Mémoire de fin d'études

présenté pour l'obtention du diplôme de maîtrise en gestion

spécialisation ingénierie financière

Comparaison de modèles factoriels linéaires et non-linéaires pour l'univers des actions du S&P500

Soumis par:

Frédéric Siino

Département de sciences de la décision

Supervisé par:

David Ardia

HEC Montréal

18 août, 2023

Table des matières

1. Introduction

Les chercheurs et les investisseurs en finance étudient et analysent les rendements d'actifs dans le but d'améliorer l'évaluation de leur valeur et ainsi générer des bénéfices. Les chercheurs ont découvert que le rendement d'un actif peut être influencé par les rendements d'un groupe d'actifs. Il existe également des facteurs financiers et des facteurs économiques qui peuvent influencer les rendements d'un actif. Les chercheurs se sont d'abord penchés sur des techniques de modélisation linéaire afin de déterminer la meilleure combinaison de facteurs. Le premier modèle factoriel essayant d'expliquer les rendements à travers différents actifs est le modèle d'évaluation des actifs financiers. Ce dernier introduit le concept de risque systématique, soit le bêta du marché. Ce facteur représente une mesure de volatilité d'un titre en comparaison au marché dans son ensemble. Il a été introduit par Jack Treynor (1961), William Sharpe (1964), John Lintner (1965 a,b) et par Jan Mossin (1966). Un bêta d'une valeur de un indique que l'actif est aussi volatile que le marché et qu'il a tendance à évoluer dans la même direction. Un bêta de plus de un indique que l'actif est plus volatile que le marché tandis qu'un beta de moins de un indique l'inverse. Davantage de tests ont révélé que la relation entre le rendement de l'actif et le bêta du marché était plus plate que prédit par le modèle, avec une grande dispersion dans les régressions, des coefficients très faibles et une ordonnée à l'origine qui impliquait un taux sans risque peu plausible. Ces travaux ont ensuite été validés empiriquement dans les années 70 par Black et al. (1972) puis par Fama et Macbeth (1973). Afin de contourner les problèmes liés au modèle, ils ont groupé les actifs dans un portefeuille permettant ainsi de réduire la variance résiduelle et d'obtenir de meilleures régressions.

Afin d'expliquer davantage le rendement des actifs, Fama et French (1993) ont formé des portefeuilles basés des facteurs propres aux firmes : la capitalisation des titres et le ratio de la valeur comptable à la valeur du marché (B/M). Ils ont trouvé qu'après avoir classé les actifs en fonction de leur capitalisation ou de leur ratio B/M et les avoir regroupés par quantile, les bêtas sont plus grands. Le bêta du portefeuille est plus grand et la rentabilité est plus grande. Ainsi, ils ont proposé leur modèle trifactoriel de valorisation des actifs. Fama et French (1996) ont ensuite avancé diverses explications afin de justifier l'amélioration de leurs résultats en comparaison à ceux de William Sharpe (1964). Notamment, le modèle trifactoriel est capable de capturer la

détresse financière via le facteur de rapport B/M et le facteur de capitalisation. En effet, une firme avec de faibles résultats financiers a généralement un rapport B/M élevé et un facteur de régression positif pour la capitalisation. De plus, cette méthode pouvait expliquer jusqu'à plus de 90 % des rendements[1]. Les méthodes multifactorielles ont continué à se développer, entrainant une quête de facteurs que Cochrane (2011) nommera le « zoo de facteurs ». Harvey et al. (2016) ont recensé une liste de 316 facteurs. Tous les facteurs ne peuvent pas être utilisés dans une même régression, certes. Ceci entrainerait des problèmes de colinéarité, d'endogénéité ou encore d'hétérogénéité. Pour leur part, Fama et French (2015) ont publié un nouvel article proposant une méthode à cinq facteurs, qui intègre les trois facteurs de leur publication de 1993, auxquels ils ajoutent les facteurs d'investissement et de profitabilité. Jensen et al. (2023) tentent de démontrer que malgré le nombre important de facteurs proposé par la littérature, ils peuvent être regroupés en thème et qu'il y a moins de quinze facteurs qui sont réellement nécessaires.

Ces méthodes empiriques sont construites en testant plusieurs nouveaux facteurs et différentes combinaisons. Il existe également d'autres méthodes visant à extraire les composantes principales qui influencent le rendement d'un actif. Au début des années 2000, des chercheurs se sont penchés sur l'utilisation de la théorie des matrices aléatoires. Cette théorie nous permet de comprendre les diverses propriétés des matrices de rendements, notamment les statistiques des valeurs propres. Plusieurs de ces études ont confirmé les facteurs principaux déterminant le rendement des actifs. Laloux et al. (1999) ainsi que Plerou et al. (1999) ont démontré que la valeur propre la plus grande représentait le marché. Ces études ont également été confirmées par Plerou et al. (2002). Ils ont également démontré que la grande capitalisation est l'une des composantes expliquant le plus les rendements. Il est clair que ces facteurs ne sont pas explicitement le résultat de la décomposition des valeurs propres. Le résultat est une « mixture » des vecteurs entrants qui peut être interprétée comme un facteur de marché.

La recherche plus récente se penche sur des méthodes non linéaires. Ce texte comparera donc des méthodes factorielles linéaires ainsi que deux méthodes non linéaires sur l'univers d'actifs du S&P500. Tel que le notent Zhang et al. (2020), l'utilisation de réseau de neurones est de plus en plus populaire dans la recherche. Ils ont ajouté à la recherche d'Heaton et al. (2017) ainsi que Ouyang et al. (2019). Ces derniers comparent plusieurs types d'autoencodeurs dont

[1] Un R^2 supérieur à 90% pour 21 des 25 portefeuilles construit avec des coefficient de régression élevé

l’autoencodeur variationnel afin de déterminer la composition d’un portefeuille en se basant sur les rendements des actifs, mais également sur des ratios comptables. Ils utilisent ainsi les signaux du marché afin de déterminer le portefeuille avec le meilleur rendement. Nous allons utiliser ces types de réseaux de neurones pour déterminer le rendement des actifs composant le S&P500. Le premier type de réseau neuronal utilisé est l’autoencodeur conditionnel. Introduit en finance par Gu et al. (2021), cette technique utilise le rendement de l’actif ainsi que d’autres ratios comptables afin de prédire le rendement de cet actif pour un ensemble de titres. Dans notre cas, nous utiliserons le rendement de différents actifs afin de prédire le rendement d’un titre. Ensuite, nous introduirons l’autoencodeur variationnel. Ce genre de modèle génératif permet, tout comme l’autoencodeur, de représenter les données en réduisant la dimensionnalité, mais il permet également de produire des données en combinant les concepts d’autoencodeur avec des techniques d’inférence probabiliste.

2. Méthodologie

Dans cette section, nous décrivons les différentes méthodologies utilisées pour modéliser les log-rendements. Pour toutes les approches, nous avons séparé les données en deux échantillons distincts : un échantillon d'entrainement et un échantillon test. Les données de l'échantillon d'entrainement nous permettent de calibrer les modèles. La performance finale du modèle est quant à elle déterminée à l'aide de la deuxième série de données. Les données tests, également appelées données hors échantillon, nous permettent d'appliquer la méthode et de comparer la prévision de cette dernière. Pour utiliser les approches impliquant les facteurs explicites, implicites ainsi que l’autoencodeur variationnel, les données ont été séparées en deux échantillons : un échantillon d'entrainement équivalant à 70% des données et un échantillon test représentant les 30% restants. Puisque ce sont des séries temporelles, la séparation des données est ordonnée. Ainsi, réalisé de manière méthodique : les données des échantillons ne sont pas choisies aléatoirement et l'ordre chronologique est conservé. Pour l'approche d'apprentissage profond de l’autoencodeur conditionel, les données ont été séparées également pour respecter les contraintes du réseau de neurones.

2.1. Facteurs explicites

Dans cette section, nous décrivons la méthode de régression avec des facteurs explicites. Afin de comparer nos résultats avec la littérature, nous avons commencé par une régression linéaire. Nous décrivons le rendement d'un actif i (r_i) comme étant une fonction de l'excès de rendement α_i, de variables prédictives x_{ik}, k=1,…, K, de l'exposition aux facteurs β_{ik}, $\mathrm{k} = 1, \dots, \mathrm{K}$ et du risque idiosyncratique (ϵ_i) :

$$r_i = \alpha_i + \sum_{k=1}^{K} \beta_{ik} x_{ik} + \epsilon_i \qquad (1)$$

Afin de comparer les différentes techniques, nous avons d'abord calibré les modèles en utilisant les données d'entrainement. Ensuite, nous avons calculé les coefficients de détermination linéaire R^2décrits dans l'article de Gu et al. (2021). Un coefficient de détermination élevé indique une meilleure explicabilité des log-rendements. La valeur du coefficient sur les données d'entrainement est comprise dans un intervalle [0,1]. Le R^2 décrit par l'équation (*2*) est la somme du carré des différences entre les rendements observés ($r_{i,s}$) et les rendements prédits ($\hat{r}_{i,s}$) dans l'échantillon d'entrainement pour tous les actifs i ($\forall i = 1, \dots, N$) jusqu'au temps s ($\forall s \in S$) soit la période couverte par l'échantillon d'entrainement divisée par la somme des rendements observés au carré.

$$R^2 = 1 - \frac{\sum_{i,s} (r_{i,s} - \hat{r}_{i,s})^2}{\sum_{i,s} {r_{i,s}}^2} \qquad (2)$$

Nous appliquons l'équation (*2*) aux données tests afin de déterminer la capacité de prévision du modèle. Nous obtenons le R^2_{OOS} décrit par l'équation (*3*). Nous calculons la somme du carré des différences entre les rendements observés dans l'échantillon test ($r_{i,u}$) et les rendements prédits ($\hat{r}_{i,u}$) pour tous les actifs i ($\forall i = 1, \dots, N$) jusqu'au temps T ($\forall \mathrm{u} \in T'$) où T' est la période couverte par l'échantillon test, divisée par la somme des rendements observés au carré. Pour le R^2_{OOS} nous pouvons observer un coefficient négatif si les rendements prédits divergent grandement des valeurs observées. La valeur du coefficient peut donc être comprise dans l'intervalle]-∞,1]. Plus la valeur du coefficient s'éloigne de 1, moins le modèle est performant.

$$R^2_{OOS} = 1 - \frac{\sum_{i,u} (r_{i,u} - \widehat{r_{i,u}})^2}{\sum_{i,u} {r_{i,u}}^2} \qquad (3)$$

Il est important de comprendre les limites de notre mesure de performance. Le R^2_{OOS} décrit par l'équation (*3*) peut être séparé en deux composantes : le numérateur et le dénominateur. Afin d'obtenir le meilleur résultat possible, nous devons minimiser le numérateur. Le numérateur est la somme des carrés de la différence entre les rendements observés et les rendements prédits. Le dénominateur est la somme des carrés des rendements observés. Notre mesure ne dépend donc pas uniquement de la valeur prédite par le modèle. Les rendements observés affectent aussi la mesure. Toutefois, la performance de la mesure ne peut être améliorée que si le numérateur diminue.

Les modèles les plus célèbres sont les modèles de Fama et French (1993, 2015) qui ont trois et cinq prédicteurs. Les trois premiers prédicteurs sont les facteurs de taille, de valeur et de marché. Les deux derniers, exclusifs au modèle de Fama-French (2015), sont les facteurs d'investissement et de profitabilité. Ces méthodes sont, en soi, une extension du modèle d'évaluation des actifs financiers (MÉDAF ou *CAPM* en anglais) de William Sharpe (1964) qui décrit le rendement d'un actif *i* (r_i) comme étant une fonction du taux sans risque (r_f), du risque systématique de l'actif (β_i), de l'excès de rendement du marché ($r_m - r_f$) et du risque idiosyncratique (ϵ_i). Pour la méthode élaborée par Fama et French (1993, 2015) le facteur *Small Minus Big* (SMB) est un effet de taille basé sur la capitalisation boursière d'une société. Le SMB mesure l'excès historique des sociétés à petite capitalisation par rapport aux sociétés à grande capitalisation. La principale raison d'être de ce facteur est que, sur le long terme, les sociétés à petite capitalisation ont tendance à enregistrer des rendements plus élevés que les sociétés à grande capitalisation. Dans notre contexte, nous incluons uniquement des actifs à grande capitalisation. L'effet de ce facteur est donc remis en question. Le facteur *High Minus Low* (HML) est une prime de valeur. Elle représente l'écart de rendement entre les sociétés dont le ratio valeur comptable/valeur de marché est élevé et les sociétés dont le ratio valeur comptable/valeur de marché est faible. Le facteur HML révèle que, sur le long terme, les actions de valeur (ratio *book-to-market* élevé) bénéficient de rendements plus élevés que les actions de croissance (ratio *book-to-market* faible). Une fois le facteur SMB et le facteur HML déterminés, son coefficient bêta peut être calculé par régression linéaire utilisant la méthode des moindres carrés ordinaires. Utilisés séparément, les facteurs SMB et HML ont peu de pouvoir explicatif.

2.2. Facteurs implicites

Dans cette section, nous décrivons la méthode de régression avec des facteurs implicites. Nous avons déterminé la capacité des facteurs implicites à expliquer les log-rendements en utilisant une analyse des composantes principales (ACP). L'ACP est une technique statistique qui permet d'explorer des données dites multivariées (données avec plusieurs variables). Notre ensemble de données T x N forme un nuage de T points (nombre de rendements quotidiens incluent dans les données utilisées) dans un espace à N dimensions. L'ACP consiste à projeter les points sur un sous-espace à K dimensions (avec K≤ N). Ce sous-espace explique la plus grande proportion de variances avec le moins de composantes possible.

Prenons la matrice R_{TxN}, chaque ligne représente une liste de rendements pour N actifs au temps *t*, t=1,…,T et i=1,…,N. Les projections des *N* observations sur le vecteur z_1 sont données par les rendements r_i, i=1,…,N et les poids ϕ_{i1}:

$$z_1 = \sum_{i=1}^{N} \phi_{i1} r_i \qquad (4).$$

Le vecteur z_1 est la première composante, celle qui a la plus grande variance. Cette composante est une combinaison linéaire des rendements et des poids. Le vecteur en colonne de poids peut être réécrit ϕ_1. Les poids sont normalisés. C'est-à-dire, la somme des carrés des poids est égale à 1 ($\sum_{i=1}^{N} \phi_{i1}^2$ =1). Nous imposons cette contrainte, car si les poids n'avaient pas de limite en valeur absolue, la variance pourrait devenir arbitrairement grande. De façon concrète, le calcul de cette composante est un problème d'optimisation. Nous cherchons les poids ϕ_{i1} i=1,…,N qui maximisent la variance pour chaque élément z_{u1}, u=1,…,T qui sont éléments de z_1 où :

$$z_{u1} = \sum_{i=1}^{N} \phi_{i1} r_{ui} \qquad (5)$$

Nous avons donc :

$$\max_{\phi_{11},\phi_{21},\dots,\phi_{N1}} \left\{ \frac{1}{T} \sum_{u=1}^{T} \left(\sum_{i=1}^{N} \phi_{i1} r_{ui} \right)^2 \right\} \; sous\ contrainte \sum_{i=1}^{N} \phi_{i1}^2 = 1 \qquad (6)$$

Ou

$$\max_{\phi_{11},\phi_{21},\dots,\phi_{N1}} \left\{ \frac{1}{T} \sum_{u=1}^{T} z_{u1}^2 \right\} \; sous\ contrainte \sum_{i=1}^{N} \phi_{i1}^2 = 1 \qquad (7).$$

L'équation (*6*) peut être résolue en calculant les valeurs propres. Ensuite, nous calculons z_2. Dans le même ordre d'idée, nous avons :

$$z_{u2} = \sum_{i=1}^{N} \phi_{i2} r_{ui} \qquad (8).$$

Le calcul de z_2 est également un problème d'optimisation cherchant à maximiser la variance. La contrainte appliquée sur les poids demeure et nous ajoutons la contrainte que z_2 est indépendant de z_1 ce qui est équivalent de dire que ϕ_2 est indépendant de ϕ_1. L'équation (*6*) peut donc être réécrite ainsi :

$$\max_{\phi_{12},\phi_{22},\dots,\phi_{N2}} \left\{\frac{1}{T}\sum_{u=1}^{T}\left(\sum_{i=1}^{N} \phi_{i2} r_{ui}\right)^2\right\} sous\ contrainte \sum_{i=1}^{N} {\phi_{i2}}^2 = 1\ \ et\ \phi_2 \coprod \phi_1 (9).$$

De la même façon, nous pouvons calculer z_3,…, z_K. Nous avons ainsi la matrice Z de taille (TxK), la matrice de rendements R de taille (TxN) et la matrice de poids Φ de taille (NxK) tel que

$$Z = R\Phi \qquad (10).$$

Une fois les composantes calculées, il est possible de reconstruire l'image des données après la réduction de dimensionnalité de la façon suivante :

$$Z\Phi^T = R\Phi\Phi^T = R' \qquad (11).$$

Puisque nous ne conservons pas toutes les composantes, l'image R′ devrait être différente des données initiales R. Cette différence est appelée l'erreur de reconstruction. Pour évaluer la capacité de prévision de cette technique, nous avons calculé la valeur des composantes sur l'ensemble des données disponibles (entrainement et test) et nous avons uniquement conservé les poids calculés utilisant l'échantillon d'entrainement. Ainsi, nous avons reconstruit les données tests en utilisant les composantes calculées avec l'échantillon test et les poids calculés avec l'échantillon d'entrainement.

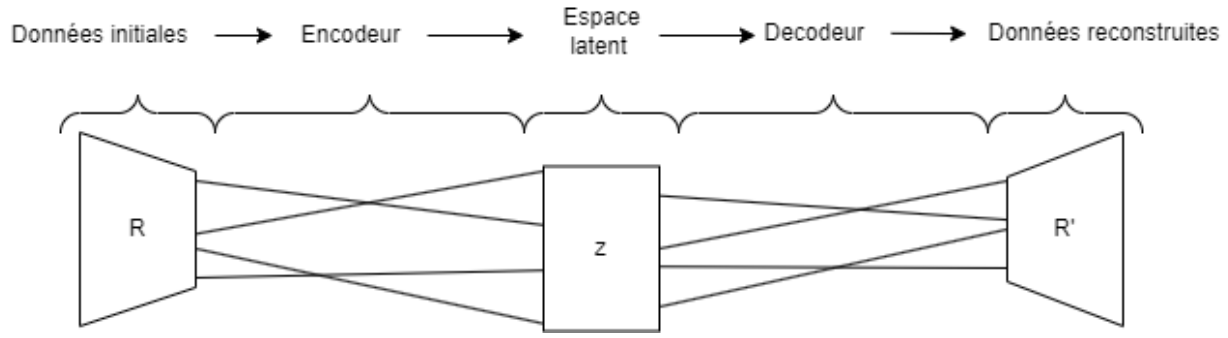
2.3. Apprentissage profond

Nous avons voulu explorer une méthode non linéaire basée sur un réseau neuronal dont l'objectif est de réduire la dimensionnalité. Pour la méthode d'apprentissage profond, nous avons utilisé un autoencodeur conditionnel.

2.3.1. Autoencodeur

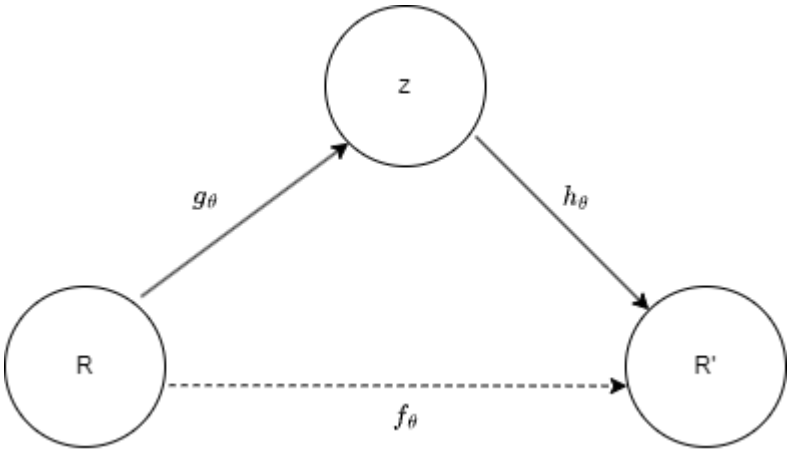
Dans cette section, nous décrivons le principe de base d'un autoencodeur. Puis, dans la section 2.3.2 nous expliquons la variante qu'est l'autoencodeur conditionnel. Un autoencodeur est une technique d'apprentissage non supervisé. Il utilise un réseau de neurones afin de reconstituer les données initiales. Comme nous pouvons l'observer dans la figure 1, la matrice de données entrantes R est projetée sur un espace latent z à dimension réduite, puis reconstituée (R') par le réseau neuronal. Il est primordial d'avoir un espace latent de dimension inférieure à l'espace initial afin d'éviter le surentrainement. En effet, si la dimension de l'espace z est trop grande, l'autoencodeur apprendra à reproduire l'ensemble de données sans avoir extrait l'information pertinente de la distribution des données. Une façon alternative de considérer ce problème est de voir l'image R' comme étant le résultat d'une transformation f_θ, où θ représente les paramètres de cette transformation. Cette transformation étant inconnue peut être approximée par une transformation g_θ, soit la fonction d'encodage puis transformée à nouveau par h_θ, soit la fonction de décodage tel qu'illustré dans la figure 2.

Figure 1 Structure de l'autoencodeur



Dans cette figure nous pouvons observer la structure d'autoencodeur utilisé afin de reconstruire les rendements

Figure 2 Représentation d'un autoencodeur à l'aide de fonction



Cette figure illustre la dynamique de la fonction f_θ qu'est un autoencodeur

2.3.2. Autoencodeur conditionnel

Pour cette technique, nous pouvons uniquement reproduire les rendements d'un vecteur fixe à la fois, car nous utilisons les rendements des autres actifs comme information pour reconstruire le rendement d'un actif. C'est-à-dire, nous devons entrainer le vecteur de rendements de chaque actif un par un. L'ensemble des données est utilisé pour entrainer le modèle et les rendements des autres actifs sont utilisés comme signal afin de reconstruire le vecteur de rendements d'un actif. Nous itérons l'entrainement pour tous les actifs et nous obtenons ainsi une reconstruction complète des données.

Pour un autoencodeur conditionnel la fonction f_θ est similaire à l'équation (1). Les données sont séparées en deux sous-ensembles : la série de rendements de l'actif *i* (r_i) et les *N-1* séries de rendements des autres actifs ($\widetilde{R}$). L'autoencodeur conditionnel utilise l'information de $\widetilde{R}$ afin de recomposer la série r_i. Nous définissons l'autoencodeur conditionnel par les équations suivantes :

$$Z_1 = \sigma\left(W_1{}^T \tilde{R} + B_1\right) \tag{12}$$

$$Z_2 = W_2 Z_1 + B_2 \tag{13}.$$

$W_1\tilde{\mathrm{R}} + B_1$ représente la couche *fully-connected*. Il s'agit d'une transformation linéaire des données initiales afin de réduire la dimensionnalité. Nous avons $\tilde{R}$ qui est la matrice des N-1 séries d'actifs, $\tilde{\mathrm{R}} \in \mathrm{R}^{\mathrm{T/2xN-1}}$. W_1, qui est de taille ($\frac{T}{2}$xK), est une matrice de poids qui sert à réduire la dimension, $W_1 \in \mathrm{R}^{\mathrm{T/2xK}}$. B_1, qui est de taille (Kx$N-1$), est le biais, $B_1 \in \mathrm{R}^{\mathrm{KxN-1}}$. Les poids initiaux sont générés de façon aléatoire. Nous avons choisi N-1>K=21, où 21 représente le nombre de jours où les marchés sont ouverts dans un mois. Ensuite, σ est la fonction d'activation, soit la fonction Unité Linéaire Rectifiée, $\sigma\left(W_1\tilde{R} + B_1\right) = max\left(0, W_1\tilde{R} + B_1\right)$. De la même manière, Z_2 est représenté par une transformation linéaire des données Z_1 afin de revenir à la taille désirée, soit une matrice de dimension N-1x$\frac{T}{2}$. Ainsi, W_2 est également une matrice de poids qui sert à réduire la dimension et B_2 est le biais.

La prochaine étape est d'encoder les rendements de l'actif *i* (r_i). De la même façon, nous réduisons la dimension initiale de r_{i} ($\frac{T}{2}$x1) à (Kx1) avec la couche *fully-connected.* Nous utilisons la même fonction d'activation σ. Ceci est illustré dans l'équation (*14*). Ensuite, nous retransformons les données afin d'obtenir la taille initiale, soit un vecteur de taille ($\frac{T}{2}$x1) tel que démontré par l'équation (*15*). La reconstruction r_{i}' est donc de la même dimension que r_{i}. L'équation (*16*) démontre l'étape finale afin de reconstruire r_{i}.

$$z_3 = \sigma(w_3 r_i + b_3) \tag{14}$$

$$z_4 = w_4 z_3 + b_4 \tag{15}$$

$$r_i' = z_4 \cdot Z_2 \tag{16}$$

Nous avons entrainé le réseau en utilisant la méthode de rétropropagation du gradient. Cette méthode consiste à mettre à jour le poids de chaque neurone en partant de la dernière vers la première. L'algorithme du gradient optimise le poids de chaque neurone afin de réduire l'erreur quadratique de reconstruction du modèle.

Afin d'évaluer la performance hors échantillon, nous avons dû alimenter le réseau avec le même nombre d'actifs et le même nombre d'observations par actif que pour les données d'entrainement. Les données ont donc été séparées de façon égale en termes de longueur de séries. Cette limitation du modèle est due au fait que ce n'est pas un modèle génératif. Il permet uniquement

de reconstruire les données. L'échantillon d'entrainement et l'échantillon test doivent donc avoir la même taille.

2.3.3. Autoencodeur variationnel

L'autoencodeur variationnel (*Variational autoencoder* ou VAE), bien qu'il soit similaire en termes de structure à un autoencodeur, est fondamentalement différent. Le VAE est un modèle génératif. Un modèle génératif a pour objectif de produire des nouvelles données réalistes en se basant sur les données d'entrainement. Tel que décrit par Pinheiro Cinelli et al. (2021), l'idée derrière ce modèle statistique est que nous avons une distribution paramétrique R pour laquelle il nous est impossible d'estimer les paramètres, donc impossible d'estimer la distribution marginale p(R) à partir des actifs $r_1, r_2, \ldots, r_N$. Ces variables ne sont pas indépendantes. La raison sous-jacente expliquant la dépendance peut être représentée par les variables latentes z. Nous obtenons ainsi la distribution conjointe p(R,z). La distribution marginale de R peut donc être écrite ainsi :

$$p(R) = \int p(R,z)\,dz = \int p(R|z)p(z)\,dz \approx \frac{1}{N}\sum_{i=1}^{N} p(R\,|z^{(i)}) \qquad (17).$$

Dans l'équation (*17*), z∼p(z) représente la distribution antérieure de l'espace latent et p(R|z) représente la fonction de vraisemblance. Une fois z estimés, nous allons pouvoir prédire des valeurs de R ou inférer sur la relation entre les données formant R. La partie gauche de l'équation (*17*) est insoluble et doit être approximée numériquement. Nous utilisons une simulation Monte-Carlo afin d'approximer la valeur de l'intégrale. Nous simulons R valeurs basées sur l'équation (*17*) S fois, s=1,…,S. Dans un espace latent à haute dimension, il est difficile d'avoir une estimation raisonnable. Nous aurions besoin de millions de tirages. Le choix de p(z) nous permettant d'obtenir une valeur plausible de z représente également un défi. Nous utilisons donc une technique de reparamétrisation et réécrivons l'équation (*17*) ainsi :

$$p(R) = \int p(\mathrm{R}|z)p(z)\,dz$$

$$= \int p(R|z)p(z)\frac{q(z|R)}{q(z|R)}dz$$

$$= \mathbb{E}_q\left[\frac{p(\mathrm{R}|z)p(z)}{q(z|R)}\right]$$

$$\approx \frac{1}{N}\sum_{i=1}^{N}\frac{p(R|z^{(i)})p(z^{(i)})}{q(z^{(i)}|R)} \qquad (18).$$

Où z∼q(z|R) et nous approximons l'intégrale avec une simulation de Monte-Carlo non biaisé de R tirages. Ceci correspond à l'approche d'échantillonnage préférentiel.

Avant de continuer plus loin, nous allons introduire deux concepts statistiques qui sont primordiaux à la suite des choses : divergence KL et ELBO. Nous avons supposé que z∼p(z) et que z∼q(z|R). Nous allons donc devoir évaluer et minimiser la dissemblance entre p(z) et $q(z|R)$. Pour ce faire, nous allons utiliser une mesure d'information relative qui évalue la dissemblance entre deux distributions, soit la divergence de Kullback-Leibler (KL divergence). Cette mesure est définie comme suit :

$$D_{KL}(q||p) = \int q(\epsilon)log(\frac{q(\epsilon)}{p(\epsilon)})d\epsilon \qquad (19).$$

ϵ représente une distribution générique de la dissemblance que nous estimons. Dans notre cas, nous estimons la dissemblance de la distribution z|R. De plus, nous avons une distribution conjointe p(R,z) où R sont les variables observables et z sont les variables latentes. La distribution a posteriori p(z|R) est intractable (i.e. elle ne peut pas être résolue). Il nous est donc impossible de minimiser la divergence KL. Nous allons devoir faire quelques manipulations algébriques :

$$D_{KL}(q(z|R)||p(z|R)) = \int q(z|R)log\left(\frac{q(z|R)}{p(z|R)}\right)dz$$

$$= -\int q(z|R)log\left(\frac{p(\mathrm{R},z)}{p(\mathrm{R})q(z|R)}\right)dz$$

$$= -\left(\int q(z|R)log\left(\frac{p(\mathrm{R},z)}{q(z|R)}\right)dz - \int q(z|R)\ log(p(\mathrm{R}))dz\right)$$

$$= -\int q(z|R)log\left(\frac{p(R,z)}{q(z|R)}\right)dz + log(p(R))\int q(z|R)\ dz$$

$$= -\mathbb{E}_q\left[log\left(\frac{p(R,z)}{q(z|R)}\right)\right] + log(p(R)) \qquad (20).$$

Nous pouvons donc réécrire l'équation (*20*) ainsi:

$$log(p(R)) = \mathbb{E}_q\left[log\left(\frac{p(R,z)}{q(z|R)}\right)\right] + D_{KL}(q(z|R)||p(z|R)) \qquad (21).$$

Nous savons que $D_{KL}(q(z|R)||p(z|R)) \geq 0$. Le deuxième terme de droite est donc une borne inférieure pour $log(p(R))$. Pour cette raison, ce terme est nommé *Evidence Lower Bound* (ELBO). Ceci nous mène à un résultat important. En effet, puisque $log(p(r))$ est une valeur fixe, en maximisant ELBO nous minimisons $D_{KL}(q(z|R)||p(z|R))$.

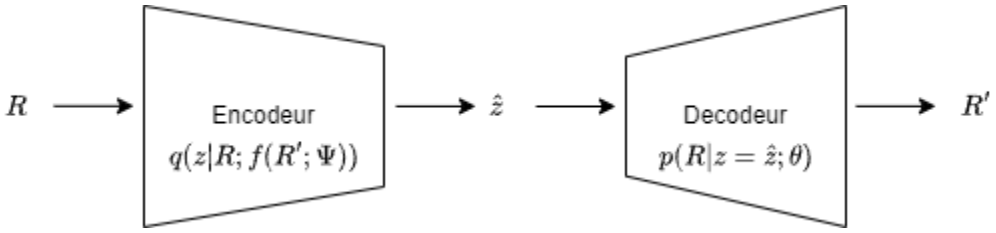
Maintenant que ces concepts ont été introduits, nous allons paramétrer la distribution p et q. Θ représente les paramètres de la distribution p et Φ représente les paramètres de la distribution q. Nous pouvons donc réécrire les distributions ainsi : $p(R|z;\Theta)$ et $q(z|R;\Phi)$. Nous devons optimiser Φ, Θ conjointement. La distribution a posteriori $q(z|R;\Phi)$ nous permet d'inférer sur la distribution latente des observations. La fonction de vraisemblance $p(R|z;\Theta)$ nous permet de générer un nouvel échantillon en appliquant la loi de probabilité a priori. Ceci implique donc que nous tirons une valeur de la distribution conjointe. Afin d'optimiser les paramètres Θ, nous utilisons la méthode de maximum de vraisemblance. Afin d'optimiser les paramètres Φ, nous aurions besoin de calculer l'estimer de $r_i \in \mathcal{D}$ pour chacun des tirages. $\mathcal{D}$ représente l'ensemble des observations. À défaut d'utiliser cette méthode, nous allons optimiser un modèle séparé, nommé modèle de reconnaissance (ce modèle est le réseau de neurones en soi), afin d'obtenir les paramètres de variations locaux Φ qui définissent la distribution a posteriori $q(z|R;\Phi)$. Chacun des nouveaux rendements r' sont transformés par la fonction f avec les paramètres Ψ, soit f(R' ; Ψ) $\rightarrow \Phi$. De cette façon, le problème devient un problème d'optimisation globale des paramètres variationnels Ψ. Nous notons que la valeur optimale de Φ est telle que $q(z|R;\Phi) = p(R|z;\Theta)$. Nous devons donc maximiser $p(R;\Theta)$. En utilisant l'équation (*18*), nous obtenons :

$$\max_{\Theta} p(R;\Theta) = \max_{\Theta} log\ p(R;\Theta) = \max_{\Theta} log\mathbb{E}_q\left[\frac{p(R|z;\Theta)p(z)}{q(z|R;\Phi)}\right] \qquad (22).$$

En utilisant l'inégalité de Jensen nous obtenons le ELBO

$$log\mathbb{E}_q\left[\frac{p(R|z;\Theta)p(z)}{q(z|R;\Phi)}\right] \geq \mathbb{E}_q\left[log\frac{p(R|z;\Theta)p(z)}{q(z|R;\Phi)}\right] = ELBO(\Theta,\Phi) \qquad (23)$$

Figure 3 Structure du VAE



Dans cette figure nous observons la structure du VAE où l'image r_i est transformée par l'encodeur du réseau de neurones. Cette transformation nous permet d'obtenir la variable latente z. Nous tirons une valeur $\hat{z}$ en simulant une valeur de la distribution de z. Ensuite, nous décodons la valeur simulée et nous obtenons la distribution $p(R|z = \hat{z}; \Theta)$ pour laquelle r_i est la valeur la plus probable

Nous pouvons comprendre de la figure 3 que $q(z|R; \mathrm{f(R'\,;\,\Psi)})$ est un encodeur probabiliste et $p(R|z = \hat{z}; \Theta)$ est un décodeur probabiliste. En effet, nous pouvons réécrire le ELBO pour un ensemble de données $\in \mathcal{D}$ avec R tirages nous avons :

$$ELBO(\Theta, \Psi) = \sum_{n=1}^{N} \mathbb{E}_q[log(p(r_n|z_n; \Theta)] - D_{KL}(q(z_n|r_n; \Psi)||p(z)) \qquad (24).$$

Le premier terme de droite a pour objectif de maximiser la vraisemblance de l'échantillon reconstruit alors que le second terme impose une structure à l'espace latent. Cette contrainte sur la structure de l'espace latent est simple ; il faut que la distribution conditionnelle à r_i soit le plus similaire possible à la distribution a priori. La divergence KL impose une structure qui différencie d'un autoencodeur traditionnel. En effet, si nous omettons la divergence KL, la fonction d'optimisation devient une optimisation du maximum de vraisemblance et nous obtenons une convergence à un point. Il serait donc possible de trouver la fonction et reconstruire de façon déterministique la transformation de h(r)→ Z.

La distribution latente entre l'encodeur et le décodeur pose un problème lors de l'optimisation avec la méthode du gradient conjugué. Il n'est pas possible de calculer numériquement le gradient d'une espérance relativement à une distribution. Nous allons utiliser une technique de reparamétrisation (*reparametrization trick*) pour lequel S simulations MC nous obtenons :

$$\widehat{ELBO}_1(\Theta, \Psi) = \sum_{n=1}^{N} \left[\frac{1}{S}\sum_{s=1}^{S} log\ p(r_n, z_n^{(s)}; \Theta) - \ log\ q(z_n^{(s)}|r_n; \Psi)\right] \qquad (25).$$

Où $z_n^{(i)} = g(\epsilon^{(i)}, r_n;\ \Psi)$ est une transformation déterministique et $\epsilon^{(i)}$ est le i^e^ échantillon de la distribution de base p(ϵ). Il est possible de choisir une famille de distributions pour représenter p(z) et q(z|R ; Ψ) tel qu'il existe une forme analytique à la divergence KL nous permettant ainsi de réécrire l'équation (25) de cette façon :

$$\widehat{ELBO}_2(\Theta, \Psi) = \sum_{n=1}^{N} \left[\frac{1}{S}\left[\sum_{s=1}^{S} log\ p(r_n|z_n^{(s)}; \Theta)\right] - D_{KL}(q(z_n|r_n; \Psi)||p(z))\right] \quad (26).$$

Dans notre cas, nous supposons que les rendements sont normaux et utilisons une fonction de vraisemblance $\mathcal{N}(\mu_i, \sigma_i^2)$ pour décrire p(R|z ; Θ) et une distribution normale diagonale centrée $\mathcal{N}(0, \mathrm{I})$ pour la distribution à priori p(z). Ainsi la divergence KL se simplifie comme suit :

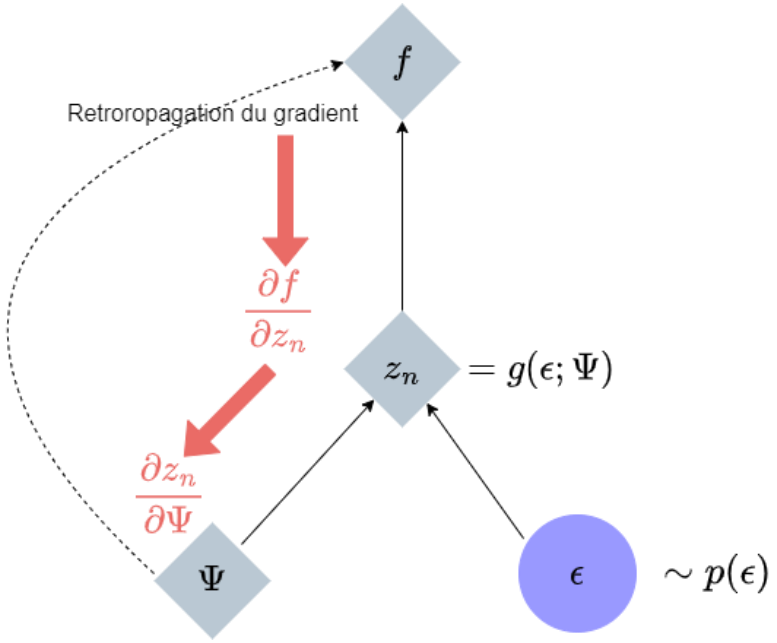
$$D_{KL}(q(z_n|r_n; \Psi)||p(z)) = \sum_i^{|z|} log\frac{1}{\sigma_i} + \frac{1}{2}(\mu_i^2 + \sigma_i^2 - 1) \quad (27).$$

De plus, la distribution a posteriori $q(z_i|R_i; \Psi)$ est une distribution gaussienne avec une matrice de covariance diagonale. La transformation déterministique $g(\epsilon, r_n;\ \Psi)$ est donc défini ainsi :

$$g(\epsilon, r;\ \Psi) = \mu(f(r; \Psi)) + \sigma(f(r; \Psi)) \odot \epsilon \quad (28).$$

Où $\epsilon \sim \mathcal{N}(0, \mathrm{I})$ et $\odot$ est l'opérateur du produit matriciel d'Hadamard.

Figure 4 Rétropropagation du gradient



Cette figure illustre le principe de rétropropagation du gradient utilisé afin d'optimiser les paramètres du VAE

Nous avons entrainé le réseau en utilisant la méthode de rétropropagation du gradient. La figure 4 illustre la méthode utilisée afin de mettre à jour le poids de chaque neurone en partant de la dernière vers la première. L'algorithme du gradient optimise le poids de chaque neurone afin de réduire l'erreur quadratique de reconstruction du modèle.

Afin d'évaluer la performance hors échantillon, tel que pour les techniques factorielles, nous avons séparé les données en deux sous-échantillons : l'échantillon d'entrainement représentant 70% des données et l'échantillon test représentant 30% des données. Le modèle est entrainé sur 90% des données d'entrainement et les paramètres sont calibrés sur les 10% restant de ce sous-ensemble. Cette séparation est unique aux autoencodeurs variationnels. Cette méthode est la seule qui requiert une période de calibration.

3. Données

Dans cette section, nous détaillons le processus de collecte de données. Nous avons utilisé deux sources de données. La première est le prix ajusté à la fermeture des marchés des constituants de l'indice S&P500 du 4 janvier 2010 au 31 décembre 2021. La liste des constituants est celle en date du 30 juin 2022. La seconde est l'ensemble de données du modèle de 13 facteurs de marché construit et estimé par Jensen et al (2023) (JKP). Les données de JKP couvrent la même période. Ces facteurs sont présentés sur le site web de Bryan Kelly et sont une reproduction des facteurs traditionnels. La fréquence des données est quotidienne.

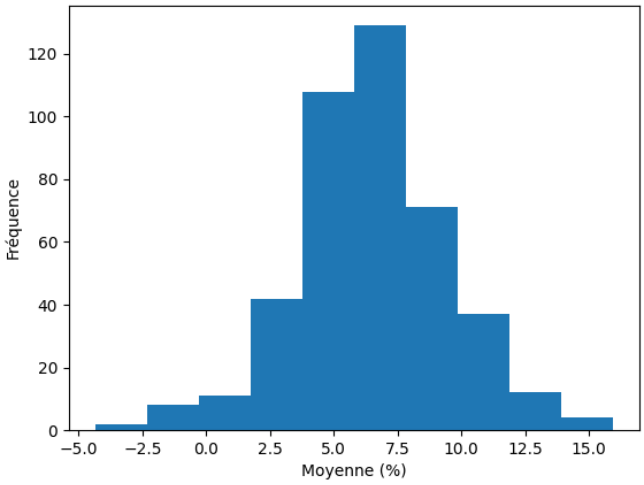
3.1. S&P500

La liste des entreprises qui composent l'indice du S&P500 a été formée à partir de Wikipédia en date du 30 juin 2022 et les données pour chacun des actifs ont été extraites de Yahoo Finance. Nous avons éliminé tous les titres ayant des données manquantes entre 2010 et 2021 afin de simplifier l'analyse ultérieure en évitant les difficultés liées aux valeurs manquantes. Ceci inclut également les titres ayant changé de symbole boursier à la suite d'une fusion ou d'une acquisition, par exemple. Nous avons une liste finale de 424 titres. Nous avons exclu un peu plus

de 70 titres. Nous avons calculé le log-rendement de chacun de ces titres sur la période et ainsi obtenu un total de 3019 rendements.

Dans la figure 5 nous pouvons observer la distribution des moyennes des rendements annualisés par actif sur la période 2010-2021. Comme nous pouvons l'observer, la plupart des actifs ont des rendements moyens supérieurs à zéro. La moyenne de rendements quotidiens pour la plupart des actifs formant l'univers du S&P500 se situe entre 2.5% et 10% sur une base annuelle. En comparaison, l'indice du S&P500 a une moyenne de rendements annualisés de 11.88% dans les dix dernières années. Dans notre cas, nous comparons tous les titres de façon égale. L'indice quant à lui prend en compte la capitalisation boursière des titres, ce qui explique la différence dans le rendement. En effet, plus la capitalisation d'une compagnie est grande, plus les rendements de cette dernière influenceront le rendement de l'indice. Par exemple, si une compagnie telle que Meta ou Google obtient des rendements supérieurs à la moyenne des rendements individuels, le rendement général de l'indice sera plus élevé que la moyenne des rendements individuels. Nous observons une asymétrie vers la gauche avec un coefficient de -0.62.

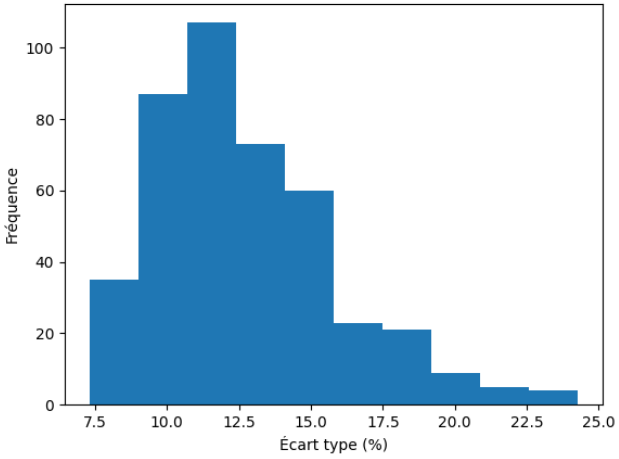
Figure 5 Moyenne annualisée des rendements



Cette figure illustre la moyenne des rendements de 424 actifs du S&P500 actifs sur la période 2010-2021. Ces actifs représentent un sous-ensemble de la liste des 500 titres formant l'indice S&P500. Aucun des actifs n'a de données manquantes. Les titres ayant été acquis ou ayant pris part à une fusion ont été exclus.

Dans la figure 6 nous pouvons observer la distribution des écarts types des rendements par actif sur la période 2010-2021. L'écart type annualisé moyen des rendements dans l'univers d'actifs se situe entre 7.5% et 17.5%. Considérant que le rendement moyen se situe entre 2.5% et 10%, nous observons une grande variabilité dans les rendements quotidiens sur cette période pour tous les titres.

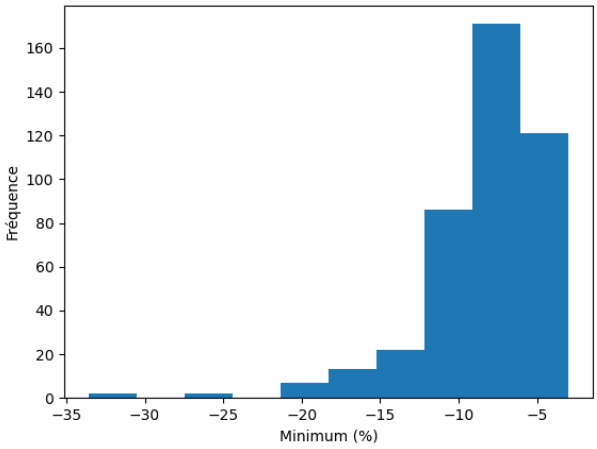
Figure 6 Écart type annualisé des rendements par actif



Cette figure illustre l'écart type des rendements de 424 actifs du S&P500 sur la période 2010-2021. Ces actifs représentent un sous-ensemble de la liste des 500 titres formant l'indice S&P500. Aucun des actifs n'a de données manquantes. Les titres ayant été acquis ou ayant pris part à une fusion ont été exclus.

Nous avons décidé de pousser notre analyse et d'observer le rendement minimum et maximum de chacun des actifs afin de déterminer si nous avions des données extrêmes qui pourraient impacter notre analyse. Nous pouvons observer dans la figure 7 un rendement particulièrement extrême de -33.60%. En effet, ce rendement de la compagnie APA, une compagnie du secteur de l'énergie, en date du 9 mars 2020, est dû à la crise de la Covid-19.

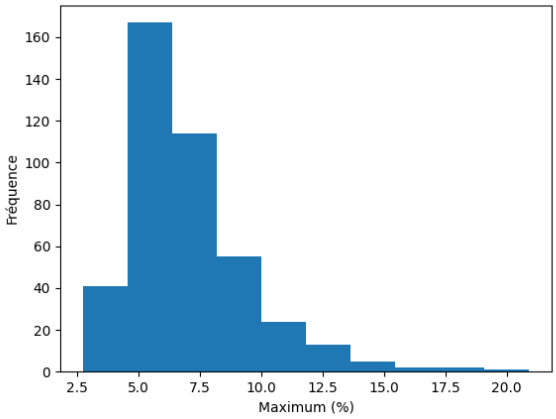
Figure 7 Rendements minimum par actif



Cette figure illustre les valeurs minimales des rendements de 424 actifs du S&P500 sur la période 2010-2021. Ces titres figuraient dans la liste des titres formant l'indice S&P500 en date du 30 juin 2022. Aucun des actifs n'a de données manquantes. Les titres ayant été acquis ou ayant pris part à une fusion ont été exclus.

Dans la figure 8 nous pouvons observer les rendements maximums par actifs. Comme attendu, nous notons une forte concentration des valeurs entre 2.5% et 10%. Nous jugeons ces valeurs comme plausibles en temps de redressement à la suite de la crise de la Covid-19.

Figure 8 Rendements maximum par actif



Cette figure illustre les valeurs maximales des rendements de 424 actifs du S&P500 sur la période 2010-2021. Ces titres figuraient dans la liste des titres formant l'indice S&P500 en date du 30 juin 2022. Aucun des actifs n'a de données manquantes. Les titres ayant été acquis ou ayant pris part à une fusion ont été exclus.

Afin d'alimenter la discussion des résultats et de les analyser de façon plus granulaire, nous avons résumé les rendements des actifs par secteur dans le tableau 1. Nous avons utilisé les secteurs définis par la classification GICS (*Global Industry Classification Standard*). Nous pouvons observer que le secteur des services et le secteur des biens de consommation ont la plus petite variance et le secteur de l'énergie a une moyenne de rendement nettement inférieur aux autres secteurs. Il est également le secteur le plus volatile. Il est important de noter le faible nombre de titres pour ce secteur. Un échantillon plus petit est moins crédible, car il peut être biaisé par des valeurs extrêmes. Le secteur de l'énergie est également un secteur complexe qui comprend des compagnies produisant, transformant ou transportant des énergies fossiles. Le 21e siècle est marqué par une transition énergétique délaissant ces énergies fossiles. Ces compagnies doivent donc diversifier leurs activités. La transition n'est pas simple étant donnée notre dépendance à ce genre d'énergie. Nous voyons donc une grande volatilité dans ce secteur et la moyenne des rendements pour ce secteur peut être plus faible.

Tableau 1 Statistiques descriptives des rendements des 424 titres par secteur entre 2010 et 2021

Secteur	Nombre d'actif	Moyenne annualisée (%)	Écart type annualisé (%)	Minimum (%)	Maximum (%)
Services de communication	16	5.50	14.10	-19.78	15.30
Biens de consommation	50	5.68	14.96	-25.80	14.39
Biens de base	28	5.23	9.65	-10.30	13.75
Énergie	15	2.48	17.92	-33.60	12.84
Financier	60	5.12	13.10	-13.57	11.96
Santé	55	6.39	13.03	-17.17	20.93
Industriel	59	5.74	12.85	-15.67	14.95
Technologie	61	6.63	14.12	-17.46	18.27
Matériaux	22	4.67	13.76	-13.97	11.29
Immobilier	30	4.79	12.05	-17.44	13.03
Services	28	4.75	9.65	-10.19	11.19

Ce tableau présente les statistiques descriptives des 424 titres du S&P500 disponible entre 2010 et 2021 par secteur. Ces titres figuraient dans la liste des titres formant l'indice S&P500 en date du 30 juin 2022. Aucun des actifs n'a de données manquantes. Les titres ayant été acquis ou ayant pris part à une fusion ont été exclus.

3.2. Facteurs explicites

Pour ce projet, nous avons utilisé deux séries de facteurs explicites. En premier lieu, nous avons utilisé les facteurs de marché, de taille et de valeur de Fama French (1993) et les facteurs de profitabilité et d'investissement de Fama French (2015). Ces facteurs sont disponibles sur le site de Kenneth French. Les données utilisées couvrent la période du 1er janvier 2010 au 31 décembre 2021.

En second lieu, les chercheurs et professeurs Jensen et al (2023) (JKP) proposent une liste de 13 facteurs explicites. La définition de ces facteurs se base sur un modèle hiérarchique bayésien. Dans ce modèle statistique, plusieurs distributions sont « emboitées » de façon hiérarchique. L'objectif est d'utiliser plusieurs indicateurs (ratio comptable et indicateurs de marché) pour créer des groupes représentant les facteurs les plus importants. Dans l'analyse globale, ils observent 153 indicateurs (p. ex.: *book-to-market ratio, dollar volume*) répartis dans 93 régions du monde. Ces indicateurs peuvent être regroupés en *J* groupes. Ces groupes sont appelés des thèmes. Au total, nous comptons 13 thèmes. Ces thèmes sont : provision (*accrual*), émission de dette (*debt issuance*), investissement (*investment*), faible ratio de levier (*low leverage*), faible risque (*low risk*), momentum (*momentum*), croissance des bénéfices (*growth profit*), profitabilité (*profitability*), qualité (*quality*), saisonnalité (*seasonality*), taille (*size*), inversion à court terme (*short-term reversal*) et valeur (*value*). Dans la figure 10, nous observons que pour chacun des thèmes, il existe une forte corrélation entre chacun des indicateurs. Entre les différents thèmes, nous observons une faible corrélation. Dans le tableau 2, nous pouvons observer que les facteurs ont une moyenne annualisée entre 0% et 4% et un écart type entre 1% et 5%.

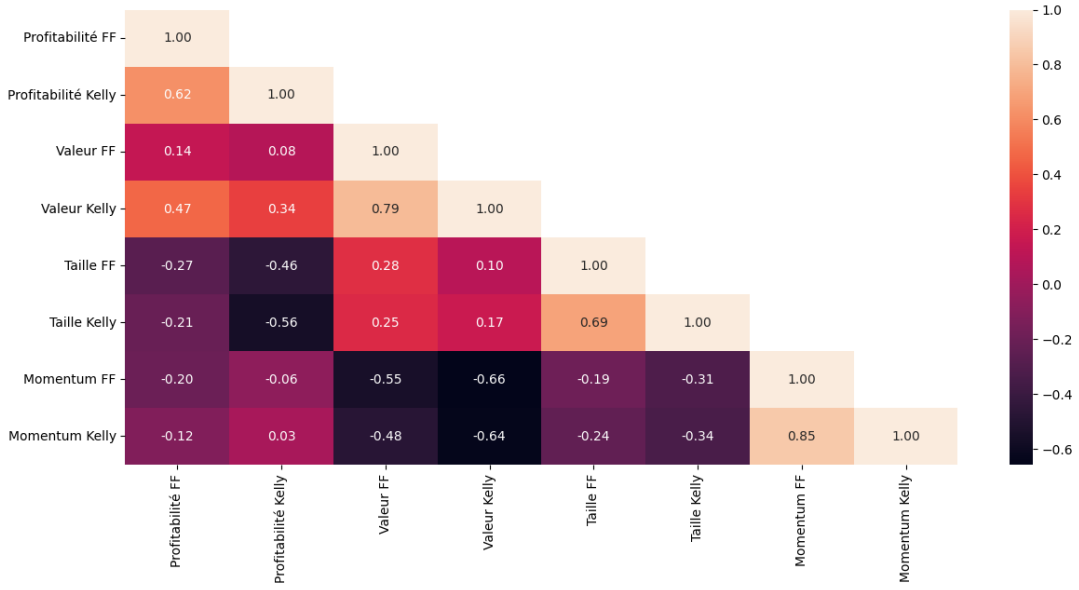
Tableau 2 Statistiques descriptives des facteurs de Bryan Kelly

Thème	Moyenne annualisée (%)	Écart type annualisé (%)	Minimum (%)	Maximum (%)
Provisions (*Accruals*)	0.81	1.75	-0.48	0.81
Émission de dette (*Debt Issuance*)	1.03	1.09	-0.35	0.36
Investissement	0.39	3.28	-1.21	2.40
Faible levier (*Low leverage*)	1.24	4.06	-1.99	1.40
Faible risque (*Low risk*)	2.14	6.08	-1.87	2.08
Momentum	4.49	6.29	-5.43	2.38
Croissance des bénéfices (*Profit Growth*)	1.67	2.12	-2.71	0.69
Profitabilité	2.31	3.11	-1.31	1.15
Qualité	4.01	3.47	-3.39	1.38
Saisonnalité (*Seasonality*)	0.28	0.99	-0.55	0.53
Taille	0.94	3.50	-1.34	1.61
Inversion à court terme (*short-term reversal*)	1.12	2.07	-0.67	0.81
Valeur	0.31	5.48	-1.89	3.29

Ce tableau présente les statistiques descriptives de la série de 13 facteurs explicites estimés par Bryan Kelly avec des indicateurs économiques et ratio comptable sur la période 2010-2021. Cette série peut être obtenue sur le site de Bryan Kelly.

Nous avons évalué la corrélation de Pearson entre les facteurs de Fama French et ceux de JKP. Nous observons les détails dans la figure 9. Une forte corrélation (>0.6) pour les facteurs similaires et une faible corrélation avec les autres facteurs. Une telle corrélation est désirable. En effet, si les facteurs sont décorrélés, nous évitons ainsi les problèmes de colinéarités.

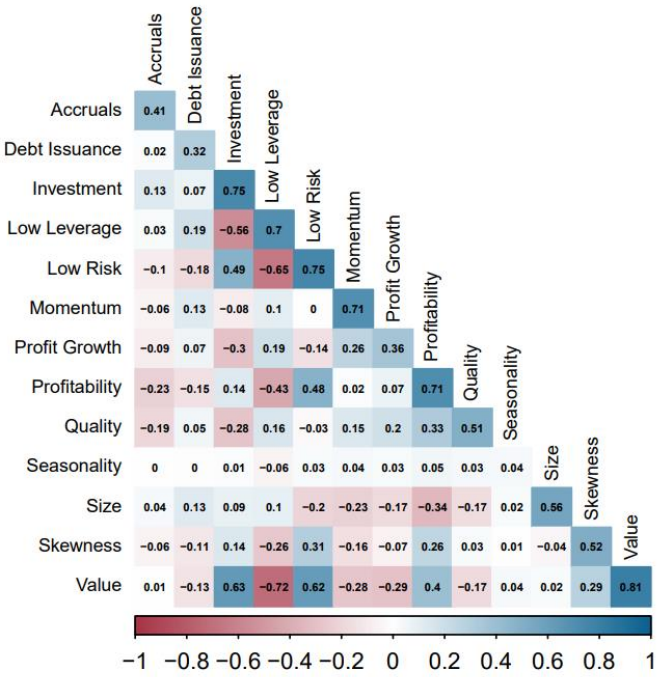
Figure 9 Corrélation facteurs explicites de Fama French et de Bryan Kelly



Cette figure représente la corrélation entre les facteurs explicites de Fama French et les facteurs explicites de Bryan Kelly sur la période 2010-2021

La corrélation entre les facteurs de Fama French et les facteurs de JKP n'est pas parfaite. Ceci peut être expliqué par la façon dont les facteurs de JKP sont construits. En effet, chaque facteur représente une distribution a posteriori construite à partir des indicateurs. Les facteurs ne sont, en moyenne, pas parfaitement corrélés avec les indicateurs. Dans la figure 10 provenant de l'article de JKP détaillant la construction des facteurs, nous observons également que la corrélation entre les facteurs de Fama French et ceux de JKP est comparable avec la moyenne des corrélations entre les facteurs de JKP et les indicateurs.

Figure 10 Moyenne des corrélations entre les indicateurs et les facteurs de Bryan Kelly



Cette figure est extraite de l'article de Jensen et al. (2023). Elle représente la moyenne des corrélations de chaque indicateur avec le facteur modélisé.

4. Résultats

Dans cette section, nous présentons les résultats des méthodes décrites dans la section 2. Nous avons tenté de déterminer si les facteurs de Fama French sont les meilleurs facteurs explicites ou si les facteurs de JKP peuvent obtenir de meilleurs résultats. Il fallait d'abord déterminer la meilleure combinaison des facteurs décrits dans la section 3.2. Nous avons donc comparé la capacité explicative de toutes les combinaisons de trois et de cinq facteurs formés à partir des treize facteurs proposés sur le site de Bryan Kelly. Pour la régression à trois facteurs, nous avons analysé les 286 combinaisons possibles et pour la régression à cinq facteurs nous avons analysé les 1287 combinaisons possibles. Afin de déterminer la meilleure combinaison, nous avons calibré les coefficients de régression en utilisant les données d'entrainement, ensuite nous avons calculé les coefficients de détermination linéaire R^2 décrits dans l'équation (*2*). La combinaison

de facteurs ayant le coefficient le plus élevé est donc considérée comme étant la meilleure. La meilleure combinaison pour la régression à trois facteurs est composée des facteurs de : Faible risque, profitabilité et qualité. Pour la régression à cinq facteurs, nous ajoutons les facteurs de : faible levier et momentum.

Pour la méthode ACP, nous avons comparé la capacité de prévision avec les méthodes précédentes en gardant les trois et cinq composantes principales. Nous avons également poussé l'analyse plus loin en gardant 50, 100 et 200 composantes afin de comparer la différence avec un grand nombre de composantes. Nous avons calculé le R^2_{OOS} utilisant l'équation (*3*) pour les résultats utilisant les données de l'échantillon test pour toutes les techniques de modélisation.

Puisque le modèle est calibré à partir des données d'entrainement maximisant le R^2, le R^2_{OOS} obtenu avec le modèle des données test devrait être plus petit dû à l'erreur de prédiction. De plus, en théorie, il ne devrait pas être négatif. La somme de l'erreur du modèle élevée au carré devrait, quant à elle, être au moins plus petite que la somme du carré des log-rendements. Dans le cas contraire, les prévisions du modèle seraient erronées à un point tel que nous pourrions supposer que tous les log-rendements sont nuls et que notre prévision serait meilleure que celle du modèle.

Nous décrivons les résultats des valeurs des R^2 avec l'échantillon d'entrainement dans la section 4.1. Nous allons également comparer nos résultats avec ceux obtenus par Gu et al. (2021) pour nos données hors échantillon. Dans leur article, ils utilisent des rendements mensuels. Leur échantillon test inclut des données de la période de janvier 1987 jusqu'à décembre 2016. Nous obtenons des résultats favorables aux leurs. Ceci peut être dû au fait que nous utilisons des rendements journaliers. L'intervalle de données étant plus petites, la distance entre les valeurs prédites et les valeurs réelles est plus petite, ainsi expliquant nos meilleurs résultats. Leur période de test inclut également deux crises financières. En comparaison, notre fenêtre de test inclut uniquement la crise financière du Covid. Cette crise a eu une relance économique beaucoup plus rapide que les autres. Notre fenêtre d'estimation est donc peut-être plus favorable, expliquant ainsi nos résultats. Il est important de noter que nous avons pu obtenir nos résultats pour cette technique en adaptant un code Github reproduisant la technique de cet article.[2]

[2] Ce code est disponible sur le github publique de D. Kyle Miller https://github.com/dkyol/Asset-Pricing-Model.

4.1. Période d'entrainement

Dans la section suivante, nous allons présenter nos résultats sur l'échantillon d'entrainement. Pour les méthodes utilisant des réseaux de neurones, le modèle utilise une méthode de descente de gradient afin de trouver l'optimum local du R^2. Pour l'autoencodeur conditionnel nous obtenons donc des résultats impressionnants avec un R^2 de plus de 99.5% pour tous les secteurs économiques. Cette technique est conçue pour reconstruire les données entrantes le mieux possible. Cette technique n'est pas nécessairement conçue dans l'optique de prédiction. Nous avons donc du surentrainement (*overfitting*). Nous observons également, en général, une meilleure performance pour les modèles ayant plus de facteurs. De plus, le R^2 est plus élevé que le R^2_{OOS} sauf pour le modèle ACP à 3 et 5 facteurs. Bien que ces exceptions ne soient pas désirables, la différence est assez négligeable surtout lorsque nous comparons avec les résultats du modèle avec 100 ou 200 facteurs. Pour les résultats globaux, nous n'observons pas de valeur négative, ce qui est attendu. Toutefois, pour le VAE nous observons des valeurs négatives pour deux secteurs. Bien que cette anomalie ne soit pas désirable, elle n'implique pas un problème dans l'implémentation de la technique.

Tableau 3 Comparaison des R^2 estimés 1/2010-5/2018

	Régression facteurs Bryan Kelly		Régression facteurs Fama-French		ACP					AE	VAE
Nombre de facteurs	3	5	3	5	3	5	50	100	200		
Tout	28.86%	30.36%	46.08%	46.45%	48.17%	50.72%	71.45%	81.21%	91.95%	99.85%	45.15%
Services de communication	18.84%	19.66%	28.34%	28.17%	29.78%	30.33%	76.64%	88.14%	96.03%	99.88%	53.01%
Biens de consommation	25.51%	27.96%	40.05%	40.59%	41.62%	42.30%	70.29%	82.86%	95.12%	99.89%	58.41%
Biens de base	14.27%	16.66%	33.47%	32.22%	34.90%	35.39%	48.57%	62.06%	78.47%	99.58%	-32.11%
Énergie	43.10%	43.76%	54.63%	52.47%	61.41%	65.15%	74.75%	84.18%	93.84%	99.92%	63.15%
Financier	36.62%	37.69%	62.26%	63.75%	59.65%	64.78%	72.29%	79.58%	89.07%	99.84%	38.90%
Santé	18.83%	19.76%	33.61%	34.21%	34.18%	38.40%	72.15%	82.65%	93.17%	99.85%	53.64%
Industriel	32.92%	34.74%	52.66%	52.55%	54.98%	56.60%	71.62%	78.98%	90.77%	99.89%	56.60%
Technologie	30.09%	30.83%	43.06%	43.69%	45.89%	48.49%	71.00%	83.08%	94.58%	99.88%	56.43%
Matériaux	36.46%	37.85%	49.70%	48.87%	52.62%	54.29%	73.26%	81.86%	92.57%	99.82%	33.92%
Immobilier	28.12%	31.06%	51.96%	51.88%	61.38%	63.10%	74.58%	81.96%	90.24%	99.76%	26.05%
Services	20.17%	24.17%	43.79%	46.21%	55.32%	56.51%	70.70%	74.17%	80.52%	99.57%	-67.95%

Ce tableau illustre les valeurs des R^2 sur la période d'entrainement s'échelonnant du 4 janvier 2010 au 29 mai 2018 pour les techniques analysées sur les 424 actifs du S&P500 disponible entre 2010 et 2021 par secteur d'activité. Ces titres figuraient dans la liste des titres formant l'indice S&P500 en date du 30 juin 2022. Aucun des actifs n'a de données manquantes. Les titres ayant été acquis ou ayant pris part à une fusion ont été exclus.

4.2. Période de test complète

Dans cette section, nous présentons les résultats des méthodes sur la période complète de l'échantillon test, soit du 30 mai 2018 au 31 décembre 2021. Dans un premier temps, dans le tableau 4 nous avons les résultats sur l'ensemble des actifs, puis nous avons calculé la métrique par secteur d'activité. Comme attendu, pour les modèles factoriels comme la régression linéaire utilisant les facteurs de Fama French ou l'ACP nous obtenons, en général, de meilleurs résultats en ajoutant des facteurs. À l'exception de la régression utilisant les facteurs de JKP pour les actifs du secteur des biens de base et du R^2 global pour la régression utilisant les facteurs de Fama French, l'ajout de facteurs apporte un gain sur la performance du modèle. Nous avons même des résultats supérieurs à 90% pour l'ACP à 200 facteurs. Cela dit, le nombre imposant de facteurs requis pour atteindre ce niveau de performance est considérable. Ce nombre de facteurs représente une réduction de dimensionnalité d'un peu plus de 50%. Bien que cette réduction soit remarquable, une telle dimensionnalité est tout de même substantielle. Ce que notre mesure ne capte pas, c'est l'ajout de variables dans notre régression. Un facteur de pénalité aurait pu être ajouté afin d'incorporer cette subtilité dans les résultats. Un R^2 ajusté est un exemple de ce genre de pénalité.

Nous observons également une différence importante entre les régressions utilisant les facteurs de JKP et les régressions utilisant les facteurs de Fama French. Celles utilisant les facteurs de Fama French performent largement mieux. Considérant la forte corrélation entre les deux types de facteurs, ce résultat est étonnant. Il est intéressant de noter que les résultats sont supérieurs à ceux de l'article de Gu et al. (2021), qui quant à eux avaient obtenu un R^2 pour les actions individuelles de 3.4% et -2.3% pour 3 et 5 facteurs Fama French. Ceci peut être dû à la différence de temps dans les fenêtres de données testées ou à la fréquence des données.

Les résultats des régressions utilisant les facteurs de Fama French sont dans la même lignée que ceux de l'ACP. Bien que l'ACP soit plus performant, l'écart est beaucoup moins notable qu'avec les régressions utilisant les facteurs de Bryan Kelly. Les résultats de la méthode ACP sont les meilleurs, tous secteurs confondus. Ils sont également mieux que ceux de Gu et al. (2021) qui, quant à eux, obtiennent un R^2 de 5.0% et 4.2% pour 3 et 5 facteurs.

Les résultats les plus décevants sont toutefois ceux de l'autoencodeur conditionnel. En effet, nous observons des valeurs négatives pour cette méthode. Ces résultats indiquent une performance médiocre du modèle. Un modèle naïf proposant des valeurs nulles pour les rendements nous aurait donné un meilleur résultat. Une analyse plus approfondie des résultats nous révèle une corrélation de 5.78% entre les rendements prédits par le modèle et les rendements réels. Cette faible corrélation explique un tel résultat. En effet, les prédictions étant décorrélées des valeurs réelles, il n'existe aucun lien entre ces deux ensembles de données. Notre mesure de performance étant basée sur la distance entre les valeurs prédites et les valeurs réelles ne peut pas obtenir de résultat favorable si les valeurs prédites sont décorrélées des valeurs réelles. Ceci va bien évidemment à l'encontre de l'objectif de modélisation. Nous aurions dû observer une forte corrélation.

Les résultats du VAE sont mieux que ceux de l'autoencodeur conditionnel. En comparaison aux autres techniques, il se situe entre les régressions utilisant les facteurs des JKP et celles utilisant les facteurs de Fama French. Cette technique est également beaucoup plus efficace en termes de temps de calcul. En effet, puisque nous avons la possibilité de prédire les rendements pour tous les actifs en comparaison à l'autoencodeur conditionnel qui ne peut que prédire un vecteur.

Tableau 4:Comparaison des R^2_{OOS} estimés 05/2018-12/2021

	Régression facteurs Bryan Kelly		Régression facteurs Fama-French		ACP					AE	VAE
Nombre de facteurs	3	5	3	5	3	5	50	100	200		
Tout	24.09%	25.77%	42.79%	42.54%	48.65%	51.64%	67.23%	75.25%	86.02%	-73.00%	30.84%
Services de communication	13.84%	15.80%	33.64%	34.13%	33.62%	35.50%	66.05%	80.47%	90.02%	-89.39%	20.92%
Biens de consommation	19.22%	23.39%	32.75%	34.35%	38.50%	38.78%	65.33%	78.43%	92.39%	-53.52%	29.54%
Biens de base	7.25%	6.61%	21.30%	23.09%	31.19%	31.27%	42.28%	51.04%	65.83%	-91.92%	19.95%
Énergie	39.69%	39.48%	43.79%	44.85%	64.66%	68.51%	77.95%	82.98%	92.07%	-35.33%	24.15%
Financier	42.23%	44.24%	66.23%	66.25%	63.64%	68.59%	74.14%	78.20%	84.31%	-57.73%	41.00%
Santé	7.05%	7.95%	30.56%	26.33%	33.92%	39.99%	61.51%	72.86%	85.61%	-140.57%	23.97%
Industriel	27.79%	29.36%	45.58%	45.26%	47.90%	51.14%	65.68%	70.97%	82.58%	-85.42%	36.64%
Technologie	23.08%	24.03%	49.36%	48.45%	49.83%	53.46%	67.92%	77.91%	89.83%	-76.66%	30.92%
Matériaux	27.97%	28.74%	44.34%	41.94%	46.75%	47.99%	67.70%	75.54%	86.82%	-70.98%	32.09%
Immobilier	18.47%	20.97%	36.68%	36.85%	54.31%	56.29%	65.91%	69.84%	80.60%	-50.41%	31.92%
Services	10.91%	12.15%	30.81%	32.42%	60.31%	61.36%	70.49%	72.54%	75.65%	-59.06%	27.48%

Ce tableau illustre les valeurs des R^2_{OOS} sur la période de test s'échelonnant du 30 mai 2018 au 31 décembre 2021 pour les techniques analysées sur les 424 actifs du S&P500 disponible entre 2010 et 2021 par secteur d'activité. Ces titres figuraient dans la liste des titres formant l'indice S&P500 en date du 30 juin 2022. Aucun des actifs n'a de données manquantes. Les titres ayant été acquis ou ayant pris part à une fusion ont été exclus.

Dans le tableau 6, nous pouvons observer les valeurs des R^2 pour les 10 titres ayant la plus petite variance entre le 1er janvier 2010 et le 31 décembre 2021. Nous observons une surperformance du R^2 individuel des actifs en comparaison au R^2 global pour quatre des dix actifs pour les méthodes utilisant des facteurs explicites. En moyenne, nous observons que la faible variance des titres n'indique pas une meilleure performance de notre métrique pour ce type de méthodes. Les actifs où le R^2 sous-performe font partie de secteurs qui eux aussi sont, en moyenne, moins performant que l'ensemble du portfolio. Pour Proctor & Gamble (PG), PepsiCo Inc. (PEP), Coca-Cola (KO) et Kimberly Clark Corp (KMB), nous observons une déviation importante de la moyenne de ce secteur pour les méthodes de facteurs explicites. Trois de ces quatre compagnies figurent dans le top 25 des compagnies en termes de capitalisation du S&P500. Cela dit Johnson & Johnson (JNJ) figure également dans ce palmarès et les résultats ne sont clairement pas aussi concluants. Pour ce qui est des résultats pour les techniques utilisant des facteurs explicites sur des actifs à variance élevée, les résultats sont également mitigés. Pour seulement cinq des dix actifs, une surperformance du R^2 individuel des actifs en comparaison au R^2 global.

En revanche, pour la méthode ACP, les performances sont considérablement meilleures. Nous observons une surperformance du R^2 individuel des actifs en comparaison au R^2 global pour huit des dix actifs à faibles variances. La nature même des facteurs explicites est d'expliquer la plus grande portion de variance. Pour des actifs avec peu de variances, il suffit de moins de facteurs afin d'expliquer les déviations des rendements. Ceci mène donc à une meilleure performance de cette méthode. Le tableau 5 nous semble confirmer cette hypothèse. Ce tableau présente les résultats des techniques pour les dix actifs avec la plus grande variance sur la période 1er janvier 2010- 31 décembre 2021 et nous observons de moins bons résultats pour cette technique.

Pour la méthode de l'autoencodeur conditionnel, nous obtenons des résultats mitigés pour les actifs à faibles variances. En effet, comme avec les méthodes utilisant des facteurs implicites, les résultats pour trois des quatre actifs incluent dans la liste, nous obtenons des résultats supérieurs à la moyenne pour ce secteur. Cela dit seulement deux actifs obtiennent de meilleurs résultats que la moyenne globale. Ceci peut être dû à la variance des résultats pour les différents secteurs. Nous avons des R^2 par secteur variant de -140.57% à -35.33%, soit une différence de plus de 100%. Pour les actifs avec une grande variance, nous observons dans le tableau 5 des résultats bien plus favorables. Pour tous les actifs sauf Netflix (NFLX), les résultats des R^2 individuels des

actifs dépassent celui du R^2 global. Cette technique est donc peut-être plus efficace avec des rendements plus volatiles. Bien que meilleurs, ces résultats sont néanmoins loin de la moyenne des autres techniques.

Avec la méthode du VAE, nous obtenons de meilleurs résultats pour les actifs les moins volatiles que ceux à grandes volatilités. Cependant, la métrique performe moins bien que presque toutes les méthodes factorielles (implicite ou explicite). Ceci est en ligne avec les résultats obtenus dans le tableau 4.

Si nous évaluons les techniques entre elles, les résultats obtenus dans le tableau 6 et le tableau 5 sont similaires à ceux obtenus dans le tableau 4 en termes d'ordre de performance.

Tableau 5 Comparaison des R^2_{OOS} estimés 05/2018-12/2021

		Régression facteurs Bryan Kelly		Régression facteurs Fama-French		ACP					AE	VAE
Symbole	Secteur	3	5	3	5	3	5	50	100	200		
MU	Technologie	29.38%	28.02%	46.86%	43.85%	59.25%	64.49%	95.96%	98.62%	99.62%	-31.96%	25.97%
RCL	Biens de consommation	34.22%	34.11%	34.03%	31.24%	39.53%	43.88%	89.05%	93.43%	93.77%	-17.54%	31.93%
UAL	Industriel	22.14%	25.02%	33.48%	26.75%	22.44%	41.36%	91.82%	93.52%	98.96%	-11.11%	19.31%
MRO	Énergie	42.69%	41.76%	41.33%	42.76%	69.71%	77.35%	87.99%	88.48%	96.96%	-27.64%	19.55%
FCX	Matériaux	38.10%	33.27%	48.31%	26.84%	37.95%	44.32%	90.52%	97.34%	99.16%	-26.41%	34.82%
NFLX	Services de communication	0.84%	0.57%	35.00%	31.76%	33.00%	35.65%	97.79%	99.01%	99.62%	-216.91%	10.35%
AAL	Industriel	22.69%	25.23%	28.88%	24.94%	20.80%	39.55%	86.83%	94.56%	99.47%	-15.31%	20.92%
PENN	Biens de consommation	18.77%	22.13%	26.12%	26.29%	24.32%	23.97%	52.26%	96.21%	99.50%	-15.84%	14.93%
APA	Énergie	33.90%	32.90%	33.82%	35.32%	57.72%	61.76%	76.29%	79.58%	97.02%	-11.28%	12.91%
AMD	Technologie	21.67%	23.97%	33.11%	33.55%	33.80%	43.14%	98.83%	99.51%	99.82%	-23.69%	14.31%
Moyenne		26.44%	26.70%	36.09%	32.33%	39.85%	47.55%	86.73%	94.03%	98.39%	-39.77%	20.50%
R^2 des 10 titres		27.94%	28.50%	34.61%	31.89%	40.14%	47.37%	82.26%	92.40%	98.11%	-26.15%	19.88%

Ce tableau illustre les valeurs des R^2_{OOS} sur la période de test s'échelonnant du 30 mai 2018 au 31 décembre 2021 pour les techniques analysées des 10 actifs avec la plus grande variance sur la période du 1er janvier 2010 au 31 décembre 2021. Ces 10 titres font partie de la liste de 424 titres du S&P500 disponible entre 2010 et 2021. Ces titres figuraient dans la liste des titres formant l'indice S&P500 en date du 30 juin 2022. Aucun des actifs n'a de données manquantes. Les titres ayant été acquis ou ayant pris part à une fusion ont été exclus.

Tableau 6:Comparaison des R^2_{OOS} estimés 05/2018-12/2021 des 10 actifs avec la plus petite variance

		Régression facteurs Bryan Kelly		Régression facteurs Fama-French		ACP					AE	VAE
Symbole	Secteur	3	5	3	5	3	5	50	100	200		
JNJ	Santé	6.95%	12.86%	35.11%	37.56%	39.03%	45.06%	57.75%	57.29%	60.62%	-96.90%	30.29%
PG	Biens de base	39.76%	40.01%	62.01%	64.01%	63.91%	64.24%	76.04%	75.38%	78.69%	-66.69%	21.42%
PEP	Biens de base	54.96%	55.52%	74.65%	77.76%	79.60%	79.71%	85.02%	84.76%	85.54%	-52.14%	31.82%
VZ	Services de communication	6.13%	4.15%	23.97%	15.65%	31.69%	31.93%	36.37%	34.95%	32.18%	-115.34%	22.28%
KO	Biens de base	35.57%	36.51%	48.20%	56.10%	67.72%	67.53%	70.82%	71.80%	70.02%	-81.11%	39.33%
CL	Biens de base	5.86%	2.33%	38.17%	39.30%	49.96%	51.27%	63.28%	63.10%	66.84%	-105.29%	23.97%
KMB	Biens de base	25.57%	24.92%	39.43%	41.23%	50.42%	51.04%	66.10%	67.41%	70.84%	-74.12%	13.21%
DUK	Services	12.72%	15.13%	28.22%	34.16%	74.29%	73.97%	83.12%	82.98%	84.03%	-100.71%	35.03%
ED	Services	11.94%	11.86%	18.11%	21.36%	64.00%	63.48%	76.35%	75.43%	75.47%	-69.59%	22.48%
WM	Industriel	15.03%	18.72%	44.69%	46.24%	58.03%	57.64%	61.96%	59.18%	58.35%	-101.02%	36.28%
Moyenne		21.45%	22.20%	41.26%	43.34%	57.87%	58.59%	67.68%	67.23%	68.26%	-86.29%	27.61%
R^2 des 10 titres		26.02%	26.74%	45.18%	47.82%	61.43%	61.98%	71.07%	70.74%	71.92%	-84.87%	27.74%

Ce tableau illustre les valeurs des R^2_{OOS} sur la période de test s'échelonnant du 30 mai 2018 au 31 décembre 2021 pour les techniques analysées des 10 actifs avec la plus petite variance sur la période du 1er janvier 2010 au 31 décembre 2021. Ces 10 titres font partie de la liste de 424 titres du S&P500 disponible entre 2010 et 2021. Ces titres figuraient dans la liste des titres formant l'indice S&P500 en date du 30 juin 2022. Aucun des actifs n'a de données manquantes. Les titres ayant été acquis ou ayant pris part à une fusion ont été exclus.

4.3. Période de test en temps de crise

Dans cette section, nous présentons les résultats des différentes méthodes pour les prédictions lors d'une période de crise. L'objectif est de déterminer si les différentes méthodes performent bien lors de périodes où les rendements sont plutôt négatifs et où la variabilité des rendements est accrue. La période de crise est la crise COVID entre le 1er janvier 2020 et le 31 décembre 2021. Bien que cette crise ait commencé en mars 2020, nous avons inclus les premiers mois de l'année. Les répercussions des décisions économiques (ex. abaissement des taux directeurs à des taux historiquement bas, réduction des imports/exports au début de la crise pour réduire l'effet de contagion, hausse des taux directeurs afin de lutter contre l'inflation) prises pour lutter contre cette crise se font encore sentir dans les marchés en 2023 et une récession est encore probable. Nous nous concentrons donc sur les effets de cette crise lors des deux premières années.

Lors de la crise sanitaire, les secteurs les plus surs étaient bien entendu les secteurs de la santé et des biens de consommation. De nombreuses compagnies pharmaceutiques ont profité d'investissement de la part des gouvernements afin de développer un vaccin et il y a eu une panique de la part des consommateurs en ce qui concerne les biens de consommation et ce secteur a connu une forte demande tout au long de la pandémie. Cela dit, tel que nous l'observons dans le tableau 1 la moyenne et la variance des rendements pour ces secteurs sont assez similaires aux autres secteurs. Nous pouvons observer dans le tableau 7 que la moyenne et la variance des rendements sont similaires pour tous les secteurs. Le secteur de la technologie est le secteur ayant le mieux performer. Nous pouvons attribuer cette performance à forte demande vers un virage numérique des entreprises. Beaucoup d'entreprises ont dû adapter leur infrastructure numérique afin de soutenir leur besoin et de permettre une transition vers le télétravail. Les entreprises ont aussi amélioré leur réseau de distribution en offrant plus de commandes en ligne. Le secteur énergétique, notamment le domaine pétrolier, a varié plus que les autres secteurs. En effet, il a été grandement affecté par le ralentissement des imports/exports et la réduction considérable des besoins de transport a causé la chute importante du prix du pétrole. Nous avons ensuite observé une remontée du prix du baril. Ceci vient évidemment affecter le prix des actifs pour lesquels leurs principales activités sont liées au pétrole. Ceci est le cas pour la plupart des compagnies dans le secteur énergétique. La chute drastique des marchés

en février et mars 2020 a été suivie d'une forte croissance jusqu'en décembre 2021. Nous avons donc une courte période de rendements négatifs suivie d'une période de forte croissance. Durant cette période, il y avait aussi une volatilité accrue.

Tableau 7 Statistiques descriptives par secteurs des actifs par secteurs pour l'année 2020-2021

	Nombre d'actif	Moyenne (%)	Variance (%)	Minimum (%)	Maximum (%)
Services de communication	16	5.50	14.10	-19.78	15.30
Biens de consommation	50	5.68	14.96	-25.80	14.39
Biens de base	28	5.23	9.65	-10.30	13.75
Énergie	15	2.48	17.92	-33.60	12.84
Financier	60	5.12	13.10	-13.57	11.96
Santé	55	6.39	13.03	-17.17	20.93
Industriel	59	5.74	12.85	-15.67	14.95
Technologie	61	6.63	14.12	-17.46	18.27
Matériaux	22	4.67	13.76	-13.97	11.29
Immobilier	30	4.79	12.05	-17.44	13.03
Services	28	4.75	9.65	-10.19	11.19

Ce tableau présente les statistiques descriptives pour l'année 2020-2021 des 424 titres du S&P500 disponible entre 2010 et 2021 par secteur. Ces titres figuraient dans la liste des titres formant l'indice S&P500 en date du 30 juin 2022. Aucun des actifs n'a de données manquantes. Les titres ayant été acquis ou ayant pris part à une fusion ont été exclus.

Afin de comparer les résultats en temps de crise aux résultats sur la période de test complète, nous avons deux choses à prendre en considération. La première étant que la fenêtre de comparaison est plus petite. Nous avons donc moins de rendements à sommer. Le dénominateur et le numérateur devraient donc être tous les deux inférieurs à celui pour la période complète. La seconde est que la période de crise est également incluse dans les résultats de la mesure pour la période complète. Nous comparons donc la mesure dans un sous-ensemble des données. Les résultats de la métrique pour les deux périodes ne sont ainsi pas indépendants.

Si nous observons que la métrique performe aussi bien en temps de crise que pour la période complète, il est possible que la diminution du dénominateur soit proportionnelle à celle du numérateur. Il est possible que la différence absolue entre les valeurs observées et celles prédites

augmente. Cela est attendu puisque la fenêtre de données d'entrainements n'inclut pas de crise. Les modèles ne sont donc pas calibrés sur des données de crise. Les rendements moyens observés dans le tableau 7 nous démontrent que les rendements sont également, en moyennes, plus grands.

Lorsque les rendements sont près de zéro, le carré des rendements inclus dans le dénominateur devient très petit. Même avec une faible différence entre les rendements prédits et les rendements observés, le ratio peut être élevé, car les valeurs sommées au dénominateur sont également petites faisant ainsi augmenter le ratio. Avec un ratio plus élevé, notre métrique performe moins bien. En période de crise, les rendements sont, en moyenne, plus élevés. Ce problème est donc évité et la performance semble donc meilleure. Notre évaluation de la performance des différents modèles est donc biaisée. En réduisant la fréquence de nos rendements, ce biais serait moindre puisque les rendements seraient moins près de zéro.

Tableau 8 Comparaison des R^2_{OOS} estimés 2020-2021

	Régression facteur Bryan Kelly		Régression facteur Fama-French		ACP					AE	VAE
	3	5	3	5	3	5	50	100	200		
Tout	25.99%	27.60%	46.34%	45.87%	52.29%	55.15%	69.63%	76.90%	86.81%	-51.44%	59.00%
Services de communication	14.55%	16.30%	38.50%	39.02%	38.84%	40.99%	67.60%	81.08%	89.87%	-67.35%	45.14%
Biens de consommation	20.60%	25.21%	35.44%	37.10%	42.20%	42.43%	68.34%	80.25%	93.09%	-36.42%	59.61%
Biens de base	10.74%	9.00%	26.82%	29.36%	35.84%	35.96%	46.91%	55.45%	68.19%	-74.71%	55.65%
Énergie	40.66%	40.44%	45.29%	45.81%	66.99%	69.89%	79.20%	83.93%	92.46%	-24.62%	61.34%
Financier	45.58%	47.56%	69.36%	69.27%	67.09%	71.17%	76.36%	80.11%	85.82%	-35.49%	65.68%
Santé	4.35%	4.72%	32.16%	25.99%	36.96%	43.11%	61.47%	72.56%	84.96%	-116.01%	47.73%
Industriel	28.93%	30.96%	47.48%	46.67%	49.48%	53.36%	69.05%	74.02%	84.19%	-58.48%	62.05%
Technologie	21.06%	21.66%	52.84%	51.64%	52.22%	55.97%	68.85%	78.08%	89.66%	-58.71%	54.17%
Matériaux	30.59%	31.53%	48.86%	45.73%	50.52%	52.02%	70.48%	77.60%	88.09%	-50.89%	59.46%
Immobilier	23.35%	24.98%	41.89%	41.86%	55.67%	57.77%	67.04%	70.75%	81.25%	-34.00%	61.29%
Services	14.93%	15.43%	37.86%	40.54%	63.85%	64.81%	73.03%	74.95%	78.12%	-43.63%	64.58%

Ce tableau illustre les valeurs des R^2_{OOS} sur la période de test s'échelonnant du 1er janvier 2020 au 31 décembre 2021 pour les techniques analysées sur les 424 actifs du S&P500 disponible entre 2010 et 2021 par secteur d'activité. Ces titres figuraient dans la liste des titres formant l'indice S&P500 en date du 30 juin 2022. Aucun des actifs n'a de données manquantes. Les titres ayant été acquis ou ayant pris part à une fusion ont été exclus.

En général, nous observons une meilleure performance de notre métrique pour toutes les techniques en comparaison à la période de test de mai 2018 au 31 décembre 2021. Nous pouvons supposer que ce problème est dû au biais que présente notre métrique lorsque la valeur des rendements est près de zéro. En effet, lorsque nous comparons la valeur moyenne de la différence entre les valeurs prédites et les valeurs observées, les résultats pour la période de crise sont généralement moins bons que pour la période complète.

Lorsque nous comparons les techniques durant la période de crise, nous observons des résultats semblables à ceux couvrant la période complète. Toutefois, le VAE semble mieux performer en temps de crise que la méthode de régression utilisant les facteurs de Fama French et que l'ACP trois et cinq facteurs. Le VAE est un modèle dynamique et les modèles factoriels sont statiques, il est possible que la performance soit meilleure en temps de crise.

Le tableau comparatif des résultats pour les différentes méthodes en temps de crise pour les 10 actifs avec la plus grande variance et les 10 actifs avec la plus petite variance peuvent être observés dans l'Annexe A et dans l'Annexe B

5. Conclusion

Pour conclure, nous allons détailler les limitations des différentes méthodes utilisées ainsi que proposer des pistes de recherches qui nous permettraient de contourner ces limitations et possiblement d'améliorer la capacité prédictive des rendements.

Le modèle factoriel de Fama French (2015), bien qu'il soit présenté comme une amélioration du modèle trifacteur n'explique quand même pas certaines anomalies : momentum et provision (*accrual*). Les anomalies de volatilité et de l'émission nette d'actions qui ne sont pas expliquées par le modèle trifacteur sont expliquées par une exposition au facteur d'investissement ou de profitabilité. Dans leur article « *Dissecting Anomalies with a Five-Factor Model* », Fama et French (2016) expliquent que bien que le facteur de momentum soit important, il est primordial que le portefeuille d'actif soit bâti sur une stratégie de momentum. Dans notre cas, cette anomalie ne nous affecte pas puisque notre univers d'actif n'est pas conçu dans l'optique d'une stratégie de momentum. En effet, avec une stratégie de momentum nous devrions vendre les

actifs perdants et acheter les actifs avec des rendements positifs. Or, notre univers d'actifs reste le même, peu importe le rendement des actifs. L'objectif d'une stratégie de momentum est d'utiliser les tendances des rendements passés afin de prédire le futur. Certains auraient une tendance de rendements positifs alors que d'autres non allant ainsi à l'encontre de la stratégie. Pour le facteur de provision, l'anomalie est présente lorsque le portefeuille est composé d'actifs à petite capitalisation (microcaps). Ce n'est pas le cas pour notre univers d'actifs, ce n'est donc pas un problème.

Il est également important de noter que notre univers d'actif est restreint aux actifs composant le S&P500. Ils sont donc tous des actifs à grande capitalisation. Le facteur de taille est donc moins intuitif. Notre méthode avec les facteurs de JKP essaie de trouver une meilleure combinaison de facteurs plus propre à notre univers d'actif. Cependant les performances sont plus décevantes.

Le plus grand désavantage de la régression linéaire est qu'elle utilise des facteurs de régression statique. C'est-à-dire que les coefficients de régression sont calculés en utilisant l'information jusqu'au temps T' (où T' est la période couverte par l'échantillon d'entrainement). Lorsque nous observons un déplacement covarié (*covariate shift*) les valeurs prédites par le modèle avec des facteurs statiques vont définitivement moins bien performer, car le modèle n'a pas été pas entrainé sur ces données. Une solution serait de tester l'utilisation de facteurs dynamiques. Nous pourrions recalculer les coefficients de régression utilisant l'information disponible jusqu'à t-1 afin de prédire les rendements des différents actifs au temps t.

Pour la méthode de facteurs implicites, Shukla et Trzcinka (1990) ont démontré que cinq facteurs étaient suffisants afin d'expliquer les rendements hebdomadaires. Nos résultats sont assez similaires aux leurs. Cinq facteurs sont suffisants afin d'obtenir de meilleurs résultats que toutes les autres techniques. Nous obtenions de meilleurs résultats en ajoutant des facteurs, car notre métrique ne prenait pas en compte le nombre total de facteurs utilisés. Une métrique pénalisant le nombre de facteurs nous aurait permis de capturer la valeur ajoutée de ces facteurs supplémentaires.

Tel que Heaton et al. (2017) le notent dans leur article, le problème fondamental avec les autoencodeurs est qu'ils sont conçus pour répliquer une tâche. En revanche, lorsque nous tentons de déterminer la performance d'un titre dans le futur, il ne s'agit pas de répliquer le passé. Nous devons prédire les variations possibles d'un actif. Un titre qui performe bien sur plusieurs mois

ne performera pas nécessairement bien ou aussi bien dans le futur. Pour cette raison, le modèle performe moins bien. De plus, la projection des données entrantes dans l'espace latent est un vecteur. L'espace latent peut ne pas être continu ou permettre une interpolation facile. Dans notre cas, il est désirable d'avoir un espace latent continu et qui permet de générer de nouvelles données. Puisque l'espace latent est de dimension plus petite, la génération de nouvelles données serait plus efficace, en théorie, que d'avoir un modèle pour toutes les variables. Gu et al. (2021) obtiennent de bons résultats, car l'information conditionnelle de l'autoencodeur est une série de facteurs macroéconomiques. Dans notre cas, nous utilisons une application plus naïve en utilisant les rendements d'actifs pour déterminer le rendement d'un actif. L'utilisation d'un autoencodeur n'est toutefois pas exclue. En effet, ce genre de réseau peut être utilisé afin de sélectionner les composantes qui seront modélisées à des fins de prédictions.

Le VAE est un modèle génératif. Il peut donc générer de nouvelles données à partir de la distribution calibrée. Les données générées peuvent être similaires aux données initiales ou avec des variations. Ces variations produisent donc des données qui n'ont jamais existé, mais qui sont plausibles. Le VAE n'est pas parfait. En effet, nous supposons que la distribution a posteriori $q(z_i|R_i;\Psi)$ est une distribution gaussienne avec une matrice de covariance diagonale. Cette hypothèse est une limitation en soi. En effet, la distribution posteriori peut être covariée. Il est également possible d'utiliser d'autres distributions. Cela dit, c'est un compromis entre la restriction imposée par ce choix et notre capacité de générer une simulation de la distribution de façon efficace ou à différencier les paramètres de la distribution pour des fins d'optimisation. Une autre possibilité serait d'ajouter un ensemble de variables latentes auxiliaires a afin d'augmenter l'inférence et de permettre à la distribution latente d'être plus expressive. C'est-à-dire, $q(z|R) = \int \mathrm{q}(a,z|R)da$. Un autre problème du VAE est qu'il peut arriver d'atteindre un optimum local pour lequel $q(z|R;\Psi) = p(z|R;\Theta) = p(z)$. Dans ce cas, aucune information de l'espace latent n'est traitée par le décodeur. L'idée derrière le modèle VAE est d'avoir une variable latente insoluble nous permettant de tirer de l'information et d'améliorer la capacité à reconstruire les données et générer de nouvelles données. Il est donc primordial que l'espace latent puisse traiter l'information. Selon le livre de Pinheiro Cinelli et al. (2021), le β-VAE nous permettrait d'éviter ce genre d'optimum local. Le facteur β multiplie la divergence KL dans l'équation (*25*). Un β <1 nous permettrait de résoudre ce problème.

Il est important de noter que toutes nos techniques ne sont pas immunes au décalage de la probabilité antérieur. Un modèle à changement de régime nous permettrait peut-être de contourner ce problème. Cela dit, un tel modèle requiert suffisamment de données dans les différents régimes pour calibrer le modèle, ce qui peut représenter un obstacle. De plus, l'utilisation de rendements journaliers apporte quelques complications. Roll et Ross (1980) ont obtenu de meilleurs résultats lorsqu'ils utilisaient un rendement sur deux. Ils ont attribué la différence dans leurs résultats à l'asymétrie des rendements quotidiens.

Annexe A Tableau de Comparaison des R^2_{OOS} estimés 2020-2021 pour les 10 actifs avec la plus grande variance

		Regression facteurs Bryan Kelly		Regression facteurs Fama-French		ACP					AE	VAE
Symbole	Secteur	3	5	3	5	3	5	50	100	200		
JNJ	Santé	11.01%	19.62%	43.56%	46.98%	47.49%	55.21%	70.43%	68.71%	72.03%	-86.93%	64.20%
	Biens de base	37.99%	37.84%	66.22%	69.31%	65.16%	65.75%	79.52%	78.17%	81.97%	-47.95%	67.80%
PEP	Biens de base	52.30%	52.66%	76.21%	80.14%	80.35%	80.51%	86.40%	86.07%	86.98%	-40.91%	75.50%
VZ	Services de communication	12.17%	6.98%	36.71%	26.30%	36.57%	36.60%	39.62%	34.67%	29.45%	-103.03%	61.45%
KO	Biens de base	34.39%	34.23%	48.04%	59.47%	70.57%	70.44%	72.27%	73.65%	70.55%	-54.75%	70.42%
CL	Biens de base	17.21%	12.33%	50.13%	52.31%	59.06%	60.81%	71.18%	70.25%	73.16%	-83.83%	67.10%
KMB	Biens de base	30.74%	28.63%	46.35%	49.34%	54.07%	55.30%	72.11%	72.61%	74.43%	-72.36%	52.09%
DUK	Services	15.38%	17.66%	33.88%	41.34%	77.72%	77.06%	84.97%	84.70%	85.92%	-72.87%	69.08%
ED	Services	14.38%	13.43%	22.00%	26.35%	65.12%	64.08%	76.41%	74.89%	75.03%	-50.88%	66.62%
WM	Industriel	23.51%	24.70%	52.36%	54.11%	63.70%	63.30%	67.81%	64.04%	61.64%	-59.86%	70.54%
Moyenne		24.91%	24.81%	47.55%	50.57%	61.98%	62.91%	72.07%	70.77%	71.12%	-67.34%	66.48%
R^2 des 10 titres		28.11%	28.09%	49.96%	53.82%	65.55%	66.19%	75.38%	74.40%	75.04%	-64.52%	67.11%

Ce tableau illustre les valeurs des R^2_{OOS} sur la période de test s'échelonnant du 1er janvier 2020 au 31 décembre 2021 pour les techniques analysées des 10 actifs avec la plus petite variance sur la période du 1er janvier 2010 au 31 décembre 2021. Ces 10 titres font partie de la liste de 424 titres du S&P500 disponible entre 2010 et 2021. Ces titres figuraient dans la liste des titres formant l'indice S&P500 en date du 30 juin 2022.Aucun des actifs n'a de données manquantes. Les titres ayant été acquis ou ayant pris part à une fusion ont été exclus.

Annexe B Tableau de Comparaison des R^2_{OOS} estimés 2020-2021 pour les 10 actifs avec la plus petite variance

		Regression facteurs Bryan Kelly		Regression facteurs Fama-French		ACP					AE	VAE
Symbole	Secteur	3	5	3	5	3	5	50	100	200		
MU	Technologie	21.21%	18.98%	49.68%	45.37%	61.63%	64.79%	95.82%	98.45%	99.60%	-22.13%	51.84%
RCL	Biens de consommation	35.15%	34.87%	33.43%	30.47%	39.21%	44.03%	90.32%	94.03%	94.13%	-10.48%	61.71%
UAL	Industriel	22.84%	25.88%	33.46%	26.26%	21.95%	41.23%	93.56%	94.68%	99.21%	-6.85%	51.09%
MRO	Énergie	43.51%	42.39%	42.09%	42.86%	72.31%	78.31%	89.00%	88.77%	97.06%	-18.44%	56.04%
FCX	Matériaux	38.66%	32.28%	52.63%	23.70%	35.78%	44.47%	91.13%	97.22%	99.17%	-15.64%	57.28%
NFLX	Services de communication	-20.36%	-22.53%	29.99%	24.67%	26.92%	29.68%	97.23%	98.82%	99.53%	-168.69%	20.57%
AAL	Industriel	22.35%	26.02%	27.57%	23.32%	16.86%	37.57%	87.10%	95.39%	99.59%	-11.00%	45.33%
PENN	Biens de consommation	18.13%	22.13%	25.71%	26.07%	24.23%	23.77%	53.18%	96.34%	99.54%	-10.25%	61.93%
APA	Énergie	34.69%	33.43%	34.77%	35.46%	59.32%	62.49%	77.57%	80.76%	97.37%	-8.12%	63.51%
AMD	Technologie	13.72%	18.36%	36.56%	37.07%	32.01%	41.48%	98.29%	99.27%	99.75%	-25.30%	31.96%
Moyenne		22.99%	23.18%	36.59%	31.53%	39.02%	46.78%	87.32%	94.37%	98.50%	-29.69%	50.13%
R^2 des 10 titres		27.36%	28.08%	34.57%	31.07%	39.59%	46.84%	81.81%	92.47%	98.13%	-16.24%	55.41%

Ce tableau illustre les valeurs des R^2_{OOS} sur la période de test s'échelonnant du 1er janvier 2020 au 31 décembre 2021 pour les techniques analysées des 10 actifs avec la plus grande variance sur la période du 1er janvier 2010 au 31 décembre 2021. Ces 10 titres font partie de la liste de 424 titres du S&P500 disponible entre 2010 et 2021. Ces titres figuraient dans la liste des titres formant l'indice S&P500 en date du 30 juin 2022. Aucun des actifs n'a de données manquantes. Les titres ayant été acquis ou ayant pris part à une fusion ont été exclus.

Bibliographie

[1] Black, F., Jensen, M., & Scholes, M. (1972). The Capital Asset Pricing Model: Some Empirical Tests. In M. C. Jensen, *Studies in the Theory of Capital Markets* (pp. 79-121). New York: Preager.

[2] Cochrane, J. (2011). Presidential address: Discount rates. *The Journal of Finance, 66*(4), 1048-1108.

[3] Fama, E., & French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics, 33*(1), 3-56.

[4] Fama, E., & French, K. (1996). Multifactor Explanations of Asset Pricing Anomalies. *The Journal of Finance, 51*(1), 55-84.

[5] Fama, E., & French, K. (2015). A five-factor asset pricing mode. *Journal of Financial Economics, 116*(1), 1-22.

[6] Fama, E., & French, K. (2016). Dissecting Anomalies with a Five-Factor Model. *The Review of Financial Studies, 29*(1), 69-103.

[7] Fama, E., & MacBeth, J. (1973). Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy, 81*(3), 753-755.

[8] French, K. (2022, 06 30). Retrieved from https ://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data library.html

[9] Goodfellow, I. J., Bengio, Y., & Courville, A. (2016). *Deep Learning.* Cambridge, MA, USA: MIT Press.

[10]Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoders Asset Pricing Models. *Journal of Econometrics, 222*(1), 429-450.

[11]Harvey, C., Liu, Y., & Zhu, H. (2016). … and the Cross-Section of Expected Returns. *The Review of Financial Studies, 29*(1), 5-68.

[12]Heaton, J., Polson, N., & Witte, J. (2017). Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry, 33*, 3-12.

[13]James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning.* Springer New York, NY.

[14]Jensen, T., Kelly, B., & Pedersen, L. (2023). Is There a Replication Crisis in Finance? *The Journal of Finance, 78*(5), 2465-2518.

[15]Kelly, B. (2022, 06 30). *Bryan Kelly Academic*. Retrieved from https ://www.bryankellyacademic.org

[16]Kelly, B., Pruitt, S., & Su, Y. (2019). Characteristics are covariances : a unified model of risk and return. *Journal of Financial Economics, 134*(3), 510-524.

[17]Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear* (5 ed.). New York: McGraw-Hill Irving.

[18]Laloux, L., Cizeau, P., Bouchaud, J.-P., & Potters, M. (1999). Noise Dressing of Financial Correlation Matrices. *PHYSICAL REVIEW LETTERS, 83*(7), 1467.

[19]Lintner, J. (1965a). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *Review of Economics and Statistics*, 13-37.

[20]Lintner, J. (1965b). Security Prices, Risk and Maximal Gains from Diversification. *Journal of Finance*, 587-615.

[21]*List of S&P 500 companies*. (2022, 6 30). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

[22]Long, L., & Zeng, X. (2022). *Beginning Deep Learning with TensorFlow: Work with Keras, MNIST Data Sets, and Advanced Neural Networks.* Springer.

[23]Mossin, J. (1966). Equilibrium in a Capital Asset Market. *Econometrica, 34*(4), 768-783.

[24]Ouyang, H., Zhang, X., & Yan, H. (2019). Index tracking based on deep neural network. *Cognitive Systems Research, 57*, 107-114.

[25]Pinheiro Cinelli, L., Araújo Marins, M., Barros da Silva, E., & Lima Netto, S. (2021). *Variational Methods for Machine Learning with Applications to Deep Networks.* Springer.

[26]Plerou, V., Gopikrishnan, P., Rosenow, B., Nunes Amaral, L., & Stanley, E. (1999). Universal and Nonuniversal Properties of Cross Correlations in Financial Time Series. *PHYSICAL REVIEW LETTERS, 83*(7), 1471.

[27]Plerou, V., Gopikrishnan, P., Rosenow, B., Nunes Amaral, L., Guhr, T., & Stanley, E. (2002). Random matrix approach to cross correlations in financial data. *PHYSICAL REVIEW E, 65*(6), 1-18.

[28]Roll, R., & Ross, S. (1980). An Empirical Investigation of the Arbitrage Pricing Theory. *The Journal of Finance, 35*(5), 1073-1103.

[29]*S&P 500 Companies by Weight*. (n.d.). Retrieved from https://www.slickcharts.com/sp500.

[30]Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance, 19*(3), 425-442.

[31]Shukla, R., & Trzcinka, C. (1990). Sequential Tests of the Arbitrage Pricing Theory: A Comparison of Principal Components and Maximum Likelihood Factors. *The Journal of Finance, 45*(5), 1541-1564.

[32]Treynor, J. (1999). Toward Theory of Market Value of Risky Assets, 1962. In R. Korajczyk, *Asset Pricing and Portfolio Performance* (pp. 15-22). London: Risk Books.

[33]Zhang, C., Liang, S., Lyu, F., & Fang, L. (2020). Stock-Index Tracking Optimization Using Auto-Encoders. *Frontier in Physincs, 8*, 1-15.