les  $\xi_i$  significatifs sont donc autour de  $\xi_{i^*} \approx \tilde{x}_j$ .  
Ainsi, l'ordre du choix est :

$$\begin{aligned} j &: 0, \dots, n-1 \\ \implies i &: \left\lfloor \frac{M_r \tilde{x}_j}{p} \right\rfloor - M_{sp}, \dots, \left\lfloor \frac{M_r \tilde{x}_j}{p} \right\rfloor + M_{sp} \quad \text{mod } M_r. \end{aligned}$$

On peut donc restreindre le calcul à  $r = 0$ , grâce à la congruence et la périodicité du noyau.

Si on pose  $i := \left\lfloor \frac{M_r \tilde{x}_j}{p} \right\rfloor + k$ , il faut donc calculer le terme :

$$\varphi\left(\frac{p}{M_r} \left\lfloor \frac{M_r \tilde{x}_j}{p} \right\rfloor + \frac{kp}{M_r} - \tilde{x}_j\right).$$

Nous pouvons aussi remarquer que le noyau est symétrique, ce qui permet de simplifier l'implémentation.

Dans notre cas, nous souhaitons évaluer les coefficients de Fourier aux fréquences entières. Il s'agit de calculer après simplifications la somme suivante pour le coefficient

$k$  :

$$\sum_{j=0}^{n-1} f_j e^{-i2\pi k \frac{\tilde{x}_j}{p}}.$$

On pose (en supposant que  $f$  est  $p$ -périodique) :

$$f(x) = \sum_{j=0}^{n-1} f_j \delta(x - \tilde{x}_j).$$

On obtient donc, suivant la convention précédente :

$$F(k) = \frac{1}{p} \sum_{j=0}^{n-1} f_j e^{-i2\pi k \frac{\tilde{x}_j}{p}}.$$

Si la transformée de Fourier du noyau est calculée selon la même convention que précédemment (convention en  $\frac{1}{T}$ ), nous obtenons par périodicité de la fonction :

$$\Phi(k) = \frac{1}{p} \hat{\varphi}\left(\frac{k}{T}\right),$$

où  $\Phi(k)$  est le coefficient de Fourier de mode  $k$  de  $\phi$ . Enfin, il suffira alors d'effectuer la démarche détaillée précédemment sans oublier la multiplication par le facteur  $p$  à la fin.

### 9.3 Implémentation numérique de l'estimateur de Hayashi Yoshida

Par souci de rapidité, il est intéressant de calculer l'estimateur avec des opérations vectorisées à l'image de [6]. On définit donc :

$$\delta_i = \begin{bmatrix} \ln(P^i(t_2^i)) - \ln(P^i(t_1^i)) \\ \vdots \\ \ln(P^i(t_{n_i}^i)) - \ln(P^i(t_{n_i-1}^i)) \end{bmatrix} \quad \text{et} \quad W = \begin{bmatrix} \mathbb{1}_{]t_1^i, t_2^i] \cap ]t_1^j, t_2^j] \neq \emptyset} & \cdots & \mathbb{1}_{]t_1^i, t_2^i] \cap ]t_{n_j-1}^j, t_{n_j}^j] \neq \emptyset} \\ \vdots & \ddots & \vdots \\ \mathbb{1}_{]t_{n_i-1}^i, t_{n_i}^i] \cap ]t_1^j, t_2^j] \neq \emptyset} & \cdots & \mathbb{1}_{]t_{n_i-1}^i, t_{n_i}^i] \cap ]t_{n_j-1}^j, t_{n_j}^j] \neq \emptyset} \end{bmatrix}.$$

$W$  est appelée matrice de Kanatagi. Avec ces expressions, on obtient :

$$\Sigma_{ij} = {}^t \delta_i W \delta_j.$$

## 9.4 Figures supplémentaires

### 9.4.1 Corrections des estimations de corrélations pour l'inversion de Dirichlet

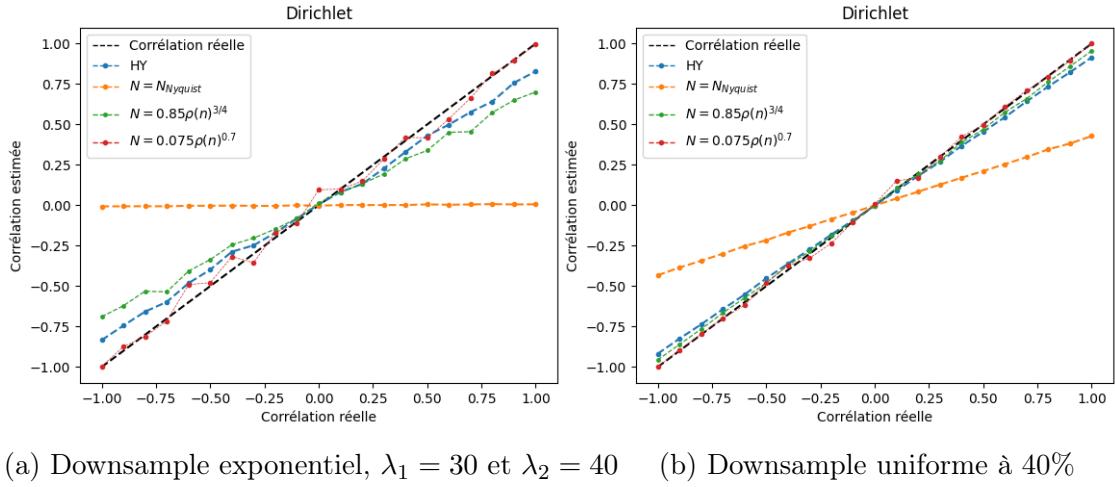


FIGURE 49 – Comparaison d'estimations de corrélations pour plusieurs fréquences de coupures

### 9.4.2 Matrices de corrélations pour $N = 0.075 \min(n_1, n_2)^{0.7}$

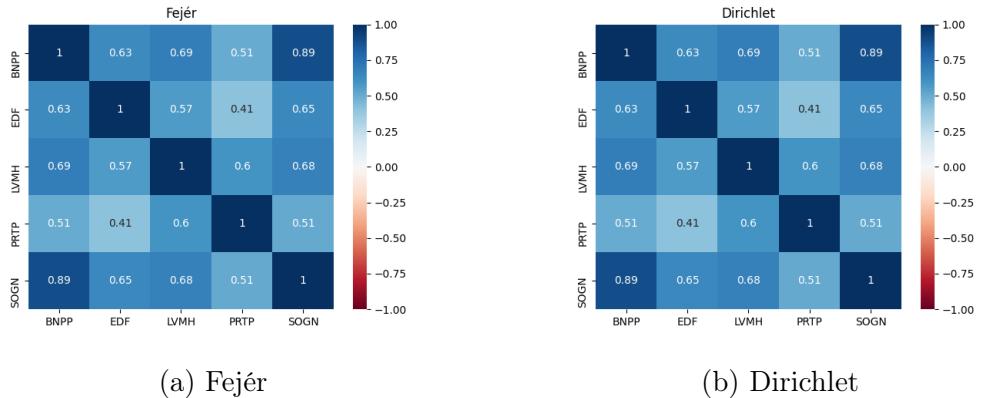


FIGURE 50 – Matrice de corrélations moyenne sur Juin 2015

## 9.4 Figures supplémentaires

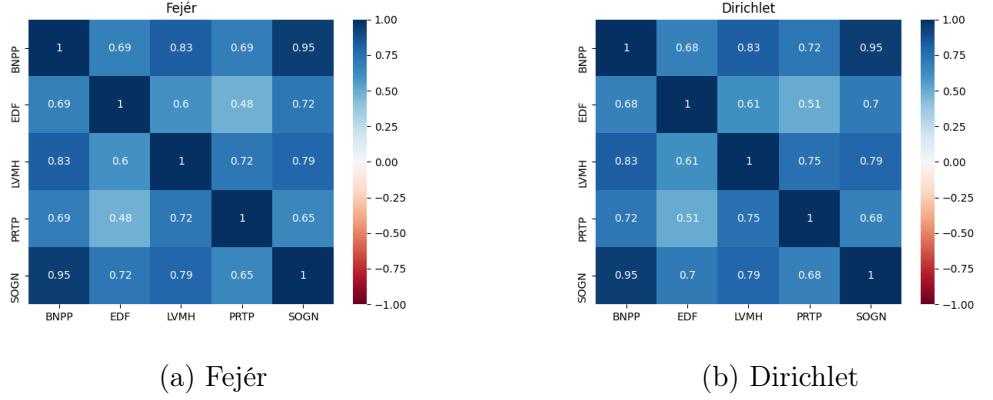


FIGURE 51 – Matrice de corrélations des données concaténées de Juin 2015

### 9.4.3 Matrices de corrélations pour l'inversion de Dirichlet

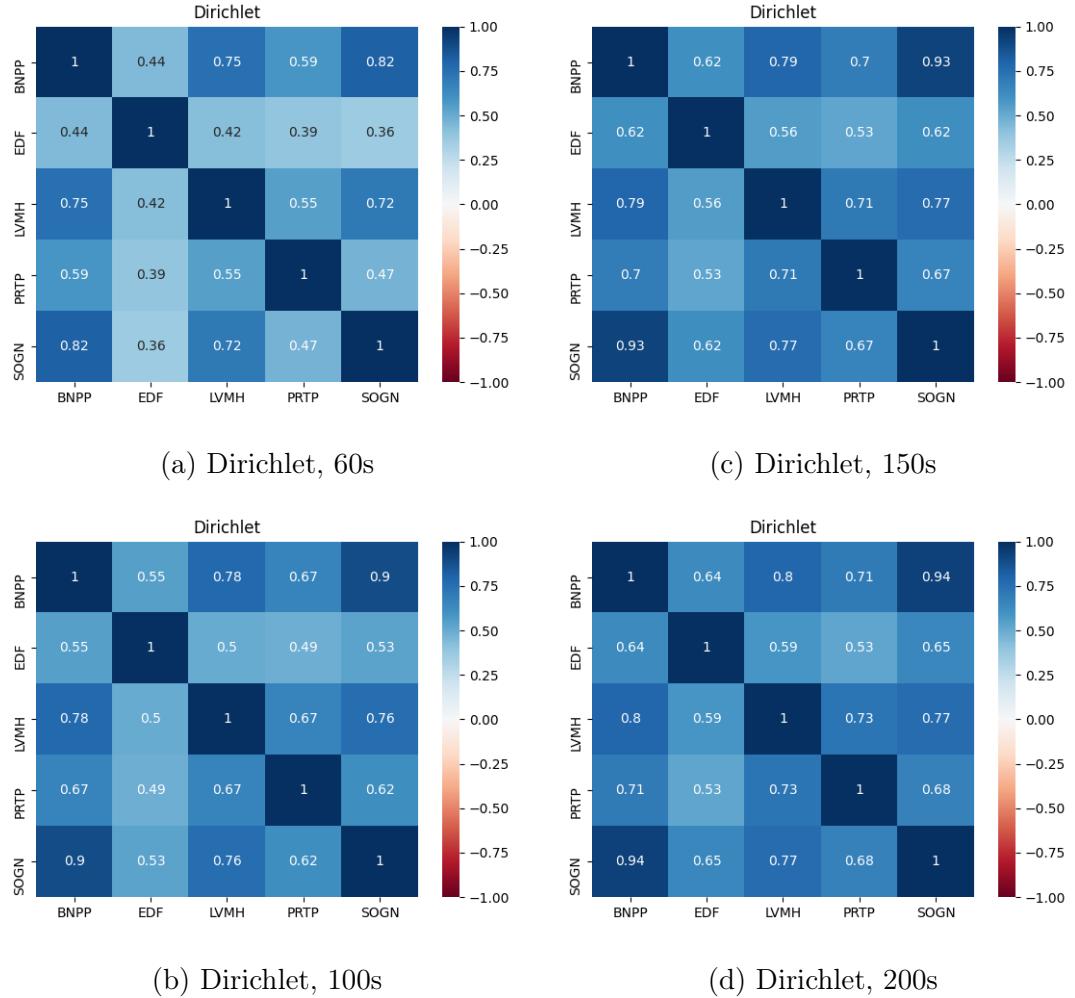


FIGURE 52 – Matrices de corrélations pour différents pas temporels

## Références

- [1] Yacine Aït-Sahalia, Per A. Mykland, and Lan Zhang. How often to sample a continuous-time process in the presence of market microstructure noise. *The Review of Financial Studies*, 18 :351–416, 2005.
- [2] H. Alexander Barnett, Jeremy Magland, and Ludvig Af Klinteberg. A parallel nonuniform fast fourier transform library based on an exponential of semicircle kernel. 2019.
- [3] L. Bergomi. Correlations in asynchronous markets. *Risk*, pages 76–82, 2010.
- [4] A. Chakraborti, I. Muni Toke, M. Patriarca, and Abergel F. Econophysics review : I. empirical facts. *Quantitative Finance*, 11(7) :991–1012, 2011.
- [5] Patrick Chang. Fourier instantaneous estimators and the epps effect. 2020.
- [6] Patrick Chang and Roger Bukuru. An exercise in R : High frequency covariance estimation using malliavin-mancino and hayashi-yoshida estimators. 2019.
- [7] Patrick Chang, Etienne Pienaar, and Tim Gebbie. Malliavin-mancino estimators implemented with non-uniform fast fourier transforms. 2020.
- [8] Emmanuelle Clement and Arnaud Gloter. Weak limit theorem in the Fourier tranform method for the estimation of multivariate volatility. *Stochastic Processes and their Applications*, 121 :1097–1124, 2011.
- [9] A. Dutt and V. Rokhlin. Fast fourier transforms for nonequispaced data. *Applied and Computational Harmonic Analysis*, 2 :85–100, 1995.
- [10] T. W. Epps. Comovements in stock prices in the very short run. *Journal of American Statistics Association*, 74 :291–298, 1979.
- [11] Torben G. Andersen and Tim Bollerslev. Intraday periodicity and volatility persistence in financial markets. *Empirical Finance*, 4 :115–158, 1997.
- [12] Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, 2003.
- [13] Leslie Greengard and June-Yub Lee. Accelerating the nonuniform fast fourier transform. *SIAM review*, 46 :443–454, 2004.
- [14] T. Hayashi and N. Yoshida. On covariance estimation of nin-synchronously observed diffusion processes. *Bernoulli*, 11(2) :359–379, 2005.
- [15] Paul Malliavin and Maria Elvira Mancino. A fourier transform method for non parametric estimation of multivariate volatility. *The Annals of Statistics*, 37(4) :4–10, 2009.
- [16] Maria Elvira Mancino, Maria Cristina Recchioni, and Simona Sanfelici. *Fourier-Malliavin Volatility Estimation : Theory and Practice*. SpringerBriefs in Quantitative Finance. Springer International Publishing, 2017.
- [17] Vanessa Mattiussi, Michele Tumminello, Giulia Iori, and Rosario N. Mantegna. Comparing correlation matrix estimators via kullback-leibler divergence. 2011.
- [18] Roberto Renò. A closer look at the epps effect. *International Journal of Theoretical and Applied Finance*, 6(1) :87–102, 2003.

## RÉFÉRENCES

---

- [19] Gayatri Tilak, Tamás Szell, Rémy Chicheportiche, and Anirban Chakraborti. Study of statistical correlations in intraday and daily financial return time series. *Econophys-Kolkata*, IV :77, 2011.