Matrices de corrélation de grande dimension en finance: estimation, régularisation, applications

Cercles dans un cercle, Vassily Kandinsky, 1923

Hugo Vigna

23 février 2026

Sommaire

Chapitre 1

Introduction

La diversification de portefeuille nécessite la connaissance des corrélations entre les actifs du portefeuille. Cette matrice de covariance est en réalité difficile à estimer à l'aide des trajectoires de prix. En effet, la matrice de covariance empirique n'est sans biais qu'asymptotiquement, c'est-à-dire lorsque le nombre d'observations T est très grand devant le nombre de variables N. Ce régime asymptotique classique ($T \gg N$) n'est pas atteignable en pratique en finance. En effet, on atteint régulièrement plusieurs milliers d'actifs avec uniquement un an de données ($T \approx 250$ observations journalières). La raison de ce manque de données tient au fait que la structure de la covariance entre actifs ne peut pas être considérée constante au-delà d'un an.

Soit $X \in \mathbb{R}^{N \times T}$ la matrice des rendements centrés. La matrice de covariance empirique s'écrit :

$$\hat{\Sigma} = \frac{1}{T} X X^\top$$

Lorsque N et T sont de même ordre de grandeur, le ratio

$$q = \frac{N}{T}$$

reste fini et non nul, ce qui correspond au régime asymptotique de grande dimension.

Les travaux de *Marchenko* et *Pastur* ont mis en évidence les biais observables dans le spectre de la matrice de covariance empirique dans ce régime.

Le problème se pose donc du développement d'estimateurs améliorés de la matrice de covariance. De nombreux travaux sont dirigés en ce sens depuis une vingtaine d'années. Ce projet a pour but de comparer ces estimateurs sur des données synthétiques et réelles selon des métriques matricielles et financières, afin d'évaluer leurs performances, leur facilité d'implémentation et leur rapidité.

Les estimateurs considérés sont :

— La méthode empirique
— La méthode de clipping de Marchenko–Pastur utilisant la borne supérieure classique λ_+
— La méthode de clipping de Marchenko–Pastur utilisant comme borne supérieure le quantile 99% des valeurs propres
— La méthode de shrinkage linéaire de Ledoit–Wolf
— La méthode de shrinkage non linéaire de Ledoit–Wolf
— La méthode de l'oracle MV (uniquement pour les données synthétiques)
— La méthode de l'oracle moyen de Bongiorno et Challet

L'évaluation porte sur les métriques suivantes :

— La norme de Frobenius entre la matrice estimée et la matrice réelle (uniquement pour les données synthétiques)
— La perte MV entre la matrice estimée et la matrice réelle (uniquement pour les données synthétiques)
— La distance de Kolmogorov–Smirnov entre les distributions des rendements générés avec la vraie matrice et ceux générés avec la matrice estimée
— Le taux de rejet du test de Kolmogorov–Smirnov
— La frontière d'efficience de Markowitz obtenue avec la matrice estimée

Nous verrons que ces métriques mesurent principalement les effets suivants :

Métrique	**Effet principal mesuré**
Norme de Frobenius	Erreur surtout sur les grandes valeurs propres
Perte MV	Erreur surtout sur les petites valeurs propres
Distance KS	Modification globale de la distribution des rendements
Taux de rejet KS	Différences statistiquement significatives entre distributions
Frontière d'efficience	Erreur combinée sur grandes et petites valeurs propres

La frontière d'efficience constitue ainsi la métrique financière privilégiée, car elle intègre l'effet global des erreurs spectrales sur l'allocation optimale. Elle est également plus proche de l'objectif final, qui est l'optimisation d'un portefeuille.

Chapitre 2

Théorie des matrices aléatoires

L'idée principale est d'utiliser la théorie des matrices aléatoires comme référence de bruit, en supposant que les rendements des actifs sont indépendants. Cette hypothèse fournit une description théorique précise du spectre des valeurs propres que l'on devrait observer s'il n'existait aucune structure économique réelle dans les données. En comparant le spectre empirique à ce spectre théorique, on peut alors identifier quelles valeurs propres relèvent du bruit et lesquelles traduisent une information réelle.

Théorème (Marchenko–Pastur). Soit $X \in \mathbb{R}^{N\times T}$ une matrice dont les coefficients sont i.i.d. de moyenne nulle et variance σ^2. Lorsque $N, T \to \infty$ avec $c = N/T$ fixé, la distribution empirique des valeurs propres de $\hat{\Sigma} = \frac{1}{T}XX^\top$ converge vers la loi de Marchenko–Pastur de densité

$$\rho_{MP}(\lambda) = \frac{1}{2\pi c\sigma^2\lambda}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}\mathbf{1}_{[\lambda_-,\lambda_+]}(\lambda),$$

avec

$$\lambda_\pm = \sigma^2(1 \pm \sqrt{c})^2.$$

Le théorème s'applique ainsi directement au spectre de la matrice de covariance empirique.

Il fournit une borne théorique du spectre dans le cas où les rendements seraient totalement indépendants et ne contiendraient aucun facteur commun : les valeurs propres devraient alors rester dans l'intervalle $[\lambda_-, \lambda_+]$. La figure suivante illustre le spectre de la matrice de covariance empirique pour p variables aléatoires *iid* normales. La matrice de covariance obtenue est appelée matrice de Wishart.

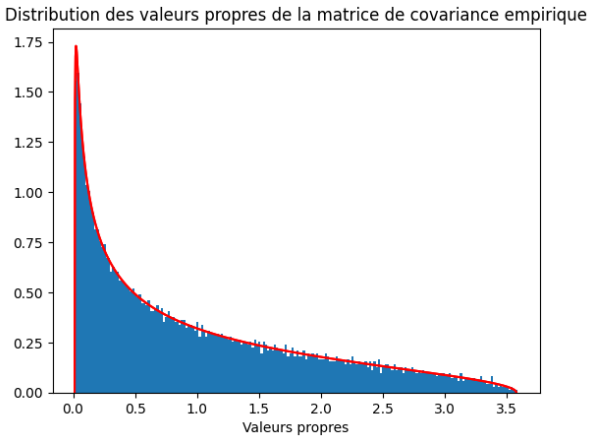

FIGURE 2.1 – Densité du spectre de $\Sigma^{Wishart}$: bulk uniquement et absence de spikes

En réalité, les données financières présentent des dépendances structurelles, et certaines valeurs propres s'étalent au delà de cet intervalle ; ces dépassements sont interprétés comme la présence de facteurs économiques réels.

On obtient donc une structure de la forme *bulk + spikes* : un bloc de valeurs propres faiblement informatives que l'on peut associer à du bruit (bulk de Marchenko–Pastur), et des valeurs propres plus grandes et informatives qui caractérisent la structure véritable de la covariance On obtient ainsi une décomposition conceptuelle de type *bulk + spikes* :

$$\underbrace{\lambda_- \quad \longleftrightarrow \quad \lambda_+}_{\text{Bulk bruit}} \qquad\qquad \underbrace{\text{Spikes}}_{\text{Valeurs propres informatives}}$$

le bulk correspond au bloc central de valeurs propres prédit par Marchenko–Pastur (bruit), tandis que les spikes correspondent aux valeurs propres isolées, interprétées comme porteuses d'information.

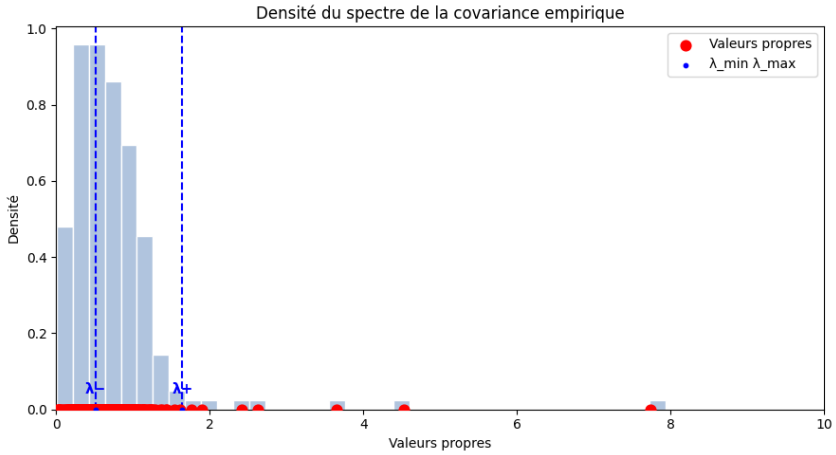

FIGURE 2.2 – Densité du spectre de $\hat{\Sigma}$

En pratique, le nombre d'actifs N est du même ordre que le nombre d'observations T, voire supérieur. Lorsque $N > T$, la matrice de covariance empirique $\hat{\Sigma}$ est de rang au plus T, et possède donc $N - T$ valeurs propres nulles.

Dans ce contexte, les valeurs propres situées dans l'intervalle inférieures à λ_+ sont considérées comme non informatives. Une stratégie simple de nettoyage consiste alors à remplacer toutes ces valeurs propres par leur moyenne :

$$\lambda_i^{\text{clean}} = \begin{cases} \lambda_i & \text{si } \lambda_i > \lambda_+, \\ \bar{\lambda}_{\text{bulk}} & \text{sinon.} \end{cases}$$

On conserve ainsi les *spikes* informatifs tout en régularisant les directions dominées par le bruit.

Dans ce cadre, une classe d'estimateurs en particulier émerge : les estimateurs invariants par rotation (RIE). Ces estimateurs reposent sur une transformation des valeurs propres de la matrice de covariance empirique, les vecteurs propres restant inchangés. L'intuition qui sous-tend cette méthode consiste à considérer que ne pas faire d'hypothèse sur les vecteurs propres est équivalent à prendre un cas général, ou de "maximum d'entropie". Il y a donc une isotropie a priori pour les vecteurs propres, et le travail n'a lieu que sur les valeurs propres.

Chapitre 3

Linear shrinkage estimator

Parmi les estimateurs envisagés pour le nettoyage de matrices de covariance se trouvent les estimateurs linéaires de la classe des esimateurs invariants par rotation (RIE). Nous avons étudié l'estimateur linéaire développé par Ledoit et Wolf.

On note la matrice de covariance empirique $S = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$.

On se place dans le régime asymptotique haute dimension $p, n \to \infty \quad \text{avec} \quad \frac{p}{n} \to q \in (0, +\infty)$.

Les hypothèses sont les suivantes :

— Il existe $K \in \mathbb{R}$ tel que pour tout n :

$$\frac{p_n}{n} < K$$

— Les variables (X_i) sont i.i.d. et centrées, i.e. $\mathbb{E}[X_i] = 0$.

— On décompose la matrice de covariance S_n en valeurs propres et vecteurs propres :

$$S_n = G_n \Lambda_n G_n^\top$$

où Λ_n est une matrice diagonale dont les éléments diagonaux sont les valeurs propres $\lambda_1^n, \ldots, \lambda_{p_n}^n$, et G_n est une matrice de rotation dont les colonnes sont les vecteurs propres $g_1^n, \ldots, g_{p_n}^n$. On définit alors :

$$Y_n = G_n^\top X_n$$

de sorte que les composantes de Y_n sont notées $Y_{i,n}$ pour $i = 1, \ldots, p_n$.

— Il existe $K_2 \in \mathbb{R}$ tel que pour tout n :

$$\frac{1}{p_n}\sum_{i=1}^{p_n} \mathbb{E}\big[Y_{i,n}^8\big] < K_2$$

—

$$\lim_{n\to\infty} \frac{p_n^2}{n^2} \cdot \frac{1}{|Q_n|} \sum_{(i,j,k,l)\in Q_n} \mathrm{Cov}(Y_{i,n}Y_{j,n},\, Y_{k,n}Y_{l,n})^2 = 0$$

où Q_n désigne l'ensemble des 4-uplets d'entiers distincts de $\{1, \ldots, p_n\}$.

Les estimateurs considérés sont invariants par rotation, c'est-à-dire qu'ils conservent les vecteurs propres empiriques et modifient uniquement les valeurs propres.

L'estimateur a pour expression :

$$S_n = \frac{b_n^2}{d_n^2}\, m_n I_n + \frac{a_n^2}{d_n^2}\, S_n$$

$$\begin{cases} m_n = \frac{1}{n}\,\mathrm{Tr}(S_n) \\ d_n^2 = \|S_n - m_n I_n\|^2 \\ \tilde{b}_n^2 = \frac{1}{n^2}\sum_{k=1}^{n} \left\|x_k x_k^\top - S_n\right\|^2 \\ d_n^2 = b_n^2 = \min\big(\tilde{b}_n^2,\, d_n^2\big) \\ d_n^2 = b_n^2 = \min\big(\tilde{b}_n^2,\, d_n^2\big) \end{cases}$$

m_n est la moyenne des valeurs propres empiriques, et $\alpha^* \in [0, 1]$ est l'intensité de shrinkage optimale (au sens de la norme de Frobenius). Cette intensité s'écrit asymptotiquement $\alpha^* = \frac{b_n^2}{d_n^2}$.

Sous ces hypothèses, l'estimateur LS (pour Linear Shrinkage) est optimal pour la norme de Frobenius parmi les estimateurs linéaires à coefficients constants ou aléatoires. Un premier avantage de cet estimateur est sa simplicité et rapidité d'impléméntation et de calcul. Par ailleurs, il est possible d'interpréter de multiple manières cet estimateur (d'un point de vue géométrique, bayésien...). Ces développements sont présents dans le papier et contribuent à une bonne compréhension globale du rôle de l'estimateur.

On peut par exemple remarquer que la correction que l'estimateur apporte à la matrice de covariance empirique augmente dans des proportions similaires au gain qu'il apporte relativement à la matrice empirique, pour la norme de Frobenius.

C'est-à-dire :

$$\text{Gain relatif} = \frac{||\hat{\Sigma}^{\text{empirique}} - \Sigma^{\text{true}}||_F - ||\hat{\Sigma}^{\text{LS}} - \Sigma^{\text{true}}||_F}{||\hat{\Sigma}^{\text{empirique}} - \Sigma^{\text{true}}||_F}$$

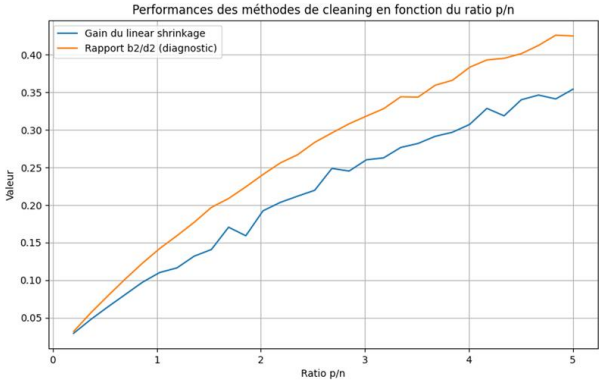

FIGURE 3.1 – Gain et correction de l'estimateur linéaire par rapport à la matrice de covariance empirique, en fonction du ratio

L'effet de l'estimateur sur les valeurs propres est le centrage autour de la moyenne des valeurs propres. L'estimateur linéaire a donc un effet global qui tend à rapprocher les spikes du bulk, limitant ainsi la dispertion des valeurs propres autour de leur moyenne.

L'estimateur linéaire est donc à la fois simple et rapide. Néanmoins il est optimal pour la norme de Frobenius, dont la pertinence est très discutable dans le contexte de matrices de covariances en finance. En effet, la norme de Frobenius pénalise très lourdement les grands écarts entre valeurs propres nettoyées et valeurs propres réelles, et très faiblement les petits écarts.

Considérons deux estimateurs candidats pour une matrice de covariance diagonale vraie :

$$\Sigma = \begin{pmatrix} 10 & 0 \\ 0 & 0.1 \end{pmatrix}.$$

Estimateur A (erreur forte sur grande valeur propre), estimateur B (erreur sur petite valeur propre) :

$$\widehat{\Sigma}_A = \begin{pmatrix} 7 & 0 \\ 0 & 0.1 \end{pmatrix}, \qquad \widehat{\Sigma}_B = \begin{pmatrix} 10 & 0 \\ 0 & 0.3 \end{pmatrix}.$$

Cet effet a pour conséquence que pour la matrice de covariance suivante, la norme de Frobenius pénalise davantage l'estimateur A que B.

Hors en finance travailler sur la matrice de covariance a souvent pour objectif d'estimer le risque d'allocation d'un portefeuille. Déterminer la frontière d'efficience d'un portefeuille fait intervenir l'inversion de la matrice de covariance. L'inversion a pour effet de faire exploser les erreurs sur les petites valeurs propres, et d'écraser celles sur les grandes. Ainsi, les deux estimateurs précédents auraient pour inverses :

$$\widehat{\Sigma}_A^{-1} = \begin{pmatrix} \frac{1}{7} & 0 \\ 0 & 10 \end{pmatrix}, \qquad \widehat{\Sigma}_B^{-1} = \begin{pmatrix} 0.1 & 0 \\ 0 & \frac{1}{0.3} \end{pmatrix}.$$

L'erreur sur la petite valeur propre est fortement amplifiée par l'inversion, ce qui dégrade davantage l'estimation du risque de portefeuille.

Utiliser un estimateur optimal pour la norme de Frobenius peut donc s'avérer inadapté.

Chapitre 4

Non linear shrinkage estimators and approximating oracle shrinkage

Des travaux récents de Ledoit et Wolf ont conduit à un estimateur RIE (*Rotationally Invariant Estimator*) de la matrice de covariance, qui repose sur une transformation non linéaire des valeurs propres de la matrice empirique.

Hypothèses

1. $c_n = \dfrac{p}{n}$ converge et $\in]0,1[$, et il existe un intervalle compact qui contient les c_n pour tout n.
2. $\Sigma_n \in \mathcal{S}_p^+(\mathbb{R})$ fixée.
3. (H_n) avec $H_n = \text{edf}\big(\text{ distribution empirique des } (\lambda_i)_i\big)$, $H_n \to H$ faiblement quand $n \to \infty$.
4. $\text{Supp}(H) = \bigcup_{k=1}^{K} I_k$, avec $\{I_k\}_k \subset]0,+\infty[^N$ fermés bornés .
5. Il existe un compact $[\underline{I}, \overline{I}] \subset]0,+\infty[$ qui contient les $\{\lambda_i\}$ empiriques pour tout n.
6. $Y_n = X_n\sqrt{\Sigma_n}$ avec $X_n \in \Pi_{n\times p}(\mathbb{R})$, i.i.d. centrées de variance 1.

Pour tout $x \in \text{Supp}(F)$ la fonctioon de répartition du spectre de Σ :

$$d^*(x) = \frac{x}{\left[\pi c\, x\, f(x)\right]^2 + \left[1 - c - \pi c\, x\, H_f(x)\right]^2}$$

Avec H_f la transformée de Hilbert définie comme suit :

$$H_f(x) = \frac{1}{\pi}\,\text{PV}\int_{-\infty}^{+\infty} \frac{f(t)}{t-x}\,dt$$

Et PV la valeur principale de Cauchy définie de la façon suivante :

$$\text{PV}\int_{-\infty}^{+\infty} \frac{\varphi(t)}{t-x}\,dt = \lim_{\varepsilon\to 0^+}\left(\int_{-\infty}^{x-\varepsilon} \frac{g(t)}{t-x}\,dt + \int_{x+\varepsilon}^{+\infty} \frac{g(t)}{t-x}\,dt\right)$$

La densité des valeurs propres de Σ étant inconnue, cet estimateur n'est pas *bona fide*. L'estimateur naturel serait la distribution empirique de S, mais celle-ci n'est pas continue. L'utilisation d'un noyau s'impose donc. ce noyau doit vérifier les propriétés suivantes :

1. κ est continu, pair, et doit être une mesure de probabilité. $\text{Supp}(\kappa) \subset [-R, R]$, de moyenne 0 et de variance 1.
2. Sa transformée de Hilbert H_κ existe et est continu sur $\mathbb{R}$.
3. κ et H_κ ont des variations bornées.

Le noyau d'Epanechnikov est retenu :

$$\kappa^E(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right)^+$$

$$H_{\kappa^E}(x) = \begin{cases} -\dfrac{3x}{10\pi} + \dfrac{3}{4\sqrt{5}\,\pi}\left(1 - \dfrac{x^2}{5}\right)\ln\left|\dfrac{\sqrt{5}-x}{\sqrt{5}+x}\right| & \text{si } |x| \le \sqrt{5} \\ -\dfrac{3}{10\pi} & \text{sinon} \end{cases}$$

Estimateurs :

$$\tilde{f}_n(x) = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{\lambda_{n,i}\, h_n}\, \beta^E\left(\frac{x - \lambda_{n,i}}{h_n}\right), \qquad \text{avec} \qquad h_n = \frac{1}{n^{1/3}}, \quad h_{n_i} = h_n * \lambda_i$$

$$H\tilde{f}_n(x) = \frac{1}{p}\sum_{i=1}^{p} \frac{1}{h_{n,i}}\, H_{\beta^E}\left(\frac{x - \lambda_{n,i}}{h_{n,i}}\right)$$

où

$$d_{n,i} = \frac{\lambda_i}{\left[\pi\lambda_{n,i}\frac{p}{n}, \tilde{f}_n(\lambda_{n,i})\right]^2 + \left[1 - \frac{p}{n} - \pi\lambda_{n,i}\frac{p}{n}, H\tilde{f}_n(\lambda_{n,i})\right]^2}$$

Si l'on note la décomposition spectrale de la matrice de covariance empirique

$$S = U\,\mathrm{diag}(\lambda_1, \ldots, \lambda_p)\, U^\top,$$

alors l'estimateur RIE s'écrit

$$\widehat{\Sigma}_{\mathrm{AS}} = U\,\mathrm{diag}\big(d_{n_1}, \ldots, d_{n_p}\big)\, U^\top,$$

Sous ces hypothèses, cet estimateur est asymptotiquement optimal pour la perte *Minimum Variance* (MV). Cette perte est définie par

$$L^{MV} = \frac{\mathrm{tr}\left(\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\right)/p}{\left(\mathrm{tr}\left(\hat{\Sigma}^{-1}\right)/p\right)^2} - \frac{p}{\mathrm{tr}\left(\Sigma_n^{-1}\right)}$$

L'intérêt de cette métrique tient au fait qu'elle fait intervenir l'inverse de la matrice nettoyée, ce qui correspond directement aux applications financières (optimisation de portefeuille). La perte MV et la norme de Frobenius peuvent donc conduire à des conclusions très différentes sur la qualité des estimateurs.

Le comportement de cet estimateur diffère profondément de celui du *Linear Shrinkage* (LS). Là où LS agit globalement sur l'ensemble du spectre, l'Analytical Shrinkage (AS) agit localement. Cet effet provient de la présence de la transformée de Hilbert dans l'expression analytique du shrinkage. Celle-ci agit comme un mécanisme d'attraction spectrale :

— positive avant un amas de valeurs propres,
— négative après cet amas,
— décroissante en 1/distance à mesure que l'on s'en éloigne.

Compte tenu de la structure typique du spectre ("bulk" et "spikes"), les valeurs propres isolées fortement informatives (spikes) sont peu affectées, car éloignées du bulk. En revanche, le bulk est légèrement resserré.

L'Analytical Shrinkage Estimator peut être comparé à l'estimateur Oracle MV, qui n'est pas un estimateur *bona fide* puisqu'il dépend de quantités inobservables. Il constitue néanmoins une borne supérieure pour la performance atteignable par les estimateurs RIE.

Si l'on note $m(z)$ la transformée de Stieltjes de la loi spectrale limite de S, alors le shrinkage oracle des valeurs propres s'écrit

$$\delta_i^{\mathrm{oracle}} = \frac{\lambda_i}{|1 - q + q\lambda_i m(\lambda_i - i0^+)|^2}.$$

L'estimateur Oracle MV est alors

$$\widehat{\Sigma}_{\mathrm{Oracle}} = U\,\mathrm{diag}\big(\delta_1^{\mathrm{oracle}}, \ldots, \delta_p^{\mathrm{oracle}}\big)\, U^\top.$$

Chapitre 5

Principal Orthogonal complement Thresholding

En finance, une grande partie de la corrélation entre actifs s'explique par des mouvements communs : le marché, le taux d'intérêt, l'inflation, la croissance...

On suppose que chaque actif i suit un modèle factoriel approximatif de la forme

$$y_{it} = b_i^\top f_t + u_{it}, \qquad i = 1, \dots, p, \quad t = 1, \dots, T.$$

Ici :

$f_t \in \mathbb{R}^K$ désigne le vecteur des facteurs communs au temps t.

$b_i \in \mathbb{R}^K$ correspond aux *loadings* de l'actif i,c'est-à-dire aux sensibilités de cet actif aux facteurs communs.

u_{it} représente la composante idiosyncratique : il capture ce qui reste une fois les mouvements communs retirés . On suppose que u_{it} est non corrélé aux facteurs au second ordre, c'est-à-dire que

$$\operatorname{cov}(f_t, u_t) = 0.$$

Donc pour nos données, on définit :

$$Y = (y_1, \dots, y_T) \in \mathbb{R}^{p \times T}. \qquad F = \begin{pmatrix} f_1^\top \\ \vdots \\ f_T^\top \end{pmatrix} \in \mathbb{R}^{T \times K}, \qquad U = (u_1, \dots, u_T) \in \mathbb{R}^{p \times T}.$$

avec :

$$y_t = \begin{pmatrix} y_{1t} \\ \vdots \\ y_{pt} \end{pmatrix} \in \mathbb{R}^p, \qquad u_t = \begin{pmatrix} u_{1t} \\ \vdots \\ u_{pt} \end{pmatrix} \in \mathbb{R}^p, \qquad B = \begin{pmatrix} b_1^\top \\ \vdots \\ b_p^\top \end{pmatrix} \in \mathbb{R}^{p \times K}.$$

Alors la forme matricielle globale du modèle est :

$$Y = BF^\top + U.$$

Sous l'hypothèse de stationnarité :

$$\Sigma = \operatorname{Cov}(y_t) = \operatorname{Cov}(Bf_t + u_t).$$

En supposant que les facteurs et les composantes idiosyncratiques sont non corrélés $\operatorname{Cov}(f_t, u_t) = 0$, on obtient :

$$\Sigma = B \operatorname{Cov}(f_t) B^\top + \operatorname{Cov}(u_t).$$

En posant

$$\Sigma_f = \operatorname{Var}(f_t) \in \mathbb{R}^{K \times K}, \qquad \Sigma_u = \operatorname{Var}(u_t) \in \mathbb{R}^{p \times p},$$

la décomposition prend la forme compacte

$$\boxed{\Sigma = B\Sigma_f B^\top + \Sigma_u.}$$

Le terme $B\Sigma_f B^\top$ est une matrice de rang au plus K. Il capture la structure systématique des co-mouvements, liée aux facteurs communs.

Le second terme Σ_u correspond à la covariance des composantes idiosyncratiques. Il représente les dépendances résiduelles entre actifs une fois les effets communs retirés. La plupart de ces corrélations résiduelles sont faibles.

$$\Sigma = \underbrace{B\Sigma_f B^\top}_{\text{faible rang (structure factorielle)}} + \underbrace{\Sigma_u}_{\text{résidu idiosyncratique}}.$$

Sous l'hypothèse de facteurs pervasifs, les valeurs propres associées à $B\Sigma_f B^\top$ sont d'ordre $O(p)$, tandis que celles de Σ_u sont bornées. Cette séparation d'échelle implique que les K plus grandes valeurs propres de Σ divergent lorsque $p \to \infty$, alors que le reste du spectre demeure borné. Cette propriété rend la structure factorielle identifiable par analyse spectrale.

On pose

$$\widehat{\Sigma} = \frac{1}{T} Y Y^\top.$$

On considère alors sa décomposition spectrale

$$\widehat{\Sigma} = \sum_{i=1}^{p} \widehat{\lambda}_i \widehat{\xi}_i \widehat{\xi}_i^\top, \qquad \widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p.$$

En pratique, la difficulté est que K est inconnu. On le sélectionne par un critère d'information de type Bai–Ng. Pour chaque entier K_1, on projette les données sur l'espace engendré par les K_1 premiers vecteurs propres et on considère l'erreur de reconstruction

$$\frac{1}{pT} \left\| Y - \widehat{B}_{K_1} \widehat{F}_{K_1}^\top \right\|_F^2,$$

qui mesure la variance résiduelle après projection sur un sous-espace de dimension K_1.

On définit alors le critère pénalisé

$$IC(K_1) = \log\left(\frac{1}{pT} \left\| Y - \widehat{B}_{K_1} \widehat{F}_{K_1}^\top \right\|_F^2 \right) + K_1 g(T,p),$$

où $g(T,p)$ est une pénalité telle que $g(T,p) \to 0$ mais $(p \wedge T) g(T,p) \to \infty$.

La pénalité impose un compromis biais-variance : on cherche le plus petit K permettant de capturer la structure systématique significative, sans absorber du bruit idiosyncratique.

On définit alors

$$\widehat{K} = \arg \min_{0 \leq K_1 \leq K_{\max}} IC(K_1).$$

Une fois $\widehat{K}$ déterminé, on définit la composante factorielle estimée

$$\widehat{\Sigma}_{\text{facteur}} = \sum_{i=1}^{\widehat{K}} \widehat{\lambda}_i \widehat{\xi}_i \widehat{\xi}_i^\top.$$

La partie résiduelle est donnée par le complément orthogonal

$$\widehat{R}_{\widehat{K}} = \widehat{\Sigma}_{\text{u}} = \sum_{i=\widehat{K}+1}^{p} \widehat{\lambda}_i \widehat{\xi}_i \widehat{\xi}_i^\top.$$

Cette matrice estime Σ_u, mais elle reste bruitée.

Sous l'hypothèse de sparsité approximative de Σ_u, on applique un seuillage

.

Pour déterminer quelles entrées doivent être conservées, on évalue l'incertitude statistique associée à chaque coefficient $\widehat{\sigma}_{ij}$. On introduit alors

$$\widehat{\theta}_{ij} = \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{u}_{it}\widehat{u}_{jt} - \widehat{\sigma}_{ij}\right)^2.$$

La quantité $\widehat{\theta}_{ij}$ évalue la précision de l'estimation de $\widehat{\sigma}_{ij}$.

Le seuil est alors choisi proportionnel à l'écart-type estimé $\sqrt{\widehat{\theta}_{ij}}$, multiplié par un facteur d'échelle dépendant de p et T :

$$\tau_{ij} = C\sqrt{\widehat{\theta}_{ij}}\left(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{T}}\right).$$

On applique alors un seuillage entrée par entrée :

$$\widehat{\sigma}_{ij}^{T} = \begin{cases} \widehat{\sigma}_{ij}, & i = j, \\ \widehat{\sigma}_{ij}\mathbf{1}\{|\widehat{\sigma}_{ij}| \geq \tau_{ij}\}, & i \neq j. \end{cases}$$

L'estimateur final POET est alors

$$\widehat{\Sigma}_{\widehat{K}} = \widehat{\Sigma}_{\text{facteur}} + \widehat{\Sigma}_{u}^{T}.$$

Chapitre 6

Génération des données (DGP)

L'approche classique consistant à tester des méthodes sur des données synthétiques "propres et contrôlées" pour ensuite les tester sur les données réelles, naturellement bruitées et plus instables, est ici source de difficultés spécifiques.

En effet, générer une matrice de covariance réaliste est complexe, sujet à de mauvaises spécifications et hypothèses, et peut facilement biaiser l'évaluation des méthodes de nettoyage.

Les données sont générées de la manière suivante :

1. Modèle factoriel pour la vraie covariance

On construit la matrice de covariance théorique $\Sigma \in \mathcal{S}_p^+(\mathbb{R})$ via un modèle factoriel à $k = 1$ facteur :

$$\Sigma = BB^\top + \Psi$$

où :

— $B \in \mathcal{M}_{p\times k}(\mathbb{R})$ est la matrice de facteurs, avec $B_{ij} \overset{\text{i.i.d.}}{\sim} 0.01 \times \mathcal{N}(0, 1)$,

— $\Psi = \text{diag}(\psi_1, \ldots, \psi_p)$ est une matrice diagonale de bruits idiosyncratiques, avec

$$\psi_i = 0.05 \times \exp(\varepsilon_i), \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(-4,\, 1).$$

2. Simulation des observations à queues lourdes

Afin de modéliser des queues plus lourdes qu'une loi gaussienne, on tire les innovations selon une loi de Student multivariée de degré de liberté $\nu = 10$:

$$\tilde{Z}_{t,j} \overset{\text{i.i.d.}}{\sim} \mathcal{T}(\nu), \qquad t = 1, \ldots, T, \quad j = 1, \ldots, p.$$

On normalise ensuite pour obtenir des innovations de variance unitaire :

$$Z_{t,j} = \frac{\tilde{Z}_{t,j}}{\sqrt{\dfrac{\nu}{\nu - 2}}}, \qquad \text{de sorte que } \operatorname{Var}(Z_{t,j}) = 1.$$

3. Décomposition de Cholesky et génération des données

On effectue la décomposition de Cholesky de Σ :

$$\Sigma = LL^\top,$$

où $L \in \mathcal{M}_p(\mathbb{R})$ est triangulaire inférieure. Les observations sont alors construites par :

$$X_t = LZ_t^\top, \qquad t = 1, \ldots, T,$$

ou, en notation matricielle pour $\mathbf{X} \in \mathcal{M}_{T\times p}(\mathbb{R})$:

$$\mathbf{X} = \mathbf{Z}L^\top,$$

Justification du choix de ν et de la méthode de génération

Queues lourdes et moments finis.

Le choix d'une loi de Student $\mathcal{T}(\nu)$ pour les innovations est motivé par le souci de réalisme : les rendements financiers présentent empiriquement des queues plus lourdes qu'une loi gaussienne. Cependant, les hypothèses théoriques du cadre étudié requièrent l'existence des moments d'ordre 8 des observations. Or, pour $X \sim \mathcal{T}(\nu)$, le moment d'ordre k est fini si et seulement si :

$$\nu > k.$$

Le choix $\nu = 10$ garantit donc l'existence des moments jusqu'à l'ordre 9, ce qui satisfait strictement la condition requise :

$$\mathbb{E}\big[X^8\big] < +\infty \iff \nu > 8, \qquad \nu = 10 > 8.$$

Justification du modèle factoriel.

La structure $\Sigma = BB^\top + \Psi$ est un modèle factoriel à $k = 1$ facteur, standard en finance (modèle de marché). Elle garantit par construction que $\Sigma \in \mathcal{S}_p^+(\mathbb{R})$, puisque pour tout $x \in \mathbb{R}^p$:

$$x^\top \Sigma x = \|B^\top x\|^2 + x^\top \Psi x \geq 0,$$

avec égalité seulement si $B^\top x = 0$ et $x^\top \Psi x = 0$, ce qui est exclu car $\Psi \succ 0$ (ses éléments diagonaux $\psi_i > 0$ p.s.). Le terme $BB^\top$ capture une structure de corrélation commune (facteur systémique), tandis que Ψ modélise le risque idiosyncratique propre à chaque actif.

Justification de la décomposition de Cholesky.

La décomposition $\Sigma = LL^\top$ permet de simuler exactement des vecteurs de covariance Σ à partir d'innovations Z_t de covariance I_p :

$$X_t = LZ_t \implies \mathrm{Cov}(X_t) = L \underbrace{\mathrm{Cov}(Z_t)}_{=I_p} L^\top = \Sigma.$$

C'est la méthode de simulation la plus stable numériquement pour des matrices Σ de grande dimension, car la décomposition de Cholesky préserve le conditionnement de la matrice et est calculable en $\mathcal{O}(p^3)$.

Exemple de spectre obtenu.

Les spectres obtenus par génération de données synthétiques et par les données réelles sont ici représentés pour confirmer le réalisme de la méthode.

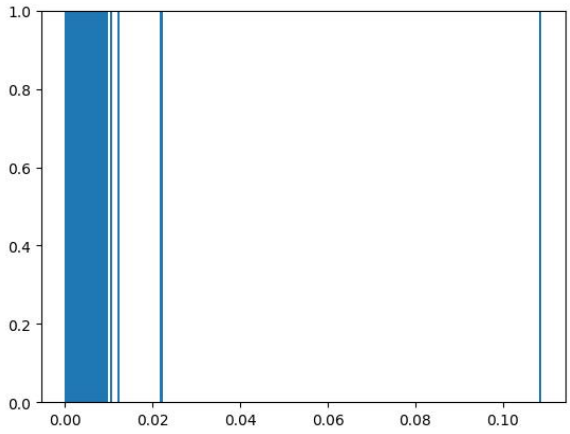

FIGURE 6.1 – Spectre obtenu pour une matrice générée

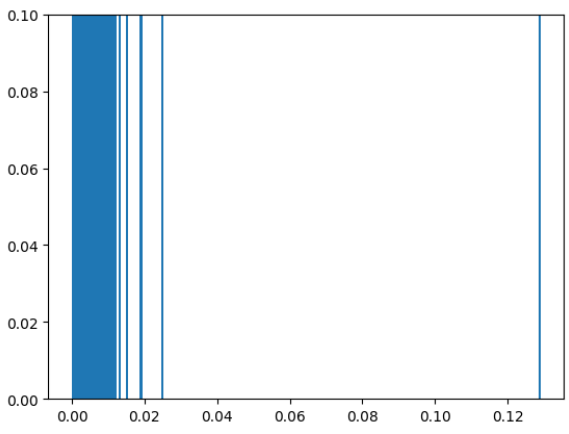

FIGURE 6.2 – Spectre obtenu pour une matrice réelle

Chapitre 7

Résultats sur les données synthétiques

Nous avons utilisé comme métriques sur les données synthétiques :
— La norme de Frobenius, pour l'impact sur les grandes valeurs propres.
— La perte MV, pour l'impact sur les faibles valeurs propres.
— La distance de Kolmogorov-Smirnov entre les distributions des rendements générés avec la vraie matrice de covariance, et ceux générés avec la matrice estimée. De même, nous avons considéré le taux de rejet de l'hypothèse nulle associée au test de Kolmogorov-Smirnov. L'objectif est d'évaluer l'impact de la transformation sur les distributions des rendements.
— La frontière d'efficience de Markowitz entre le rendement et le risque du portefeuille construit avec la matrice estimée.

Norme de Frobenius

On rappelle que le gain d'une méthode par rapport à la matrice empirique est calculé de la manière suivante :

$$\text{Gain relatif} = \frac{||\hat{\Sigma}^{\text{empirique}} - \Sigma^{\text{true}}||_F - ||\hat{\Sigma}^{\text{méthode}} - \Sigma^{\text{true}}||_F}{||\hat{\Sigma}^{\text{empirique}} - \Sigma^{\text{true}}||_F}$$

Ce gain est donc compris entre 0 (performance similaire à celle de la matrice empirique) et 1 (erreur négligeable devant l'erreur de la matrice empirique). Il est à maximiser.

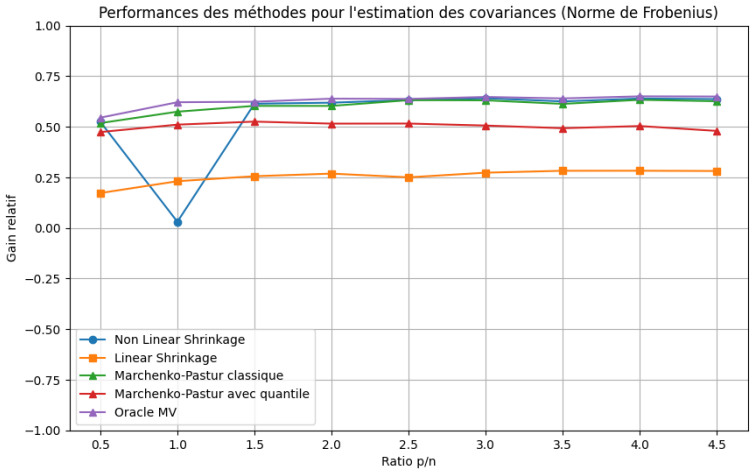

FIGURE 7.1 – Gain des méthodes pour la norme de Frobenius

Plusieurs observations peuvent être faites pour cette métrique :

— La méthode non linéaire est très proche de l'oracle MV, ce qui indique une excellente performance de AS relativement à ce que sa classe d'estimateurs est capable d'atteindre. Le décrochage d'AS quand le ratio vaut exactement 1 est dû à la divergence du modèle quand $n = p$.

— La méthode de shrinkage linéaire est inférieure aux autres méthodes. Cette observation est étonnante au premier abord dans la mesure où l'intensité du shrinkage linéaire est calculée en optimisant justement la norme de Frobenius. La performance médiocre ne vient donc pas d'une mauvaise calibration des paramètres, mais plutôt d'une inadéquation de la méthode par rapport à la structure du spectre. En effet, shrinker les valeurs propres de manière uniforme vers leur moyenne est pertinent lorsqu'elles sont dispersées autour d'une moyenne pertinente. Or les spikes du spectre de la matrice de covariance ne doivent surtout pas être considérés comme une dispersion dû à du bruit. LS est donc meilleure dans le cas d'un bulk simple résultant d'un étalement des valeurs propres, et devient mauvais lorsque des spikes apparaissent "loin" du bulk. Si ces spikes caractérisent la structure de la covariance, on pourrait faire l'hypothèse que plus portefeuille est structuré, moins LS sera performant.

— Notre implémentation du LS shrink vers $\bar{\lambda} I_p$ où $\bar{\lambda} = \mathrm{tr}(S)/p$, ce qui suppose que toutes les variances sont identiques. Or, notre modèle factoriel génère des variances hétérogènes (ratio max/min ≈ 7).

Perte MV (minimum variance)

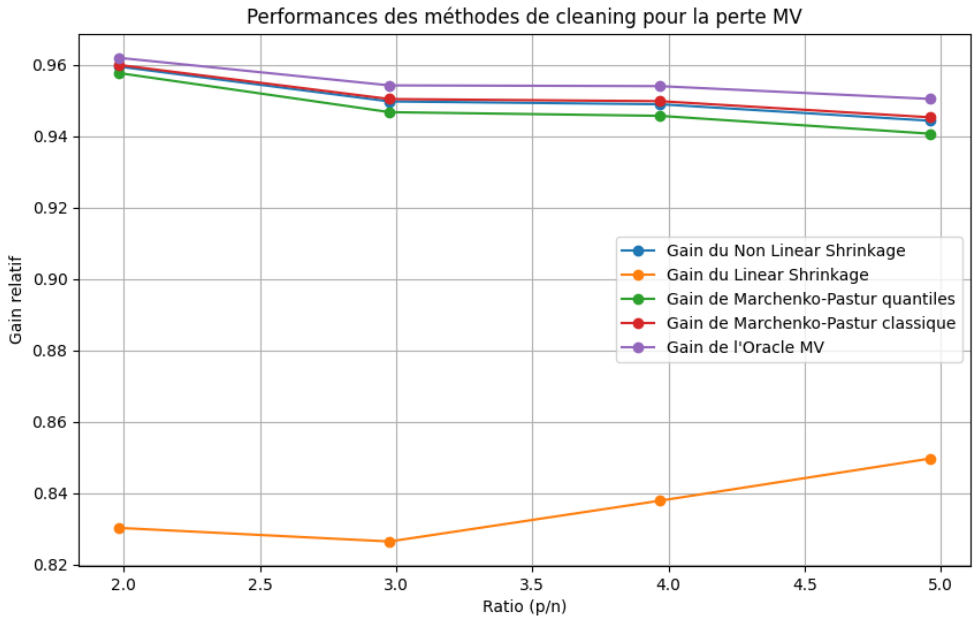

FIGURE 7.2 – Gain des méthodes pour la perte MV (minimum variance)

La performance du shrinkage non linéaire est excellente pour la perte MV, conformément au résultat d'optimalité de cet estimateur pour cette perte.

Les deux méthodes de clipping de Marchenko Pastur montrent d'excellentes performances également. Ceci peut s'expliquer par le fait que le clipping recentre ce qui est considéré comme du bruit autour de sa moyenne (les valeurs propres inférieures à λ_{max}). Ceci a pour effet de légèrement augmenter les valeurs propres très proches de 0, ce qui pourrait stabiliser l'inversion de la matrice et participer à une bonne perte MV.

La méthode de shrinkage linéaire est nettement inférieure, mais a tout de même un réel apport relatif à la matrice empirique. En effet les petites valeurs propres sont légèrement augmentées aussi, ce qui a le même effet stabilisateur que les méthodes de clipping. Le shrinkage des hautes valeurs propres a néanmoins le même effet négatif en ce qu'il gomme les facteurs structurants de la covariance, ce qui explique sa performance inférieure.

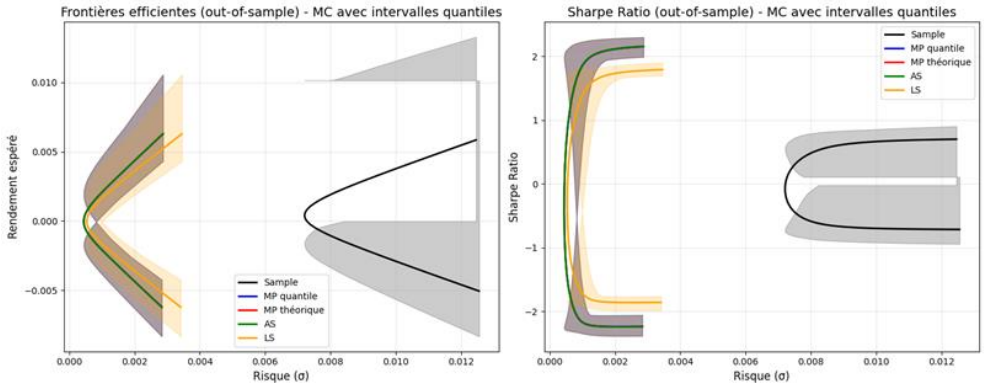

FIGURE 7.3 – Frontières de Markowitz avec celle de la matrice empirique

Frontières d'efficience de Markowitz

Toutes les méthodes permettent une grande amélioration des performances des portefeuilles générés. Pour des questions de clarté, le même graphe est affiché ci-dessous en retirant la frontière issue de la matrice empirique.

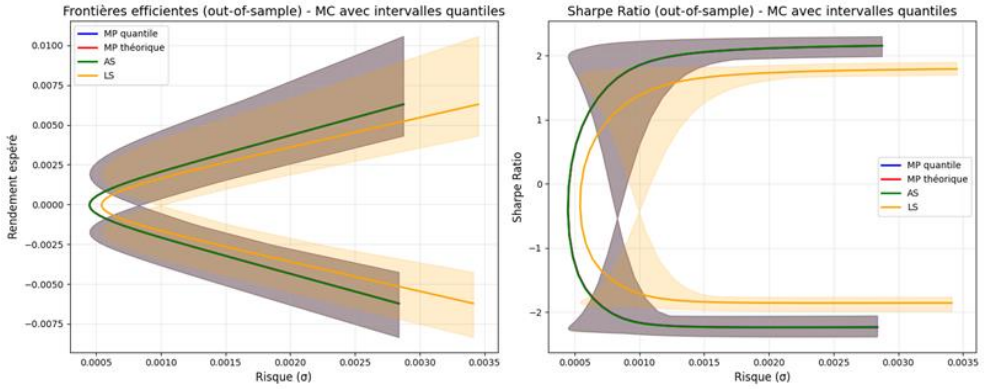

FIGURE 7.4 – Frontières de Markowitz sans celle de la matrice empirique

Nous remarquons que la méthode de shrinkage non linéaire et les méthodes de Marchenko-Pastur sont équivalentes. La méthode de shrinkage linéaire est légèrement inférieure, ce qui confirme la prédiction obtenue par les performances sur la perte MV.

La structure de la covariance supposée par l'estimateur de shrinkage linéaire explique cette plus faible performance.

On peut également souligner que les résultats sur la perte MV sont similaires à ceux obtenus sur la frontière d'efficience. Ceci confirme l'intérêt de cette métrique de perte, plus simple et rapide à calculer qu'une frontière complète.

Conclusions temporaires

L'étude sur données synthétiques permet de supposer la supériorité des méthodes de shrinkage non linéaire et de clipping de Marchenko Pastur (qui sont par ailleurs d'efficacité comparable) sur la méthode de shrinkage linéaire. Toutes les méthodes sont néanmoins très positives en ce qu'elles améliorent largement les performances des portefeuilles optimaux générés, par rapport à la matrice empirique. Ces conclusions sont soumises à la bonne génération des données qui, comme mentionné plus haut, n'est pas complètement assuré.

Chapitre 8

Résultats sur les données financières réelles

Origine des données

Les données réelles sont obtenues par YahooFinance et proviennent de 1974 actifs du $S\&P500$, du Nasdaq et d'Euronext. La liste complète des tickers des actifs utilisés est en annexe de ce rapport. Ceci nous donne un rapport $\frac{p}{n} = 7.83$ pour un an de données ($n = 252$). Cet ordre de grandeur se rapproche des cas usuels en finance. Ces données sont échantillonnées sur la période 2015 - 2025 afin de permettre suffisamment d'itérations sur une période d'un an.

Dans un régime où $p > n$ nous avons $p - n$ valeurs propres nulles, ce qui implique la présence d'un pic important de la densité des valeurs à 0, suivi du bulk classique, puis des spikes.

Effet de chaque méthode sur le spectre

Nous avons commencé par évaluer l'effet de chaque méthode sur le spectre d'une matrice de covariance empirique réelle.

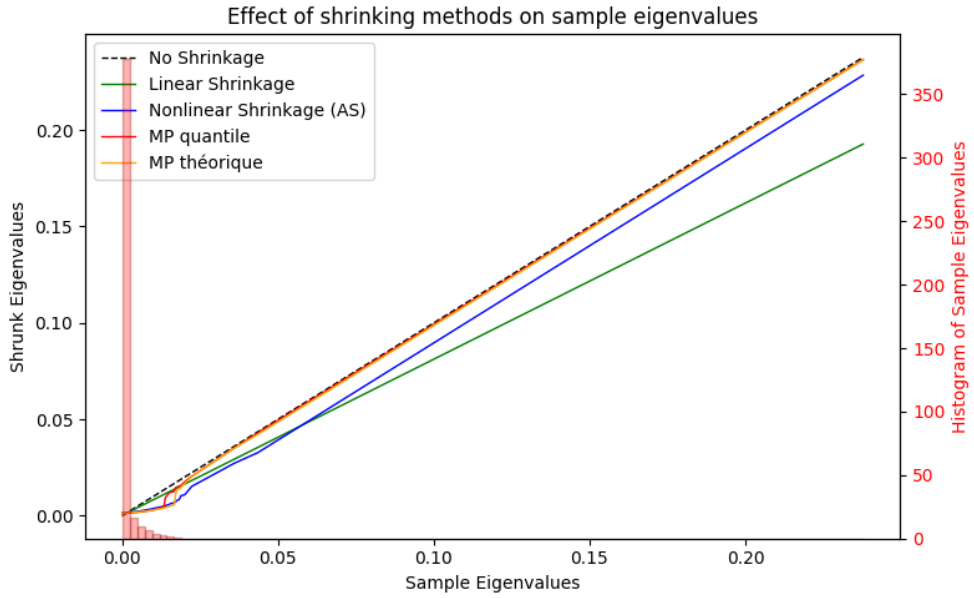

FIGURE 8.1 – Correction des méthodes sur les valeurs propres, avec un spectre concentré

Nos observations sont les suivantes :

1. Nous observons l'effet du clipping pour les deux méthodes de Marchenko Pastur, dont les seuils sont différents.

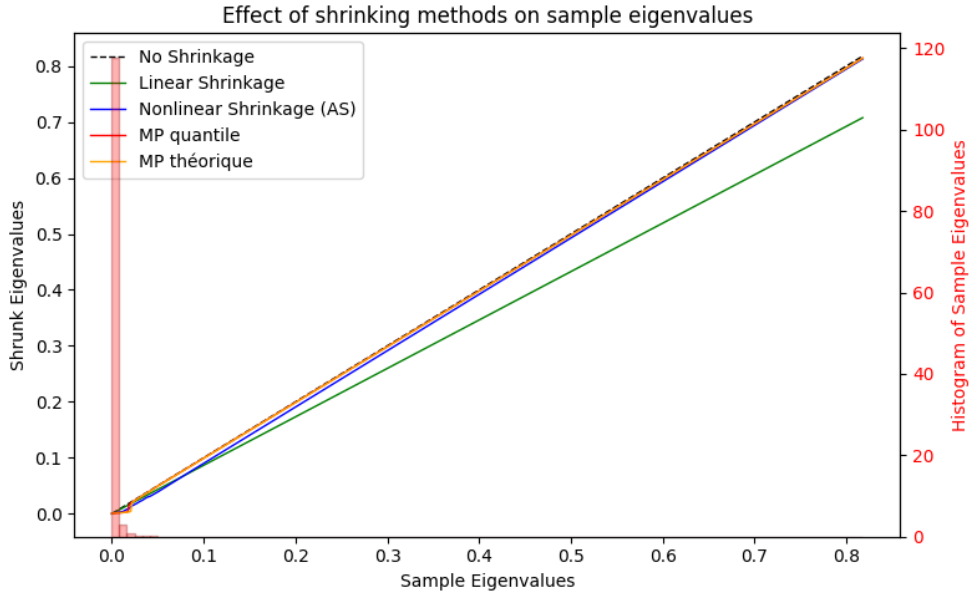

FIGURE 8.2 – Correction des méthodes sur les valeurs propres, avec un spectre plus étalé

2. L'effet global de la méthode de shrinkage linéaire apparemment nettement, avec un recentrage autour de la moyenne (qui en l'occurence est proche de 0). Les grandes valeurs propres sont ainsi réduites.
3. L'effet local de la méthode de shrinkage non linéaire est également mis en évidence. En effet, lorsque toutes les valeurs propres sont relativement proches, toutes les valeurs propres sont "attirées" par le centre de masse proche de 0, et l'effet atteint toutes les valeurs propres (figure 8.1),. A l'inverse, lorsque les plus grandes valeurs propres sont relativement éloignées du centre de masse proche de 0, l'effet du shrinkage est presque nul (figure 8.2).

Comparaison des frontières d'efficience de Markowitz

La métrique de référence pour les données réelles est la frontière d'efficience de Markowitz. La meilleure méthode est celle dont la frontière se situe la plus à gauche, et idéalement contient les autres. Les résultats sont les suivants :

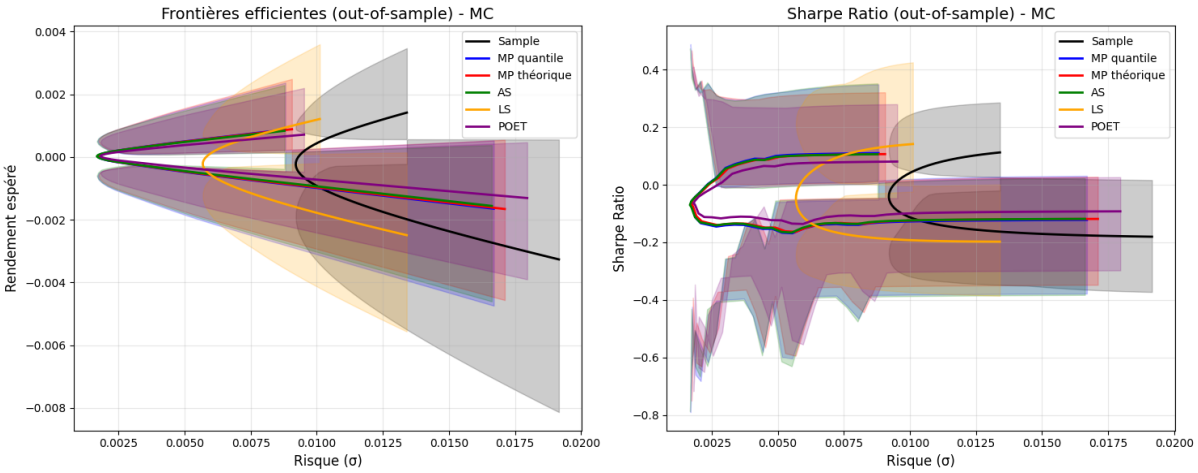

FIGURE 8.3 – Frontières d'effience de Markowitz pour chaque méthode, et Sharpe ratio

Les méthodes sont désignées par : AS pour analytical shrinkage (shrinkage non linéaire), LS pour linear shrinkage, MP pour le clipping de Marchenko Pastur selon le quantile 99 des valeurs propres, MP2 pour le clipping classique de Marchenko Pastur. Les barres d'erreurs correspondent aux quantiles 99/1.

Plusieurs constats peuvent être faits :

— Tout d'abord, toutes les méthodes ont des performances supérieures à celles de la sample matrix.

— La méthode de shrinkage linéaire a des performances inférieures à celles des méthodes de Marchenko Pastur et de shrinkage non linéaire. Cette conclusion est conforme aux observations sur les données synthétiques, et fait sens compte tenu que la méthode de shrinkage linéaire est optimale pour la norme de Frobenius uniquement. Les spikes informatifs et structurants sont ramenés à la moyenne, ce qui n'est pas souhaitable.

— Les méthodes de Marchenko Pastur avec quantile et selon la borne théorique sont très proches, ce qui supporte l'hypothèse que l'on peut considérer comme des spikes le quantile 99 des valeurs propres.

— Les méthodes de Marchenko Pastur et la méthode de shrinkage non linéaire sont similaires, ce qui suggère que l'effet local du shrinkage non linéaire sur les spikes est similaire à celui d'un clipping, qui les laisse inchangés. Ainsi, les valeurs propres significatives sont conservées dans un cas parce qu'elles font partie du quantile 99, et dans un autre cas parce qu'elles sont suffisamment "loin" du bulk. Cette observation seule permet de dire que les spikes sont éloignés du bulk, et ne sont pas simplement dans sa continuité en étant supérieurs.

— Les méthodes de Marchenko Pastur et du shrinkage non linéaire sont similaires en performance, mais le clipping de Marchenko Pastur étant plus simple conceptuellement et pratiquement, on pourrait privilégier cette méthode.

Chapitre 9

Conclusion et perspectives

Les résultats obtenus amènent plusieurs conclusions :

1. **Apport de la théorie des matrices aléatoires.** L'utilisation de la théorie des matrices aléatoires permet d'obtenir des résultats empiriquement convaincants. Elle fournit un cadre théorique adapté au régime de grande dimension (N, T grands avec N/T fini) et justifie l'utilisation d'estimateurs spectraux fondés sur la structure *bulk + spikes*. Ces résultats encouragent l'intégration de ces estimateurs dans les procédures d'optimisation de portefeuilles.
2. **Limites des données synthétiques.** L'étude des données synthétiques s'avère délicate. D'une part, ces données peuvent être significativement éloignées des données réelles, ce qui interroge sur la pertinence de leur utilisation pour évaluer les estimateurs. D'autre part, même dans le cas où elles reproduisent correctement certaines propriétés stylisées des marchés financiers, leur génération présente une complexité temporelle élevée, typiquement en $\mathcal{O}(p^3)$, ce qui constitue un obstacle computationnel important en grande dimension.
3. **Limites de la norme de Frobenius.** L'utilisation de la norme de Frobenius comme critère d'évaluation, ainsi que des estimateurs optimisés selon cette métrique, apparaît discutable dans un contexte d'allocation de portefeuille. En effet, cette norme pénalise principalement les erreurs portant sur les grandes valeurs propres, alors que l'optimisation moyenne-variance est particulièrement sensible aux petites valeurs propres. Cet écart entre critère mathématique et objectif financier explique vraisemblablement les performances inférieures observées pour l'estimateur par shrinkage linéaire.
4. **Cas de l'oracle moyen.** La méthode de l'oracle moyen présente des performances du même ordre de grandeur que les autres méthodes avancées. Toutefois, sa complexité temporelle très élevée rend son utilisation difficile en pratique, en particulier en grande dimension. Pour cette raison, elle a été écartée dans une perspective opérationnelle. Sa structure non paramétrique conserve néanmoins un intérêt théorique important, car elle permet de vérifier que les performances des méthodes paramétriques ne résultent pas d'une mauvaise spécification des modèles.
5. **Comparaison des estimateurs spectraux.** Les performances des estimateurs fondés sur le clipping de Marchenko–Pastur et sur le shrinkage non linéaire sont comparables et nettement supérieures à celles du shrinkage linéaire. Toutes ces méthodes permettent néanmoins une amélioration significative des performances des portefeuilles par rapport à l'utilisation directe de la matrice de covariance empirique, confirmant l'importance de la régularisation spectrale.

Les perspectives que nous envisageons sont les suivantes :

1. **Augmenter le nombre d'actifs.** Utiliser un nombre plus important d'actifs permettrait d'approcher des ordres de grandeur utilisés en pratique en banque. Idéalement, nous souhaiterions arriver à 10.000 actifs pour un an de données.

Bibliographie

[1] Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, **88**(2), 365–411.

[2] Ledoit, O., & Wolf, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Annals of Statistics*, **48**(5), 3043–3065.

[3] Bun, J., Bouchaud, J.-P., & Potters, M. (2017). Cleaning large correlation matrices : tools from random matrix theory. *Physics Reports*, **666**, 1–109.

[4] Laloux, L., Cizeau, P., Bouchaud, J.-P., & Potters, M. (1999). Noise dressing of financial correlation matrices. *Physical Review Letters*, **83**(7), 1467–1470.

[5] Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **75**(4), 603–680.

[6] Bai, Z. D., & Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer.

[7] Bongiorno, C., Challet, D., & Loeper, G. (2021). Cleaning the covariance matrix of strongly nonstationary systems with time-independent eigenvalues. *arXiv preprint*, arXiv :2111.13109. https://arxiv.org/abs/2111.13109