Université de Montréal

Critères d'optimisation d'algorithmes d'apprentissage en gestion de portefeuille

par
Nicolas Chapados
Département d'informatique et de recherche opérationnelle
Faculté des arts et sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maîtrise
Informatique

31 mars 2000

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé:

Critères d'optimisation d'algorithmes d'apprentissage en gestion de portefeuille

présenté par:

Nicolas Chapados

a été évalué par un jury composé des personnes suivantes:

Pierre L'Écuyer
(président-rapporteur)

Yoshua Bengio
(directeur de recherche)

Felisa Vázquez–Abad
(membre du jury)

Mémoire accepté le:

Sommaire

Les systèmes adaptatifs, tels que les réseaux de neurones, jouent un rôle croissant en gestion de portefeuille. Nous considérons leur utilisation dans le problème de l'allocation d'actifs.

Pour ce problème, les systèmes adaptatifs sont traditionnellement utilisés pour modéliser les quelques premiers moments (la moyenne et la variance, par exemple) de la distribution jointe future du rendements des actifs. La prévision effectuée sur le comportement des actifs est ensuite fournie à un système de décision classique, comme l'allocation moyenne–variance, qui, sous certaines conditions, rend une décision d'allocation optimale qui respecte les contraintes d'aversion au risque de l'investisseur.

Une nouvelle utilisation des systèmes adaptatifs dans le problème de l'allocation d'actifs consiste à les employer à *rendre directement une décision*, sans passer par l'étape intermédiaire de la prévision.

Un premier objectif de ce mémoire est de comparer ces deux paradigmes d'utilisation des systèmes adaptatifs. Nous pouvons considérer que chacun optimise un critère (fonction de coût) différent : un critère d'erreur quadratique pour le modèle de prévision, et un critère de rendement financier pour le modèle de décision.

Un second objectif de ce mémoire est d'explorer l'utilisation de méthodes récentes de *combinaison de modèles*, qui construisent un « comité » à partir d'un certain nombre de modèles sous-jacents. Plusieurs résultats théoriques et expérimentaux indiquent que la performance du comité est généralement

supérieure à celle d'un des modèles sous-jacents choisi au hasard. Nous utilisons les comités pour systématiser le choix souvent problématique des hyperparamètres dans l'entraînement de réseaux de neurones.

Nos expériences sont effectuées dans le cadre de l'allocation par *contrôle de la valeur à risque*. Cette méthodologie, dont la popularité est croissante en gestion de trésorerie, contrôle le risque subi par portefeuille en limitant la perte maximale possible (avec une certaine probabilité) à une valeur fixée par le gestionnaire.

Les résultats expérimentaux sont les suivants :

1. Comparant la performance des modèles sous-jacents, le modèle de prévision produit des résultats statistiquement significativement supérieurs à ceux du modèle de décision. Cependant, pour certaines configurations des hyperparamètres, le modèle de décision fournit la même performance que le modèle de prévision.
2. Les méthodes de combinaison de modèles éliminent ces différences brutes entre les sous-jacents : nous ne trouvons aucune différence significative entre les comités formés par les modèles de prévision ou de décision. De plus, la performance des comités n'est jamais significativement pire que celle *du meilleur* de leurs modèles sous-jacents. Finalement, nous observons fréquemment que la performance des comités est significativement supérieure à celle d'un de leurs modèles sous-jacents tiré au hasard.

Summary

Adaptive systems, such as neural networks, play an increasingly important role in portfolio management. In this thesis, we apply them to the problem of asset allocation.

Traditionally, in solving this problem, adaptive systems are used to model the first few lower-order moments (for instance, the mean and the variance) of the joint distribution of asset returns. The forecast made by the adaptive system about the asset behaviour is then passed to a classical decision module, such as a mean–variance allocator, which, under certain assumptions, produces an optimal allocation decision while remaining consistent with the investor's degree of risk aversion.

A new use of adaptive systems in the problem of asset allocation is to have them *make direct decisions*, without the intervening step of forecasting the asset returns.

A first goal of this thesis is to compare those two paradigms for using adaptive systems. One way to view this problem is to consider that each optimizes a different criterion (cost function) : a mean-squared error criterion for the forecasting model, and a financial return criterion for the decision model.

A second goal of this thesis is to explore recent methods of *model combination*, which construct a « committee » from a number of underlying models. Several theoretical and experimental results indicate that the performance of the committee should generally be superior to that of one of the underlying

models picked randomly. In this thesis, we use committees to systematize the problem of choosing hyperparameters for training neural networks.

Our experiments are performed within the *value-at-risk control* framework, which is increasingly popular in short-term asset management. This methodology controls the risk to which a portfolio is exposed by setting the maximal loss (within a certain probability) that can be incurred by the portfolio.

Our experimental results are as follows :

1. Strictly comparing the underlying models, the forecasting model provides statistically significantly better performance than the decision model. However, for certain specific settings of the hyperparameters, the decision model exhibits the same performance as the forecasting model.
2. The model combination methods completely obliterate these differences between the underlying models : we find no significant difference between the committees formed by the forecasting or the decision model. Furthermore, the performance of a committee is never significantly worse than that of the *best* of its underlying models. Finally, we frequently observe that the performance of a committee is significantly better than that of one of its underlying models picked randomly.

À mes parents
et à mes petites soeurs

Préface

Par un beau vendredi d'octobre 1987, vous passez un coup de fil à votre conseiller financier pour lui demander avis sur « l'évolution des marchés » dans les prochains mois. Après une brève analyse, celui-ci vous déclare péremptoirement :

> *Notre bureau d'études stratégiques conclut à une hausse vraisemblable des taux d'intérets sous peu, hausse qui devrait conduire à une fourchette réduite d'occasions d'investir sur le marché boursier. De plus, nos analystes techniques perçoivent des pressions liquidatives imminentes qui devraient mener à une légère correction du prix des actions. Mon conseil : vendez tout.*

Non content de ces semonces paternalistes, vous arrangez un rendez-vous pour le jour-même avec votre voyante préférée, la mystérieuse Irma. Dans son obscur réduit, où ne filtre aucune lumière du jour, Irma jette un coup d'oeil dans sa boule de cristal, et à la lumière frémissante d'une bougie, annonce : « Je vois la sécheresse, la faim, la misère... Je vois le désarroi, les ambitions brisées, les rêves déchirés. Je vois... » Sous le coup d'une intense émotion, Irma doit s'interrompre un instant, avant d'ajouter : « Achetez de l'or, beaucoup d'or. »

Ambivalent de nature, ne sachant trop en qui, de votre conseiller ou de votre voyante, avoir confiance, vous prenez une décision fidèle à vous-même. Vous téléphonez à votre courtier et lui ordonnez : « Vendez toutes mes actions, *et* achetez de l'or ».

Le lundi suivant, 19 octobre 1987, coup de théâtre ! L'indice Dow Jones perd pour cette seule journée plus de 500 points. Vous vous en tirez cependant avec brio, votre décision de vendredi vous protégeant entièrement contre ce hoquet inattendu. Votre entourage vous prend soudainement pour un génie...

Prévoir ou décider ?

Ce bref récit illustre les deux grands thèmes de ce mémoire. Le premier pose la question : « Pour pouvoir prendre de bonnes décisions, est-il nécessaire de faire d'abord de bonnes prévisions ? » Dans les applications financières, une *prévision* est un pronostic effectué sur le comportement futur de variables d'intérêt, par exemple, les rendements du marché boursier. La *décision* est l'action prise sur les marchés, comme l'action d'investir dans certains titres plutôt que d'autres.

La décision peut faire suite à une prévision explicite, mais elle peut aussi être le résultat de processus complexes qui ne sont pas explicités, comme le résultat du traitement effectué par un réseau de neurones. Dans la petite histoire précédente, le conseiller financier incarne l'exemple de la prévision, qu'il peut vous expliquer, avant de vous suggérer une action. La voyante, quant à elle, ne vous donne aucune explication (que vous pouvez comprendre rationnellement) avant de vous indiquer dans quel chemin vous orienter—la décision sans prévision.

Combiner, combiner...

Manifestement, si vous aviez eu à *choisir* entre la recommandation de votre conseiller et celle de votre voyante, vous auriez obtenu un moins bon résultat qu'en *faisant un peu des deux*. Cette observation se confirme pour un grand nombre de situations pratiques : il est généralement avantageux de *combiner* plusieurs décisions alternatives que d'en choisir une à l'exclusion des autres.

Le deuxième thème de ce mémoire porte sur la combinaison de décisions : nous explorons plusieurs méthodes de combinaison de modèles et analysons leur performance relatives.

Remerciements

Avant tout, je désire remercier mon directeur de recherches, Professeur Yoshua Bengio, pour ses conseils judicieux et l'orientation qui m'a permis de m'aventurer sur des voies profitables. Merci aussi pour son dévouement, sa patience et son support, qui rendent exceptionnel l'accomplissement d'études supérieures sous sa direction.

J'aimerais souligner l'apport du CIRANO pour le support financier qui m'a grandement aidé dans la poursuite de mes études.

Merci à tous les membres du laboratoire d'informatique des systèmes adaptatifs (LISA) pour la patience de supporter un confrère aux prises avec des séances de déverminage souvent chargées émotionnellement ; particulièrement à Samy Bengio, Réjean Ducharme et Joumana Ghosn pour leur soutien lors mon expédition à travers des montagnes de code parfois inhospitalières, ainsi qu'à Charles Dugas pour son regard pénétrant sur la structure « des marchés ».

Merci à mes bons copains Louis-Martin Rousseau, Éric Méthot, et Alexandre Le Bouthilier pour des discussions toujours animées sur les éternels dérèglements boursiers, et les dernières aubaines à ne pas laisser passer.

Une pensée toute spéciale va à ma tendre amie Yasmina Chaibi, dont les visites inattendues rendirent chaque fois l'accomplissement de ce travail un peu moins insupportable.

Merci à mes parents, à mon père pour ses conseils et son expérience du milieu académique, et à ma mère pour n'avoir jamais douté de mes capacités en entreprenant un retour aux études.

Finalement, merci à Jean-François Blanchette, sans l'encouragement indéfectible de qui je serais encore programmeur pour une grosse compagnie, sis dans un petit cubicule gris...

Table des matières

Sommaire **iii**

Summary **v**

Préface **viii**

Table des matières **xi**

Liste des figures **xv**

Liste des tableaux **xvi**

1 Introduction **1**
1.1 Définitions et notation 2
1.1.1 Rendements simples 2
1.1.2 Rendements d'un portefeuille 3
1.1.3 Actifs spéciaux 4
1.1.4 Autres notations 4
1.2 Gestion moderne de portefeuille 5
1.2.1 Choix de portefeuille 7
1.2.2 Systèmes adaptatifs et allocation quadratique 7
1.3 Systèmes adaptatifs pour la décision 8
1.4 Combinaison de modèles 8
1.5 Aperçu du mémoire 9

2 Valeur à risque **11**
2.1 Principes de valeur à risque 11
2.1.1 Utilisations de la VàR 12
2.2 Estimation de la VàR 12

2.2.1 Distribution empirique 13
2.2.2 Approximation normale 14
2.3 La VàR comme modèle de placement 18
2.4 Équations de rééchelonnement 20
2.4.1 Estimateur de β_t 21
2.4.2 Distribution échantillonale de $\hat{\beta}_t$ 21
2.5 Évaluer la performance selon la VàR 26
2.5.1 Mesures classiques de performance 27
2.5.2 Une mesure propre à l'allocation selon la VàR 28
2.5.3 Frais de transaction 29
2.6 Calcul de la volatilité entre les actifs 30
2.6.1 Définition 30
2.6.2 Modèle localement constant 30
2.6.3 Variance historique simple 31
2.6.4 Variance historique pondérée exponentiellement 32
2.6.5 Le cas pour plusieurs actifs 34
2.6.6 Comment choisir le facteur d'oubli ? 34
2.6.7 Autres modèles de volatilité 37
2.6.8 À propos de la volatilité implicite 38

3 Systèmes adaptatifs pour la gestion de portefeuille 40
3.1 Un bref survol des réseaux de neurones 40
3.1.1 Pourquoi ? 41
3.1.2 Topologie d'un perceptron multi-couches 42
3.1.3 Entraînement d'un MLP 45
3.1.4 Réseaux récurrents 54
3.1.5 Les réseaux de neurones comme sous-systèmes 55
3.2 Deux paradigmes 56
3.2.1 Intérêt pratique de ces paradigmes 57
3.3 Modèle de prévision 58
3.3.1 Schéma général 58
3.3.2 Maximisation de l'utilité 59
3.3.3 Utilité quadratique 60
3.3.4 Équations d'allocation 62
3.3.5 Au-delà de l'allocation moyenne–variance 65
3.3.6 Stratégies d'entraînement d'un bon prédicteur 66
3.4 Modèle de décision 69
3.4.1 Schéma général 69
3.4.2 Équations de rétropropagation 72
3.4.3 Régularisation de la fonction de coût 76

4 Cadre expérimental **81**
4.1 Sommaire des deux paradigmes 81
4.1.1 Topologie 82
4.1.2 Prise de décisions 82
4.1.3 Entraînement 83
4.2 Estimation de la performance 86
4.2.1 Validation séquentielle 87
4.2.2 Pourquoi une validation *séquentielle* ? 89
4.2.3 Taille minimale d'entraînement 89
4.2.4 Ensemble de validation 91
4.3 Contrôle de la capacité 93
4.3.1 Pénalisation sur la norme des poids 93
4.3.2 Pénalisation sur la norme des entrées 94
4.4 Combinaisons de modèles 96
4.4.1 Comité : définition 97
4.4.2 Performance de généralisation d'un comité 98
4.4.3 Application au contexte de gestion de portefeuille . . . 101
4.4.4 Sélection des poids 102
4.5 Schéma des expériences 105
4.5.1 Comparaison entre les types de modèles 106
4.5.2 Comparaison entre les méthodes de combinaison 106
4.6 Ensembles de données 106
4.6.1 Description générale 106
4.6.2 Variables explicatives et prétraitements 107

5 Résultats et analyse **109**
5.1 Brèves remarques statistiques 109
5.1.1 Comparaison de séries de rendements 109
5.1.2 Analyse de la variance 110
5.1.3 Généralisation à plusieurs facteurs 112
5.1.4 Vérification des hypothèses 113
5.2 Comparaison entre les types de modèles 113
5.2.1 Résultats bruts 113
5.2.2 Analyse 121
5.3 Combinaisons de modèles 126
5.3.1 Résultats bruts 126
5.3.2 Analyse 132
5.3.3 Gradient exponentiel contre sous-jacents 133

6 Conclusion **135**
6.1 Contributions théoriques 135
6.2 Contributions expérimentales 136
6.3 Pistes futures 137

Références **139**

Liste des figures

1.1 Schéma des conventions de temps 3
1.2 Frontière efficiente et utilité quadratique 6

2.1 Rendements de l'indice TSE 300 15
2.2 Système adaptatif pour la gestion de portefeuille 19
2.3 Comparaison entre la variance simple et exponentielle 35
2.4 Erreur quadratique moyenne de l'estimation des rendements . 37

3.1 Calcul de la « passe avant » dans un MLP 48
3.2 Calcul de la « passe arrière » dans un MLP 48
3.3 Schéma d'un réseau récurrent 54
3.4 Réseau récurrent déplié à travers le temps 54
3.5 MLP comme sous-système adaptatif 56
3.6 Système adaptatif pour prévision versus décision 57
3.7 Entraînement d'un système de prévision 68
3.8 Entraînement d'un système de décision 70
3.9 Graphe de flot déplié d'un système récurrent 73
3.10 Fonction de coût non-régularisée 77
3.11 Fonction de coût régularisée 77

4.1 Validation séquentielle 87
4.2 Prédicteur naïf pour l'allocation de trois actifs 92
4.3 Pénalisation sur la norme des entrées 95
4.4 Coût de la pénalisation sur la norme des entrées 96

5.1 Exemple d'autocorrélation des rendements 111
5.2 Inspection de l'homogénéité de la variance 114
5.3 Inspection de la normalité des rendements 115
5.4 Exemple détaillé de la gestion de portefeuille 120
5.5 Effet de chaque facteur sur le rendement mensuel moyen . . . 123

5.6 Effet des hyperparamètres sur la performance 124
5.7 Résultats de trois méthodes de combinaison de modèles 130
5.8 Évolution des pondérations attribuées par un comité 131

Liste des tableaux

4.1 Résumé du processus de prise de décisions 84

5.1 Résultats pour les modèles de décision sans récurrence 117
5.2 Résultats pour les modèles de décision avec récurrence 118
5.3 Résultats pour les modèles de prévision sans récurrence 119
5.4 ANOVA pour le modèle de décision sans récurrence 125
5.5 ANOVA pour le modèle de décision avec récurrence 125
5.6 ANOVA pour le modèle de prévision sans récurrence 125
5.7 ANOVA sur les types de modèles 127
5.8 Comparaisons entre les rendements de différents modèles . . . 127
5.9 Comparaisons entre les rendements de différents modèles . . . 127
5.10 Résultat des comités pour les modèles de décision sans récurrence 129
5.11 Résultat des comités pour les modèles de décision avec récurrence 129
5.12 Résultat des comités pour les modèles de prévision sans récurrence 129
5.13 ANOVA pour le comité formé par le gradient exponentiel 132
5.14 ANOVA comparant les méthodes de combinaison de modèle . . 132
5.15 Comparaison entre le comité et le meilleur sous-jacent 134
5.16 Comparaison entre le comité et un sous-jacent ctiré au hasard 134

Chapitre 1

Introduction

L'allocation d'actifs est un problème important pour les gestionnaires de portefeuille. Ce problème consiste à partager un capital à investir entre différents actifs (par exemple, le marché boursier, les obligations corporatives, ou les obligations gouvernementales « sans risque »), de façon à maximiser un objectif financier.

Un certain nombre de méthodologies d'allocation ont été proposées et employées, dont la gestion « moderne » de portefeuille (Markowitz 1959), qui est populaire chez les praticiens, et presque complètement dominante chez les théoriciens ; cette méthodologie fut historiquement la première à asseoir l'art alchimique de la composition d'un portefeuille sur des fondements théoriques crédibles.

Plus récemment, d'autres méthodologies d'allocation, basées sur les *systèmes adaptatifs*, ont été appliquées avec succès. Les systèmes adaptatifs, tels que les réseaux de neurones (McClelland et Rumelhart 1986) sont utilisés depuis plusieurs années dans les applications financières, incluant la gestion de portefeuille (Weigend, Abu-Mostafa et Refenes 1997), mais leur usage s'est souvent trouvé confiné à celui de la *prévision* de séries chronologiques (Weigend et Gershenfeld 1993), sans considérer le but plus général d'un

système de gestion financière, qui est de *prendre des décisions* pour maximiser un profit ou une utilité économique (Moody et Wu 1997; Bengio 1997).

Ce mémoire se propose de répondre aux questions suivantes :

- Comparer les rôles possibles d'un système adaptatif en gestion de portefeuille. Le premier rôle est entièrement classique : utiliser le système adaptatif pour faire une prévision, laquelle servira ensuite à une méthode traditionnelle, telle que la gestion moderne de portefeuille (décrite plus bas) pour prendre une décision. Le second rôle est plus récent : appliquer le système adaptatif à prendre directement une décision qui maximise explicitement un objectif financier.
- Examiner le problème du choix des hyperparamètres dans l'entraînement d'un réseau de neurones, et explorer l'utilisation de méthodes de combinaison de modèles pour automatiser ce choix et le rendre plus robuste.
- Traiter le problème de la régularisation des fonctions de coût, pour obtenir des critères pour lesquels le minimum recherché est unique.
- Considérer le problème de la gestion de portefeuille dans le cadre de l'allocation par contrôle de la valeur à risque, qui est une méthodologie de gestion du risque de plus en plus fréquemment utilisée en gestion de trésorerie.

Ces contributions sont mises en perspective dans les pages qui suivent.

1.1 Définitions et notation

1.1.1 Rendements simples

Dans ce mémoire, nous ne considérons que le scénario en temps discret, dans lequel s'écoule *une période* (par exemple, une semaine ou un mois) entre les temps t et $t+1$, $t \geq 0$ et entier. Par convention, la période t est celle écoulée entre les temps $t-1$ et t.

Soit $\{P_t\}, P_t \geq 0$, le processus stochastique des prix d'un actif. Pour un t

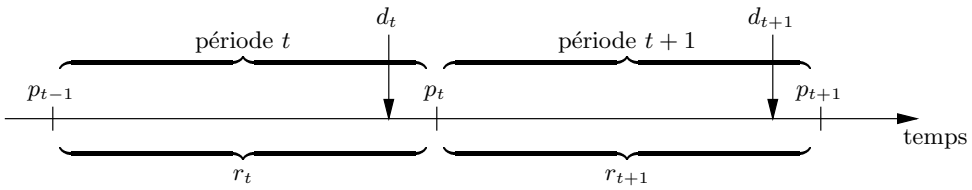


Figure 1.1: *Schéma des conventions de temps.*

donné, P_t est une variable aléatoire qui peut être mesurée étant donné l'ensemble des informations disponibles au temps t, que nous dénotons $\mathcal{I}_t$.

Définition 1.1 *Le* **rendement simple** *d'un actif à la période t est défini par*

$$R_t = \frac{P_t}{P_{t-1}} - 1. \tag{1.1}$$

Pour un actif qui verse des dividendes, nous considérons les dividendes au temps t, D_t, comme étant versés **immédiatement avant** l'enregistrement du prix P_t. Le rendement simple de l'actif tenant compte du dividende est

$$R_t = \frac{P_t + D_t}{P_{t-1}} - 1. \tag{1.2}$$

Le rendement simple est un rendement relatif qui est souvent exprimé sous forme de pourcentage. À cause de la contrainte de non-négativité des prix, $R_t \geq -1$.

Une réalisation particulière du processus des rendements est dénotée par $\{r_t\}$ (les réalisations considérées seront clairement spécifiées selon le contexte).

La figure 1.1 illustre les conventions de temps qui sont retenues.

1.1.2 Rendements d'un portefeuille

Soit un ensemble $\mathcal{A}$ de N actifs, dont les rendements simples à la période t sont donnés par $\mathbf{R}_t = (R_{1t}, R_{2t}, \ldots, R_{Nt})'$ (où $'$ dénote la transposée d'une matrice ou d'un vecteur).

Définition 1.2 *Un* **portefeuille** $\mathbf{x}_t$ *défini par rapport l'ensemble d'actifs* $\mathcal{A}$

est le vecteur des montants x_{it} investis dans chaque actif à un temps t donné :

$$\mathbf{x}_t = (x_{1t}, x_{2t}, \ldots, x_{Nt})', \quad x_{it} \in \mathbb{R} \ et \ -\infty < x_{it} < \infty. \tag{1.3}$$

Les montants x_{it} sont fonction de $\mathcal{I}_t$. Ils n'obéissent pas nécessairement à une contrainte de somme-à-un ; ils représentent des positions nettes (par exemple, en \$) prises dans chaque actif.

Le **rendement total** du portefeuille $\mathbf{x}_{t-1}$ pour la période t est

$$R_t(\mathbf{x}_{t-1}) = \sum_{i=1}^{N} x_{i(t-1)} \, R_{it} = \mathbf{x}'_{t-1}\mathbf{R}_t. \tag{1.4}$$

Ce rendement est un rendement « brut » (exprimé, par exemple, en \$), contrairement aux rendements relatifs des actifs $\mathbf{R}_t$.

Le **rendement espéré** du portefeuille $\mathbf{x}_{t-1}$ pour la période t est simplement

$$\mathrm{E}[R_t(\mathbf{x}_{t-1}) \mid \mathcal{I}_{t-1}] = \mathbf{x}'_{t-1} \, \mathrm{E}[\mathbf{R}_t \mid \mathcal{I}_{t-1}]. \tag{1.5}$$

De même que pour les rendements d'actifs simples, nous dénotons une réalisation particulière du processus des rendements d'un ensemble d'actifs par $\{\mathbf{r}_t\}$; les réalisations considérées seront clairement spécifiées selon le contexte.

1.1.3 Actifs spéciaux

Nous définissons de plus quelques « actifs » spéciaux : $\{r_{0t}\}$ dénote les rendements de l'actif sans risque (par exemple, des obligations gouvernementales à court terme), et $\{r_{Mt}\}$ dénote les rendements du « portefeuille de marché ». r_{0t} et r_{Mt} sont fonction de $\mathcal{I}_t$. Ces actifs ne font pas nécessairement partie des actifs $\mathcal{A}$ d'un portefeuille, mais sont utilisés comme référence, en particulier par les mesures d'évaluation de la performance (§ 2.5).

1.1.4 Autres notations

Dans la mesure du possible, nous respectons les conventions notationnelles suivantes, similaires à celles de Campbell, Lo et MacKinlay (1997) :

- Les matrices et les vecteurs sont typographiés en **gras** ; les scalaires sont en *italiques*. De façons générale, nous utilisons des lettres majuscules pour les matrices, et minuscules pour les vecteurs.
- Les matrices de variance–covariance (par exemple, entre les actifs d'un portefeuille) sont dénotées par $\mathbf{\Gamma}$ et $\mathbf{\Sigma}$, avec pour éléments respectifs γ_{ij} et σ_{ij}.
- Il est quelquefois nécessaire de faire appel à un vecteur qui ne contient que des 1. Nous dénotons un tel vecteur, d'une longueur appropriée au contexte, par $\boldsymbol{\iota}$.
- Les sommations utilisent les indices suivants :

  i (de 1 à N) pour les actifs d'un portefeuille.

  t (de 0 à T) pour les périodes de temps.

  m (de 0 à M) pour le nombre de « retards » dans le calcul de moyennes mobiles.

  j, k, ℓ comme indices généraux.

1.2 Gestion moderne de portefeuille

La gestion moderne de portefeuille est apparue au cours des années 1950 avec les contributions fondamentales de Markowitz (1952) et Markowitz (1959).[1] La question posée par Markowitz est la suivante : « Si l'investisseur croit posséder de l'information relative aux des rendements d'actifs individuels, comment peut-il concevoir un portefeuille optimal en utilisant cette information ? » Son apport fut de définir, pour un ensemble d'actifs, une famille de *portefeuilles efficients*, qui sont définis suivant les deux premiers moments de la distribution (jointe) du rendements des actifs.

Un portefeuille $\mathbf{x}_t$ est *efficient* s'il remplit les trois conditions suivantes (Markowitz 1959, p. 140) :[2]

[1] Pour un traitement récent, on pourra se référer à Markowitz (1987).

[2] Dans la formulation traditionnelle de Markowitz, la contrainte de positivité est aussi

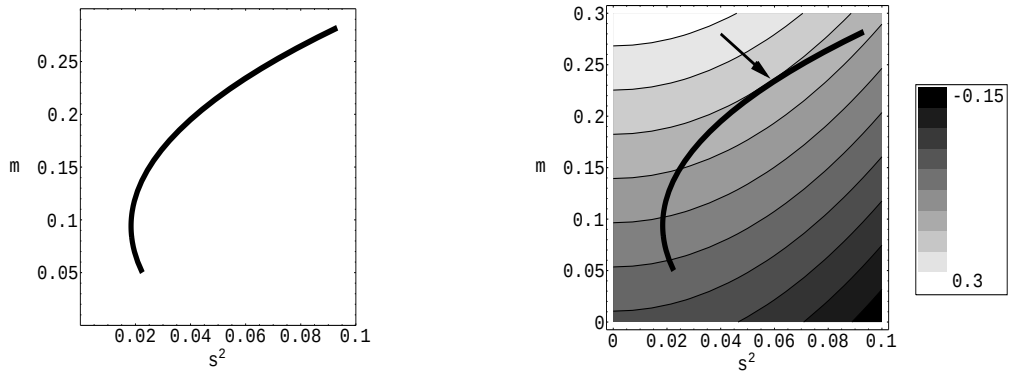


Figure 1.2: Gauche *: exemple de frontière efficiente dans le plan moyenne–variance, pour un ensemble d'actifs non-spécifié.* **Droite** *: frontière efficiente superposée avec la fonction d'utilité $U = m - 15s^2$ d'un investisseur hypothétique, où m est la moyenne et s^2 la variance du rendement du portefeuille ; le portefeuille efficient correspondant à l'utilité maximale pour l'investisseur est noté par la flèche.*

1. Tout portefeuille autre que $\mathbf{x}_t$ dont le rendement espéré est égal à celui de $\mathbf{x}_t$ aura une variance supérieure ou égale à celle de $\mathbf{x}_t$ pour la période $t+1$.
2. Tout portefeuille autre que $\mathbf{x}_t$ dont la variance est égale à celle de $\mathbf{x}_t$ aura un rendement espéré inférieur ou égal à celui de $\mathbf{x}_t$ pour la période $t+1$.

Dans le plan moyenne–variance (qui positionne un portefeuile en fonction des deux premiers moments de la distribution de ses rendements), l'ensemble des portefeuilles efficients forme une courbe appellée *frontière efficiente*, similaire à celle de la figure 1.2 (gauche). Tout point situé à droite de cette frontière représente un portefeuille inefficient, alors que tout point situé à sa gauche représente un portefeuille non-admissible ou irréalisable.

imposée, i.e. $x_{it} \geq 0$, interdisant ainsi les ventes à découvert ; nous n'imposons pas cette contrainte dans ce mémoire, pour des raisons qui se feront claires au chapitre suivant. D'autres contraintes d'admissibilité sont parfois imposées, par exemple certaines pouvant forcer le portefeuille à retourner une certaine proportion de dividentes ; nous ne nous préoccuppons pas de telles contraintes.

1.2.1 Choix de portefeuille

Dans la méthodologie classique, après avoir déterminé la frontière efficiente correspondant à un ensemble d'actifs, l'investisseur effectue le choix de portefeuille en sélectionnant le point « qu'il préfère » sur cette frontière. Ce choix est obtenu à partir de la *fonction d'utilité* de l'investisseur, qui formalise le compromis entre le rendement et le risque (la moyenne et la variance de la distribution des rendements) d'un façon spécifique à l'investisseur. Étant donné une telle fonction, l'investisseur parcourt la frontière efficiente et arrête son choix de portefeuille sur celui qui maximise l'utilité (ce point est unique pour une utilité quadratique du type fréquemment utilisé en pratique ; voir chapitre 3). Ce processus est illustré à la figure 1.2 (droite).

Cette méthodologie d'investissement est appellée, pour des raisons évidentes, *allocation moyenne–variance*, ou allocation *quadratique*.

1.2.2 Systèmes adaptatifs et allocation quadratique

L'allocation moyenne–variance ne spécifie pas comment doit être obtenue la distribution du rendement des actifs qui constituent un portefeuille. Cette distribution peut être, naïvement, la distribution historique des rendements ; elle peut aussi être le produit d'un groupe d'analystes financiers qui emploient une méthodologie établie (et postulée sur des bases empiriques), comme l'analyse fondamentale (Bodie, Kane et Marcus 1996).

Plus récemment, les systèmes adaptatifs ont été appliqués pour estimer cette distribution (Magdon-Ismail et Atiya 1998). Comme nous le verrons au chapitre 3, un système adaptatif peut jouer l'un des rôles suivants :

- Modéliser la distribution jointe du rendement de tous les actifs dans le portefeuille.
- Modéliser les quelques premiers moments inféreurs de cette distribution.

À partir de l'estimation de la distribution produite par le système adaptatif, nous pouvons extraire une prévision de la moyenne et de la variance des rendements, et l'introduire dans un système d'allocation moyenne–variance

pour obtenir un choix de portefeuille. C'est là le premier paradigme d'allocation d'actifs que nous considérons.

1.3 Systèmes adaptatifs pour la décision

Le second paradigme d'allocation d'actifs consiste à entraîner un système adaptatif à optimiser un critère de performance financière comme le rendement. Contrairement à l'allocation moyenne–variance, nous omettons ici entièrement l'étape de la prévision : le système adaptatif (lequel est, dans la majorité des applications, un réseau de neurones) produit directement les décisions d'allocation déterminées strictement en fonction du critère.

Cette voie de recherches connaît une popularité croissante depuis quelques années. Choey et Weigend (1997) ont appliqué la maximisation directe du ratio de Sharpe (Sharpe 1966) au problème restreint du choix entre un actif risqué et un sans risque, avec de bons résultats. De même, Moody et Wu (1997) l'appliquent dans un contexte d'apprentissage par renforcement. Finalement, Bengio (1997) considère cette avenue dans un problème réaliste d'allocation de 35 actifs, et obtient d'excellentes performances par rapport à un modèle entraîné à minimiser l'erreur de prévision (erreur quadratique).

Dans ce mémoire, nous comparons ces méthodes « directes » de prise de décision avec l'allocation moyenne–variance, qui demeure l'une des méthodes les plus efficaces et éprouvées utilisées en pratique.

1.4 Combinaison de modèles

Dans la mise au point de systèmes adaptatifs, il est fréquent d'entraîner un certain nombre de modèles et de choisir celui qui offre la meilleure performance sur un ensemble de validation. Cette pratique comporte plusieurs inconvénients : premièrement, l'ensemble de validation est nécessairement affecté par une certaine composante de bruit, et nous n'avons aucune garan-

tie que le modèle sélectionné par rapport à cet ensemble offrira la meilleure performance de généralisation ; deuxièmement, tout l'effort déployé pour l'entraînement d'un grand nombre de modèles est perdu lorsqu'on n'en choisit qu'un seul.

Les méthodes de combinaison de modèles contournent le problème en combinant plusieurs modèles pour former un *comité* (Perrone 1993; Perrone et Cooper 1993). Le mode d'opération d'un comité est très simple. Soit M modèles « sous-jacent » $f_m(\cdot)$ (qu'on cherche à combiner), et des pondérations w_m telles que $w_m \geq 0$ et $\sum_m w_m = 1$ (ces pondérations w_m sont déterminées par différentes méthodes de combinaison ; nous décrivons trois de ces méthodes au chapitre 4). La sortie du comité est donnée par :

$$\mathbf{y}^{\mathsf{comité}} = \sum_{m=1}^{M} w_m f_m(\mathbf{x}),$$

où $\mathbf{x}$ est un vecteur d'entrées.

Nous appliquons les comités à résoudre le problème du choix des hyperparamètres[3] dans l'entraînement des réseaux de neurones : nous regroupons dans un même comité les modèles de même topologie qui diffèrent dans les valeurs des hyperparamètres utilisés pour l'entraînement.

Nous expliquons en détails la construction de comités et dérivons plusieurs de leurs propriétés théoriques au chapitre 4.

1.5 Aperçu du mémoire

Ce mémoire se présente comme suit.

Le chapitre 2 présente les concepts de gestion de portefeuille qui sont employés au long du mémoire. Nous introduisons la notion de *valeur à risque*,

[3]Un hyperparamètre contrôle la distribution des paramètres dans un modèle ; il ne fait pas partie des paramètres « de base » à estimer, mais il gouverne l'entraînement en affectant la forme de la fonction de coût. Il est généralement déterminé par essais–erreurs, ou par une procédure de validation croisée.

qui est utilisée comme cadre d'évaluation de stratégies de placement, ainsi que comme stratégie de placement à part entière. Nous considérons de plus la nature de certains estimateurs de volatilité qui jouent un rôle crucial dans le calcul de la valeur à risque.

Le chapitre 3 introduit la théorie fondamentale d'une classe particulière de réseaux de neurones, les perceptrons multi-couches, nécessaire à la compréhension de ce mémoire. Il explicite de plus les paradigmes de *prévision* et de *décision* dans lesquels peuvent s'incorporer les systèmes adaptatifs pour la gestion de portefeuille. Nous dérivons les équations nécessaires à l'implantation et à l'entraînement efficaces de réseaux de neurones pour chaque paradigme. De plus, nous traitons le problème de la régularisation de critères financiers qui sont utilisés pour l'entraînement, de manière à garantir l'existence d'un minimum unique.

Le chapitre 4 précise les détails de l'entraînement des réseaux de neurones que nous utilisons comme systèmes adaptatifs. Nous expliquons le problème du choix des hyperparamètres nécessaires au contrôle des pénalisations sur la norme des poids et la norme des entrées, et nous traitons des méthodes de combinaison de modèles qui sont utilisées pour y remédier. Nous décrivons de plus les prétraitements appliqués aux ensembles de données d'entraînement, et dressons le schéma des expériences effectuées.

Le chapitre 5 présente de façon détaillée tous nos résultats expérimentaux, et entreprend une analyse statistique approfondie de leur signification.

Finalement, le chapitre 6 tire quelques conclusions et explore des pistes futures pour prolonger ce travail.

Chapitre 2

Valeur à risque

Ce chapitre introduit la notion de la *valeur à risque* d'un portefeuille. Il explique comment elle peut s'utiliser pour guider les décisions de placement et évaluer leur performance. Il présente de plus différentes méthodes d'estimation de cette valeur à risque pour un portefeuille donné.

2.1 Principes de valeur à risque

Définition 2.1 *La* **valeur à risque** *(VàR) avec probabilité α du portefeuille $\mathbf{x}_{t-1}$ pour la période t, est la valeur $V_t \geq 0$ telle que*

$$\Pr[\mathbf{R}_t'\mathbf{x}_{t-1} < -V_t \mid \mathcal{I}_{t-1}] = 1 - \alpha. \tag{2.1}$$

La VàR d'un portefeuille est la perte maximale que ce portefeuille peut encourir avec une probabilité α donnée, pour une certaine période de temps. La VàR donne une indication du degré de risque auquel un portefeuille est exposé. Au contraire d'autres mesures de risque, la VàR est une mesure absolue, donnée, par exemple, en dollars ; la VàR réduit le risque à un seul chiffre :

la perte maximale (en dollars) sur une certaine période, associée avec une probabilité donnée.

2.1.1 Utilisations de la VàR

La VàR est généralement utilisée de deux façons distinctes : en premier lieu, elle peut servir à calculer (à posteriori) le risque auquel a été exposé un portefeuille dans le passé ; cette mesure peut servir, par exemple, à comparer la performance de différents placements. En second lieu, la VàR peut servir à prévoir le risque auquel sera exposé un portefeuille dans le futur. Cette prévision peut permettre de choisir lequel, d'entre deux placements, offrira le rendement espéré le plus élevé pour un niveau de risque fixé.

Ces deux utilisations de la VàR sont complémentaires. La première est pertinente à l'évaluation des performances réalisées, alors que la deuxième sert dans la constitution des stratégies de placement. Nous faisons usage de ces deux points de vue dans ce mémoire.

2.2 Estimation de la VàR

La valeur à risque V_t du portefeuille $\mathbf{x}_{t-1}$ est une quantité qu'on ne peut généralement mesurer directement, car le contraire supposerait (*cf.* éq. (2.1)) une connaissance exacte de la distribution conditionnelle des rendements des actifs pour la période t, $\mathbf{R}_t$. Puisque la « véritable » distribution est généralement inconnue, toute estimation de V_t doit se faire en fonction d'un *modèle* de celle-ci.

Nous considérons les deux modèles les plus fréquemment utilisés par les praticiens de la VàR (Jorion 1997) : la distribution empirique des rendements et l'approximation normale.

2.2.1 Distribution empirique

Par simplicité, nous considérons dans cette section un portefeuille constitué d'un seul actif ; les positions (fixes) prises dans le portefeuille à chaque temps t sont notées x_t.

L'utilisation de la distribution empirique est fondée sur deux hypothèses concernant la distribution sous-jacente de l'actif. Pour déterminer l'estimateur $\hat{V}_t$ de V_t, nous supposons que :

1. La distribution conditionnelle du rendement R_t, étant donnée l'information au temps $t-1$, est **stationnaire**.
2. Les rendements sont **indépendants**, i.e. R_{t_1} est indépendant de R_{t_2}, $\forall t_1 \neq t_2$.

Sous ces hypothèses, l'emploi de la distribution empirique se justifie à partir du fait que les réalisations passées seront indicatrices du comportement statistique futur de la série.

Soit $\hat{F}_{t-1}(r)$ la fonction de répartition estimée à partir d'une réalisation $\{r_\tau\}_{\tau=0}^{t-1}$ des rendements de l'actif. Nous supposons qu'il existe une fonction inverse $\hat{F}_{t-1}^{-1}(p)$.[1]

Nous estimons la VàR de niveau α de la façon suivante. Sous les hypothèses et selon les définitions précédentes, nous avons :

$$\widehat{\Pr}[R_t < r_t \mid \mathcal{I}_{t-1}] = \hat{F}_{t-1}(r_t) \tag{2.2}$$

et

$$\widehat{\Pr}[R_t x_{t-1} < -\hat{V}_t \mid \mathcal{I}_{t-1}] = 1-\alpha. \tag{2.3}$$

Suivant la définition de l'éq. (2.1), nous choisissons $r_t = \hat{F}_{t-1}^{-1}(1-\alpha)$, d'où nous obtenons

$$\widehat{\Pr}[R_t x_{t-1} < \hat{F}_{t-1}^{-1}(1-\alpha)x_{t-1} \mid \mathcal{I}_{t-1}] = 1-\alpha, \tag{2.4}$$

[1] La fonction $\hat{F}_{t-1}(r)$ peut être estimée en utilisant des méthodes statistiques standard d'estimation de densité, comme les méthodes à noyau (Silverman 1986; Simonoff 1996).

et finalement, comparant les éq. (2.3) et (2.4),

$$\hat{V}_t = -\hat{F}_{t-1}^{-1}(1-\alpha)x_{t-1}. \tag{2.5}$$

Exemple d'application : le TSE 300

À titre d'exemple, imaginons un portefeuille constitué d'un actif imitant le comportement de l'indice TSE 300. La figure 2.1 montre la distribution empirique des rendements mensuels du TSE 300 sur une période de plus de trente ans.

À un temps t ultérieur à février 1997, nous désirons estimer la VàR d'un portefeuille investissant $x_{t-1} = 1\$$ dans cet actif, sur une échéance de un mois, avec une probabilité de 95%, $\alpha = 0.95$. Nous supposons que la distribution des rendements au cours du mois demeure inchangée par rapport à l'historique, et que les rendements sont indépendants. Nous estimons le $5^{\text{ième}}$ percentile de la distribution empirique et obtenons une valeur de -0.0646. L'estimateur de la VàR est donc :

$$\begin{aligned} \hat{V}_t &= -(-0.0646)x_t \\ &= 0.0646\ \$. \end{aligned}$$

2.2.2 Approximation normale

Portefeuille comportant un actif

Un modèle très fréquemment employé pour les calculs « simples » de la VàR d'un portefeuille est basé sur un modèle normal conditionnel de la distribution des rendements,[2] dans lequel nous supposons que, pour t donné, la variable R_t est distribuée conditionnellement à $\mathcal{I}_{t-1}$ selon

$$R_t \ \sim\ \mathcal{N}(\mu_t, \sigma_t^2), \quad \sigma_t^2 > 0, \tag{2.6}$$

[2]Ce modèle suppose que le rendement total du portefeuille peut être approximé de façon raisonnable par une distribution normale, ce qui exclut par exemple les portefeuilles composés d'options.

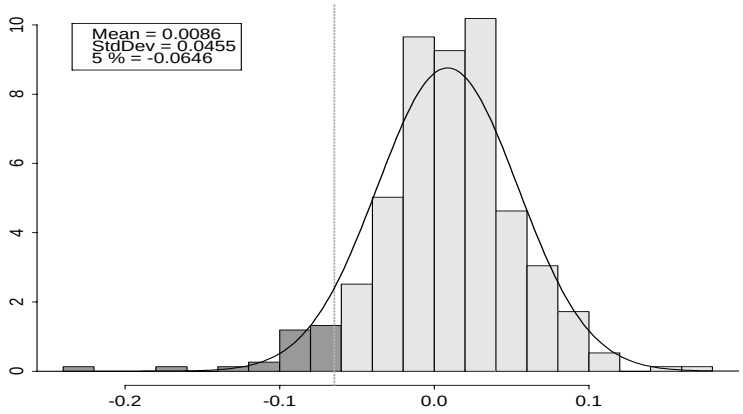


Figure 2.1: *Histogramme des rendements mensuels de l'indice TSE 300 de janvier 1965 à février 1997. La partie ombragée à gauche de la ligne verticale illustre les rendements inférieurs au 5e percentile. Une approximation normale de la distribution empirique, de moyenne et d'écart-type tels que ci-haut, est aussi fournie.*

ce qui équivaut à

$$\Pr[R_t < r_t \mid \mathcal{I}_{t-1}] = \Phi\left(\frac{r_t - \mu_t}{\sigma_t}\right), \tag{2.7}$$

où $\Phi(\cdot)$ est la fonction de répartition de la distribution normale centrée réduite, et μ_t et σ_t sont respectivement la moyenne et l'écart-type de la distribution conditionnelle des rendements.

Selon ce modèle, nous calculons ainsi la VàR V_t de niveau α : soit x_{t-1} la position (fixe) prise dans l'actif au temps $t-1$. Choisissons $r_t = \sigma_t \Phi^{-1}(1-\alpha) + \mu_t$ que nous substituons dans l'équation précédente pour obtenir

$$\Pr[R_t < \sigma_t \Phi^{-1}(1-\alpha) + \mu_t \mid \mathcal{I}_{t-1}] = 1 - \alpha, \tag{2.8}$$

d'où

$$\Pr[R_t x_{t-1} < (\sigma_t \Phi^{-1}(1-\alpha) + \mu_t) x_{t-1} \mid \mathcal{I}_{t-1}] = 1 - \alpha, \tag{2.9}$$

et, comparant les éq. (2.3) et (2.9),

$$\begin{aligned} V_t &= -(\sigma_t \Phi^{-1}(1-\alpha) + \mu_t) x_{t-1} \\ &= (\sigma_t \Phi^{-1}(\alpha) - \mu_t) x_{t-1}, \end{aligned} \tag{2.10}$$

utilisant le fait que $\Phi^{-1}(1-\alpha) = -\Phi^{-1}(\alpha)$ à cause de la symétrie de la distribution normale.

Estimation de V_t Soit $\hat{\mu}_t$ et $\hat{\sigma}_t$ des estimateurs des paramètres de la distribution de R_t, calculés selon l'information $\mathcal{I}_{t-1}$ (nous traitons du choix des estimateurs à la section 2.6). L'estimateur de V_t est donné par :

$$\hat{V}_t = (\hat{\sigma}_t \Phi^{-1}(\alpha) - \hat{\mu}_t) x_{t-1}. \tag{2.11}$$

Si les estimateurs $\hat{\mu}_t$ et $\hat{\sigma}_t$ sont non-biaisés, alors $\hat{V}_t$ est sans biais, car :

$$\begin{aligned} \mathrm{E}[\hat{V}_t \mid \mathcal{I}_{t-1}] &= \mathrm{E}[(\hat{\sigma}_t \Phi^{-1}(\alpha) - \hat{\mu}_t) x_{t-1} \mid \mathcal{I}_{t-1}] \\ &= (\mathrm{E}[\hat{\sigma}_t \mid \mathcal{I}_{t-1}] \Phi^{-1}(\alpha) - \mathrm{E}[\hat{\mu}_t \mid \mathcal{I}_{t-1}]) x_{t-1} \\ &= (\sigma_t \Phi^{-1}(\alpha) - \mu_t) x_{t-1} \\ &= V_t. \end{aligned}$$

Portefeuille comportant plusieurs actifs

Le modèle normal précédent s'étend naturellement au cas à plusieurs actifs. Soit les rendements des actifs pour la période t distribués conditionnellement à $\mathcal{I}_{t-1}$ selon

$$\mathbf{R}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Gamma}_t), \tag{2.12}$$

où $\boldsymbol{\Gamma}_t$ est définie-positive, et soit $\mathbf{x}_{t-1}$ les positions (fixes) prises dans chaque actif au temps $t-1$. Nous obtenons la VàR de niveau α pour la période t comme étant

$$V_t = \Phi^{-1}(\alpha) \sqrt{\mathbf{x}'_{t-1} \boldsymbol{\Gamma}_t \mathbf{x}_{t-1}} - \boldsymbol{\mu}'_t \mathbf{x}_{t-1}. \tag{2.13}$$

Dans certaines circonstances (entres autres, lorsqu'on considère des actifs boursiers sur de courtes échéances), les rendements espérés des actifs sont près

de zéro. En posant $\boldsymbol{\mu}_t = \mathbf{0}$, nous pouvons simplifier l'équation précédente à

$$V_t = \Phi^{-1}(\alpha)\sqrt{\mathbf{x}'_{t-1}\boldsymbol{\Gamma}_t\,\mathbf{x}_{t-1}}. \tag{2.14}$$

Estimation de V_t Soit $\hat{\boldsymbol{\mu}}_t$ et $\hat{\boldsymbol{\Gamma}}_t$ des estimateurs des paramètres de la distribution normale multivariée qui modélise les rendements $\mathbf{R}_t$, calculés selon l'information $\mathcal{I}_{t-1}$. L'estimateur de la VàR suivant ce modèle est une simple extension de l'éq. (2.11) :

$$\hat{V}_t = \Phi^{-1}(\alpha)\sqrt{\mathbf{x}'_{t-1}\hat{\boldsymbol{\Gamma}}_t\,\mathbf{x}_{t-1}} - \hat{\boldsymbol{\mu}}_t{}'\,\mathbf{x}_{t-1}. \tag{2.15}$$

L'estimateur de V_t lorsque les rendements espérés sont nuls est :

$$\hat{V}_t = \Phi^{-1}(\alpha)\sqrt{\mathbf{x}'_{t-1}\hat{\boldsymbol{\Gamma}}_t\,\mathbf{x}_{t-1}} \tag{2.16}$$

Exemple (suite)

Nous concluons l'exemple du calcul de la VàR pour le portefeuille imitant le comportement du TSE 300 en démontrant l'utilisation de l'approximation normale. Nous estimons la moyenne et l'écart-type historiques comme étant $\hat{\mu} = 0.0086$ et $\hat{\sigma} = 0.0455$.

Utilisant ces estimateurs conjointement à l'éq. (2.10), nous obtenons pour une VàR à $\alpha = 0.95$ sur un mois ($x_{t-1} = 1\$$), pour t ultérieur à février 1997 :

$$\begin{aligned}\hat{V}_t &= (\hat{\sigma}\Phi^{-1}(0.95) - \hat{\mu})x_{t-1} \\ &= 0.0455 \times 1.645 - 0.0086 \\ &= 0.0662\ \$,\end{aligned}$$

où la valeur -1.645 correspond au $5^{\text{ième}}$ percentile de la distribution normale standardisée. Des valeurs appropriées devront être utilisées pour connaître la VàR à d'autres probabilités.[3]

[3]Dans le présent mémoire, nous nous sommes limités aux VàR à 95%, car l'approximation normale de la distribution des rendements perd généralement de sa validité pour des probabilités plus élevés (RiskMetrics 1996).

2.3 La VàR comme modèle de placement

La présentation précédente de la VàR s'effectuait dans le contexte restreint où le portefeuille était déterminé à priori, et où on ne cherchait qu'à estimer passivement la VàR. Il est aussi possible de l'appliquer à un contexte plus actif, où nous l'utilisons pour contrôler le risque encouru par un portefeuille.

Nous présentons ici un modèle de placement basé sur l'emploi de la VàR pour obtenir un contrôle actif du risque :

1. À chaque temps t, une *VàR cible* $\tilde{V}_{t+1}$ est fixée (par exemple, par le gestionnaire de portefeuille). Le but de notre stratégie est de construire un portefeuille $\mathbf{x}_t$ ayant cette VàR.
2. Nous consultons un système adaptatif, tel qu'un réseau de neurones, pour obtenir des *recommandations* de placement parmi un ensemble de N actifs possibles. Ces recommandations prennent la forme d'un vecteur $\mathbf{y}_t$ donnant les *pondérations relatives* que les actifs devraient avoir dans le portefeuille. Il n'y a aucune contrainte sur les y_{it}, par exemple de positivité ou de somme à 1.
3. Nous *rééchelonnons* les recommandations $\mathbf{y}_t$ par un facteur homogène (voir ci-bas) pour produire les positions finales $\mathbf{x}_t$ (en dollars) à prendre sur chaque actif au début de la période t. Cet ajustement est effectué de telle manière que l'estimateur $\hat{V}_{t+1|t}$ (calculé au temps t) de la VàR du portefeuille $\mathbf{x}_t$ au cours de la période $t+1$ soit égal à la VàR cible $\tilde{V}_{t+1}$.
4. Nous investissons au temps t dans les positions $\mathbf{x}_t$ pendant exactement une période, en empruntant la somme nécessaire $\sum_{i=1}^{N} x_{it}$ au taux sans risque r_{0t}.

La figure 2.2 illustre l'implantation de ces étapes, de même que l'évaluation de la performance décrite à la section 2.5.

Il est à noter que ce modèle de placement ne correspond pas à une notion classique d'« investissement », pour deux raisons.

Premièrement, nous ne disposons pas d'un capital initial qu'on doit partager entre plusieurs actifs ; selon le modèle présenté ici, la position *nette* de l'investisseur dans tous les actifs (en incluant l'actif sans risque duquel on

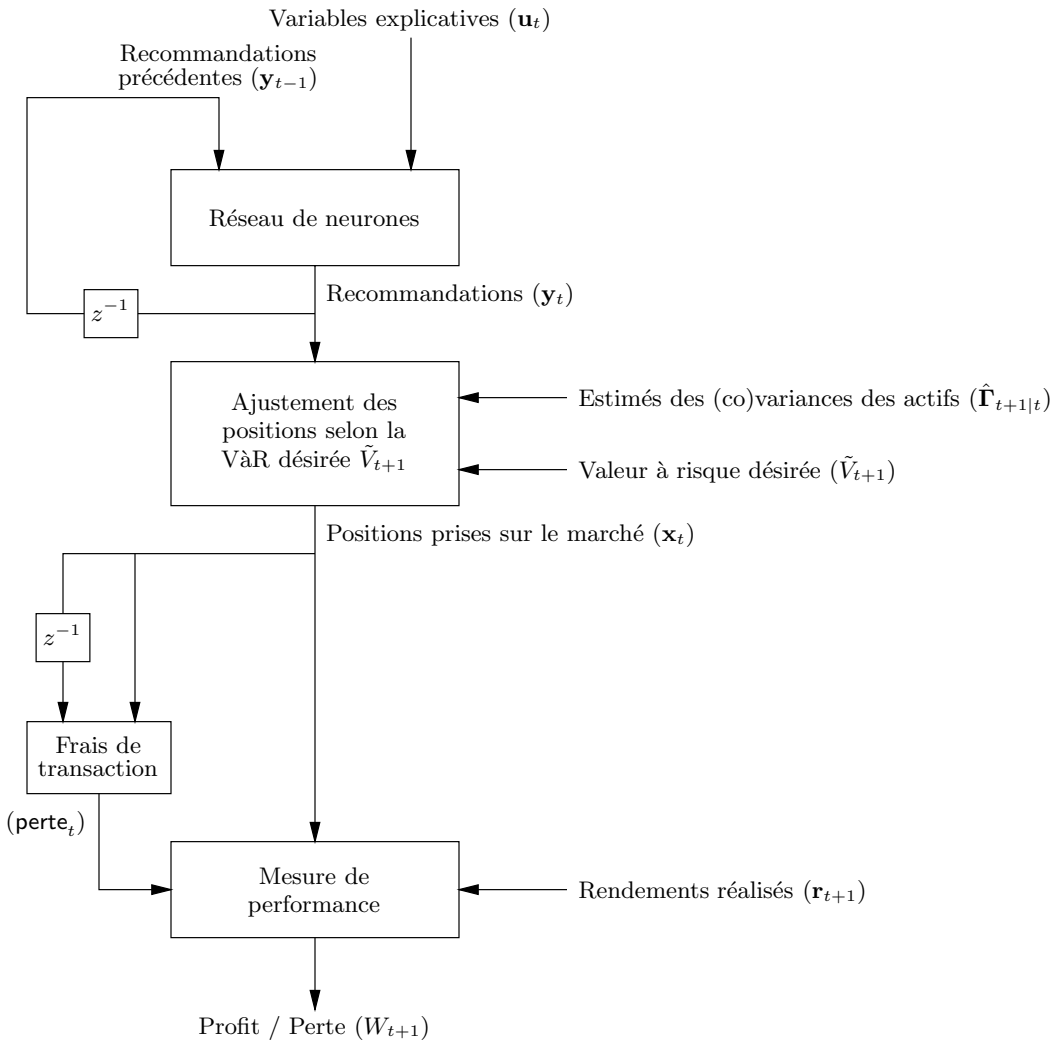


Figure 2.2: *Utilisation d'un système adaptatif, comme un réseau de neurones, dans un paradigme de gestion de portefeuille basé sur le contrôle de la VàR. Les z^{-1} représentent des délais unitaires.*

emprunte) est nulle au début de chaque période. De plus, rien ne contraint l'investisseur à des positions x_{it} positives ; des x_{it} négatifs correspondent simplement à des ventes à découvert, pour lesquelles nous supposons qu'il n'existe aucune restriction réglementaire.

Deuxièmement, les profits générés à une période ne sont pas systématiquement réinvestis au cours de la période suivante. La raison en est que l'unique facteur qui détermine la taille du placement à la période t est la VàR désirée $\tilde{V}_{t+1}$; les profits dégagés au cours des périodes précédentes ne sont pas considérés dans cette opération. (Évidemment, les profits générés comptent pour beaucoup dans l'évaluation de la performance ; nous verrons plus bas comment comparer de façon réaliste la performance de deux systèmes adaptatifs utilisant cette stratégie de placement).

2.4 Équations de rééchelonnement

À partir de recommandations $\mathbf{y}_t$ données (pour t fixé), rendues par le réseau de neurones, nous souhaitons les ajuster pour obtenir une position finale $\mathbf{x}_t$ dont l'estimé par rapport à $\mathcal{I}_t$ de la VàR, $\hat{V}_{t+1}^{(\mathbf{x}_t)}$, est égal, idéalement, à la VàR cible $\tilde{V}_{t+1}$:

$$\mathrm{E}[\hat{V}_{t+1}^{(\mathbf{x}_t)}] = \tilde{V}_{t+1}. \tag{2.17}$$

Pour simplifier les calculs de la valeur à risque, nous faisons l'hypothèse que l'espérance des rendements des actifs est nulle, $\mathrm{E}[\mathbf{R}_{t+1}|\mathcal{I}_t] = \mathbf{0}$.

Proposition 2.1 *Si les rendements des actifs du portefeuille pour la période $t+1$ sont distribués conditionnellement à $\mathcal{I}_t$ selon*

$$\mathbf{R}_{t+1} \;\sim\; \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}_{t+1}), \tag{2.18}$$

avec $\mathbf{\Gamma}_{t+1}$ définie-positive, alors le rééchelonnement d'une recommandation $\mathbf{y}_t$ (fixée), supposant $\|\mathbf{y}_t\| > 0$, donné par

$$\mathbf{x}_t = \beta_t \, \mathbf{y}_t, \tag{2.19}$$

où

$$\beta_t = \frac{\tilde{V}_{t+1}}{\Phi^{-1}(\alpha)\sqrt{\mathbf{y}_t' \, \mathbf{\Gamma}_{t+1} \, \mathbf{y}_t}} \tag{2.20}$$

produit un portefeuille $\mathbf{x}_t$ *dont la VàR de niveau* α, $V_{t+1}^{(\mathbf{x}_t)}$, *est* $\tilde{V}_{t+1}$, *conditionnellement à* $\mathcal{I}_t$.

Preuve Par hypothèse, $\mathbf{y}_t' \, \mathbf{\Gamma}_{t+1} \, \mathbf{y}_t > 0$, car $\mathbf{\Gamma}_{t+1}$ est supposée définie-positive et $\|\mathbf{y}_t\| > 0$.

Substituant $\mathbf{x}_t$ défini par les éq. (2.19) et (2.20) dans (2.14), nous obtenons

$$\begin{aligned} V_{t+1} &= \Phi^{-1}(\alpha)\sqrt{\left(\frac{\tilde{V}_{t+1}}{\Phi^{-1}(\alpha)\sqrt{\mathbf{y}_t'\mathbf{\Gamma}_{t+1}\mathbf{y}_t}}\right)\mathbf{y}_t'\mathbf{\Gamma}_{t+1}\left(\frac{\tilde{V}_{t+1}}{\Phi^{-1}(\alpha)\sqrt{\mathbf{y}_t'\mathbf{\Gamma}_{t+1}\mathbf{y}_t}}\right)\mathbf{y}_t} \\ &= \tilde{V}_{t+1}. \end{aligned}$$

■

2.4.1 Estimateur de β_t

Le facteur de rééchelonnement β_t peut être estimé simplement en remplaçant la matrice de covariance $\mathbf{\Gamma}_{t+1}$ dans l'éq. (2.20) par un estimateur :

$$\hat{\beta}_t = \frac{\tilde{V}_{t+1}}{\Phi^{-1}(\alpha)\sqrt{\mathbf{y}_t' \, \hat{\mathbf{\Gamma}}_{t+1} \, \mathbf{y}_t}}. \tag{2.21}$$

Malheureusement, même si $\hat{\mathbf{\Gamma}}_{t+1}$ est sans biais, $\hat{\beta}_t$ se trouve biaisé (car en général, pour une variable aléatoire $X > 0$, $\mathrm{E}[1/X] \neq 1/\mathrm{E}[X]$). Nous analysons ce biais dans la section suivante, démontrons qu'il est asymptotiquement nul, et proposons un nouvel estimateur qui corrige pour le biais en échantillon fini.

2.4.2 Distribution échantillonale de $\hat{\beta}_t$

Dans cette section, nous considérons la distribution échantillonale de $\hat{\beta}_t$ pour *un seul actif*. Nous supposons que les rendements de l'actif sont de

moyenne nulle et distribués i.i.d. selon une normale :

$$R_t \ \sim \ \mathcal{N}(0, \sigma^2), \quad \sigma > 0, \quad 1 \leq t \leq T. \tag{2.22}$$

Soit s_t^2 un estimateur non-biaisé de σ^2, calculé selon l'information $\mathcal{I}_t$,

$$s_t^2 = \frac{1}{t} \sum_{\tau=1}^{t} r_\tau^2. \tag{2.23}$$

(Nous divisons par t plutôt que par $t-1$ car la moyenne est posée et non estimée ; cet estimateur de la variance est non-biaisé dans ce cas).

Selon la théorie échantillonale normale standard, nous avons

$$\frac{t}{\sigma^2} s_t^2 \ \sim \ \chi_t^2, \tag{2.24}$$

avec pour fonction de densité

$$f_{\chi_t^2}(x) = \frac{1}{2^{t/2}\Gamma(t/2)} e^{-\frac{x}{2}} x^{\frac{t}{2}-1}, \quad x > 0, \tag{2.25}$$

et la fonction gamma $\Gamma(z)$ définie par l'intégrale

$$\Gamma(z) = \int_0^\infty t^{z-1}\, e^{-t}\, dt. \tag{2.26}$$

Cette fonction obéit à la récurrence $\Gamma(z+1) = z\,\Gamma(z)$.

Distribution de $Y = 1/X$

Soit $X > 0$ une variable aléatoire continue ayant une fonction de répartition $F_X(x) = \Pr[X < x]$. La variable aléatoire $Y = h(X) = 1/X$ a pour fonction de répartition :

$$\begin{aligned} \Pr[h(X) < y] &= \Pr[h^{-1}(h(X)) > h^{-1}(y)] \\ &= \Pr[X > x] \\ &= 1 - \Pr[X < x], \quad x, y > 0, \end{aligned}$$

où la première étape découle de la monotonicité décroissante de $h(\cdot)$. Donc,

$$F_Y(y) = 1 - F_X(1/y). \tag{2.27}$$

Soit $f_X(\cdot)$ la densité de X. La densité de Y est :

$$\begin{aligned} f_Y(y) &= \frac{dF_Y(y)}{dy} \\ &= -\frac{dF_X(1/y)}{dy} \\ &= \frac{f_X(1/y)}{y^2}, \quad y > 0. \end{aligned} \tag{2.28}$$

Distribution de $Y = \sqrt{X}$

Suivant un développement semblable à celui de la section précédente, nous trouvons pour $Y = \sqrt{X}$ (et $X > 0$)

$$F_Y(y) = F_X(y^2), \quad y > 0, \tag{2.29}$$

d'où

$$f_Y(y) = 2y f_X(y^2), \quad y > 0. \tag{2.30}$$

Comportement asymptotique de l'espérance inverse

Définissons les variables $U_t, W_t > 0$ suivantes, pour t fixé :

$$U_t = \sqrt{\frac{t}{\sigma^2} s_t^2} \qquad \text{et} \qquad W_t = \frac{1}{U_t}. \tag{2.31}$$

Appliquant les transformations de variables aléatoires dérivées ci-haut, nous trouvons comme densité de U_t :

$$f_{U_t}(x) = 2x f_{\chi_t^2}(x^2) = \frac{1}{2^{\frac{t}{2}-1}\Gamma(t/2)} e^{-\frac{x^2}{2}} x^{t-1}, \tag{2.32}$$

et comme densité de W_t :

$$f_{W_t}(x) = \frac{f_{U_t}(1/x)}{x^2} = \frac{1}{2^{\frac{t}{2}-1}\Gamma(t/2)} e^{-\frac{x^{-2}}{2}} x^{-t-1}. \tag{2.33}$$

Les espérances de U_t et W_t sont, respectivement,

$$\begin{aligned} \mathrm{E}[U_t] &= \int_0^\infty x\, f_{U_t}(x)\, dx & \mathrm{E}[W_t] &= \int_0^\infty x\, f_{W_t}(x)\, dx \\ &= \frac{\sqrt{2}\,\Gamma\left(\frac{t+1}{2}\right)}{\Gamma\left(\frac{t}{2}\right)} & &= \frac{\Gamma\left(\frac{t-1}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{t}{2}\right)} \end{aligned} \tag{2.34}$$

d'où nous déduisons, tel que mentionné précédemment,

$$\frac{1}{\mathrm{E}[U_t]} = \frac{\Gamma\left(\frac{t}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{t+1}{2}\right)} \neq \frac{\Gamma\left(\frac{t-1}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{t}{2}\right)} = \mathrm{E}[W_t] = \mathrm{E}\left[\frac{1}{U_t}\right]. \tag{2.35}$$

Proposition 2.2 *Dans la limite où $t \to \infty$, $1/\,\mathrm{E}[U_t] = \mathrm{E}[W_t]$.*

Preuve Nous souhaitons montrer que

$$\lim_{t\to\infty} \frac{1/\,\mathrm{E}[U_t]}{\mathrm{E}[W_t]} = 1. \tag{2.36}$$

Substituant dans l'éq. (2.34), et posant $u + \frac{1}{2} = t/2$, nous obtenons :

$$\begin{aligned} \lim_{t\to\infty} \frac{1/\,\mathrm{E}[U_t]}{\mathrm{E}[W_t]} &= \lim_{u\to\infty} \frac{\Gamma^2(u+\frac{1}{2})}{\Gamma(u+1)\Gamma(u)} \\ &= \lim_{u\to\infty} \frac{\Gamma^2(u+\frac{1}{2})}{u\,\Gamma^2(u)}, \end{aligned} \tag{2.37}$$

et, prenant $\sqrt{\cdot}$ des deux côtés,

$$\begin{aligned} \lim_{t\to\infty} \sqrt{\frac{1/\,\mathrm{E}[U_t]}{\mathrm{E}[W_t]}} &= \lim_{u\to\infty} \frac{\Gamma(u+\frac{1}{2})}{\sqrt{u}\,\Gamma(u)}, \\ &= \lim_{u\to\infty} \frac{\sqrt{u}}{\sqrt{u}} \left(1 - \frac{1}{8u} + \frac{1}{128u^2} + \frac{5}{1024u^3} - \frac{21}{32768u^4} + \cdots\right) \\ &= 1, \end{aligned} \tag{2.38}$$

où l'expansion en série de $\Gamma(u+\frac{1}{2})/\Gamma(u)$ est bien connue (Graham, Knuth et Patashnik 1994). Comme nous voulons le démontrer, ce résultat établit bien que $\lim_{t\to\infty} 1/\operatorname{E}[U_t] = \lim_{t\to\infty} \operatorname{E}[1/U_t]$. ■

Absence asymptotique de biais de $\hat{\beta}_t$

Proposition 2.3 *L'estimateur de β_t (pour le cas à un seul actif),*

$$\hat{\beta}_t = \frac{\tilde{V}_{t+1}}{\Phi^{-1}(\alpha)} \frac{1}{|y_t| \sqrt{s_t^2}}, \tag{2.39}$$

est asymptotiquement non-biaisé.

Preuve Nous faisons appel au résultat précédent. Nous avons

$$\begin{aligned}
\lim_{t\to\infty} \operatorname{E}[\hat{\beta}_t] &= \lim_{t\to\infty} \frac{\tilde{V}_{t+1}}{\Phi^{-1}(\alpha)\,|y_t|} \operatorname{E}\left[\frac{1}{\sqrt{s_t^2}}\right] \\
&= \lim_{t\to\infty} \frac{\tilde{V}_{t+1}}{\Phi^{-1}(\alpha)\,|y_t|} \frac{1}{\operatorname{E}[\sqrt{s_t^2}]} \\
&= \frac{\tilde{V}_{t+1}}{\Phi^{-1}(\alpha)\,|y_t|} \frac{1}{\sigma} \\
&= \beta_t,
\end{aligned}$$

ce qui établit l'absence de biais asymptotique de $\hat{\beta}_t$. ■

Correction du biais en échantillon fini

Nous pouvons aussi construire un estimateur de β_t qui ne souffre pas de biais en échantillon fini, comme le démontre le corollaire suivant.

Corollaire 2.4 *Sous les hypothèses de normalité, de stationarité et d'indépendance énoncées précédemment, l'estimateur de β_t,*

$$\tilde{\beta}_t = \left(\frac{\tilde{V}_{t+1}}{\Phi^{-1}(\alpha)} \frac{1}{|y_t| \sqrt{s_t^2}}\right) \left(\sqrt{\frac{2}{t}} \frac{\Gamma(\frac{t}{2})}{\Gamma(\frac{t-1}{2})}\right), \tag{2.40}$$

est sans biais.

Preuve Utilisant le résultant préalable concernant la distribution échantillonale de la quantité $\sqrt{\sigma^2/(t\,s_t^2)}$, nous multiplions l'équation précédente par $\sqrt{\sigma^2/t}$ et prenons l'espérance :

$$\begin{aligned}
\mathrm{E}\left[\sqrt{\frac{\sigma^2}{t}}\,\tilde{\beta}_t\right] &= \mathrm{E}\left[\left(\frac{\tilde{V}_{t+1}}{\Phi^{-1}(\alpha)\,|y_t|}\right)\left(\sqrt{\frac{2}{t}}\frac{\Gamma(\frac{t}{2})}{\Gamma(\frac{t-1}{2})}\right)\left(\sqrt{\frac{\sigma^2}{t\,s_t^2}}\right)\right] \\
&= \left(\frac{\tilde{V}_{t+1}}{\Phi^{-1}(\alpha)\,|y_t|}\right)\left(\sqrt{\frac{2}{t}}\frac{\Gamma(\frac{t}{2})}{\Gamma(\frac{t-1}{2})}\right)\left(\frac{\Gamma(\frac{t-1}{2})}{\sqrt{2}\Gamma(\frac{t}{2})}\right) \\
&= \frac{\tilde{V}_{t+1}}{\sqrt{t}\,\Phi^{-1}(\alpha)\,|y_t|}\,,
\end{aligned}$$

d'où

$$\begin{aligned}
\mathrm{E}[\tilde{\beta}_t] &= \frac{\tilde{V}_{t+1}}{\Phi^{-1}(\alpha)\,|y_t|\,\sigma} && (2.41)\\
&= \beta_t, && (2.42)
\end{aligned}$$

ce qui établit bien que $\tilde{\beta}_t$ est sans biais. ■

En pratique, nous devons travailler avec des échantillons de longueur finie, mais de taille suffisante pour ne pas remarquer un biais évident dans $\hat{\beta}_t$; nous employons l'estimateur (2.21) sans modification, et observons un comportement raisonnable.

2.5 Évaluer la performance selon la VàR

Une mesure de performance doit fournir un cadre réaliste pour comparer plusieurs systèmes « concurrents » de prise de décision. Dans le cas d'une stratégie basée sur la VàR, nous désirons mettre en opposition différents modèles (par exemple, plusieurs réseaux de neurones) offrant leurs recommandations.

L'une des thèses fondamentales de la théorie financière moderne est le compromis inhérant existant entre le rendement et le risque : on ne peut

généralement pas augmenter le rendement d'un placement sans simultanément augmenter le risque de pertes auquel il est exposé (Bodie, Kane et Marcus 1996). Une mesure de performance adéquate doit donc tenir compte conjointement des deux aspects du rendement et du risque.

2.5.1 Mesures classiques de performance

Certaines approches classiques d'évaluation de la performance d'un placement considèrent le « rendement total » d'un actif initial au cours de la période complète couvrant le placement ; un exemple simple est le taux de rendement composé moyen annualisé, qui est défini comme suit pour un placement effectué sur un total de N mois :

$$r_{\mathsf{ann}} = \sqrt[N/12]{\frac{X_{\mathsf{final}}}{X_{\mathsf{initial}}}} - 1,$$

où X_{final} est la valeur finale du placement, et X_{initial} est la valeur initiale (nous supposons ces valeurs mesurées à l'échéance, donc fixées).

Cependant, cette mesure souffre d'inconvénients majeurs pour la stratégie de placement basée sur le contrôle actif de la VàR telle que décrite plus haut. Premièrement, la stratégie ne fait même pas appel à la notion d'un « actif initial » qui est réinvesti de période en période ; elle détermine, au début de chaque période, la somme à investir qui rencontre les contraintes de valeur à risque, et calcule le profit à la fin de la période. Ce profit n'est pas réinvesti à la période suivante. Donc, pour cette stratégie, la notion de taux de rendement *composé* n'a pas de sens.

La méthode d'évaluation de la performance que nous retenons rejoint une mesure bien connue en finance, celle du *ratio de Sharpe*, qui considère de façon indépendante les rendements réalisés à chaque période. Soit $\{r_{Pt}\}_{t=1}^{T}$ les taux de rendement réalisés par le portefeuille P (lequel peut changer de période en période), et $\{r_{0t}\}_{t=1}^{T}$ les taux de rendement de l'actif sans risque. Ces rendements sont supposés connus *ex post*, c'est-à-dire mesurés par rapport à l'information $\mathcal{I}_T$.

Le ratio de Sharpe est simplement le taux de rendement moyen du por-

tefeuille en surplus de l'actif sans risque, normalisé par une mesure de risque (Sharpe 1966; Sharpe 1994; Bodie, Kane et Marcus 1996) :

$$\widehat{SR} = \frac{\bar{r}_P - \bar{r}_0}{\hat{\sigma}_P}, \tag{2.43}$$

où $\bar{r}_P = \frac{1}{T}\sum_{t=1}^{T} r_{Pt}$ et $\bar{r}_0 = \frac{1}{T}\sum_{t=1}^{T} r_{0t}$. La mesure de risque retenue par ce ratio est la variance empirique des rendements du portefeuille,

$$\hat{\sigma}_P = \frac{1}{T-1}\sum_{t=1}^{T}(r_{Pt} - \bar{r}_P)^2.$$

2.5.2 Une mesure propre à l'allocation selon la VàR

Dans le cadre d'allocation selon le contrôle de la VàR, nous avons retenu une modification du ratio de Sharpe, qui utilise la VàR du portefeuille comme mesure de risque, et dans laquelle nous tenons explicitement compte des frais de transaction. Le rendement total d'une stratégie S est simplement la moyenne arithmétique du profit dégagé à chaque période :

$$W^S = \frac{1}{T}\sum_{t=1}^{T} W_t^S, \tag{2.44}$$

où W_t^S est le profit (aléatoire) net dégagé par la stratégie S à la période t, déterminé comme suit (nous donnons l'équation de W_{t+1}^S pour alléger la notation) :

$$W_{t+1}^S = \frac{(\mathbf{R}_{t+1} - \boldsymbol{\iota}\, r_{0t})'\mathbf{x}_t^S + \mathsf{perte}_t}{V_{t+1}}, \tag{2.45}$$

où $\mathbf{x}_t^S$ est le portefeuille choisi par la stratégie S au temps t. Le numérateur de W_{t+1}^S donne le profit net (en dollars) obtenu au cours de la période. Il est composé de trois parties : premièrement, il calcule le rendement obtenu par chaque actif au cours de la période, pondéré par sa proportion dans le portefeuille ; ensuite, il incorpore les frais d'emprunt du montant $\mathbf{x}_t^S$, au taux sans risque r_{0t} prévalant au début de la période ; finalement, il inclut les pertes occasionnées par les frais de transaction, telles que décrites plus bas.

Il est nécessaire de normaliser le profit par la valeur à risque V_{t+1}, car l'éq. (2.20) montre clairement qu'il est possible d'obtenir un profit (en dollars) aussi grand que désiré en ayant un V_{t+1} suffisamment grand.

Estimation des W^S et W_t^S

Pour estimer les quantités W^S et W_t^S, nous substituons des estimateurs calculés selon l'information disponible à la fin de la dernière période, $\mathcal{I}_T$:

$$\hat{W}^S = \frac{1}{T}\sum_{t=1}^{T} \hat{W}_t^S \tag{2.46}$$

et

$$\hat{W}_{t+1}^S = \frac{(\mathbf{r}_{t+1} - \boldsymbol{\iota}\, r_{0t})'\mathbf{x}_t^S + \mathsf{perte}_t}{\tilde{V}_{t+1}}, \tag{2.47}$$

où nous utilisons les rendements réalisés par les actifs, $\{\mathbf{r}_t\}$, et faisons appel à la VàR cible $\tilde{V}_{t+1}$ comme estimateur de V_{t+1}. Tout comme pour l'estimateur $\hat{\beta}_t$ de la section 2.4, nous ignorons le biais en échantillon fini associé à $\hat{W}_{t+1}^S$ car il est peu important pour les tailles d'échantillon que nous utilisons en pratique.

2.5.3 Frais de transaction

Les frais de transactions perte_t sont modélisés par une simple perte multiplicative :

$$\mathsf{perte}_t = -\mathbf{c}'\,|\mathbf{x}_t - \tilde{\mathbf{x}}_t| \tag{2.48}$$

où $\mathbf{c} = (c_1, \ldots, c_N)'$, c_i la perte relative associée à un changement de position (en dollars) sur l'actif i, et $\tilde{\mathbf{x}}_t$ la position du portefeuille dans chaque actif *immédiatement avant* que la transaction ne soit effectuée au temps t. Il est à noter que cette position est différente de celle prévalant au temps précédent, à cause des rendements $\mathbf{r}_t$ entre les temps $t-1$ et t :

$$\tilde{x}_{it} = (r_{it} + 1)\, x_{i(t-1)}. \tag{2.49}$$

2.6 Calcul de la volatilité entre les actifs

Comme le montre l'éq. (2.20), la matrice $\mathbf{\Gamma}_t$ des variances et covariances des actifs joue un rôle fondamental dans le calcul de la VàR fondé sur l'approximation normale : c'est cette matrice qui joue le rôle de modèle de volatilité des actifs. Il est donc d'une grande importance de choisir un bon estimateur $\hat{\mathbf{\Gamma}}_t$, selon les facteurs suivants :

- Nous considérons les critères habituels du choix d'estimateurs statistiques, c'est-à-dire l'absence de biais, l'efficacité (variance minimale), et la consistance asymptotique.
- Nous introduisons le critère supplémentaire de répondre « rapidement » aux changements éventuels (non-stationarités) dans la distribution des rendements des actifs sur la période considérée, $\{\mathbf{R}_t\}_{t=1}^{T}$.

Il est à noter que tous ces objectifs ne peuvent être remplis simultanément ; par exemple, un estimateur tenant compte des non-stationarités sera probablement moins efficace qu'un autre qui suppose la stationarité, pour le cas où la distribution des rendements est effectivement stationnaire.

2.6.1 Définition

La matrice $\mathbf{\Gamma}_t$ que nous désirons estimer (sous le modèle normal des rendements) est définie ainsi

$$\mathbf{\Gamma}_t = \mathrm{E}[(\mathbf{R}_t - \boldsymbol{\mu}_t)(\mathbf{R}_t - \boldsymbol{\mu}_t)' \mid \mathcal{I}_{t-1}] \tag{2.50}$$

où $\boldsymbol{\mu}_t = \mathrm{E}[\mathbf{R}_t \mid \mathcal{I}_{t-1}]$.

2.6.2 Modèle localement constant

L'emploi des moyennes mobiles simples et exponentielles présentées ci-bas se justifie à partir d'un modèle *localement constant* de la matrice $\mathbf{\Gamma}_t$, dans lequel nous supposons que, pour tout τ situé dans l'intervalle $t_1 \leq \tau \leq t_2$:

$$\mathbf{\Gamma}_\tau = \mathbf{A}, \tag{2.51}$$

où **A** est une constante sur l'intervalle. (Cette matrice peut évidemment varier d'un intervalle à l'autre).

Ce modèle nous fournit une méthode extrêmement simple de calculer l'estimateur $\hat{\mathbf{\Gamma}}_{t+k}$ à partir de l'estimateur $\hat{\mathbf{\Gamma}}_t$ calculé en fonction de $\mathcal{I}_t$ (posant $t_1 = t$ et $t_2 = t + k$) :

$$\hat{\mathbf{\Gamma}}_{t+k} = \hat{\mathbf{\Gamma}}_t. \tag{2.52}$$

Nous dénotons cet estimateur par $\hat{\mathbf{\Gamma}}_{t+k|t}$. L'horizon de prévision que nous utilisons le plus fréquemment est $k = 1$.

Des modèles plus complexes permettent de raffiner l'approximation localement constante, utilisant, par exemple, un modèle localement linéaire ou quadratique (Brown 1962; Gourieroux et Monfort 1997). Nous n'avons pas fait usage de ces modèles plus complexes par souci de simplicité, en grande partie parce qu'ils s'avèrent relativement peu utilisés par les praticiens de la VàR (RiskMetrics 1996).

2.6.3 Variance historique simple

Considérant pour le moment un seul actif, l'estimateur le plus simple de la volatilité de cet actif est la *variance historique simple* sur une fenêtre de M périodes :

$$\hat{\sigma}^{2\ (\mathsf{S})}_{t,M} = \frac{1}{M-1} \sum_{j=0}^{M-1} (r_{t-j} - \bar{r}_{t,M})^2, \tag{2.53}$$

où r_t est le rendement de l'actif au temps t et $\bar{r}_t$ est une moyenne mobile des rendements sur les M dernières périodes :[4]

$$\bar{r}_{t,M} = \frac{1}{M} \sum_{j=0}^{M-1} r_{t-j}. \tag{2.54}$$

La théorie statistique classique montre que $\hat{\sigma}^{2\ (\mathsf{S})}_{t,M}$ est sans biais pour des

[4]Nous pouvons aussi prendre la moyenne historique sur toute la séquence, ou encore, pour des actifs boursiers mesurés sur de courtes échéances, fixer $\bar{r}_t = 0$ (Figlewski 1994). Généralement, nous préférons considérer des prédicteurs qui sont calculables de manière causale au temps t à partir de l'information disponible au temps t.

rendements i.i.d.. Il est de plus consistant pour $M \to \infty$ (et une séquence d'observations de longueur infinie).

2.6.4 Variance historique pondérée exponentiellement

Bien que l'estimateur (2.53) de la volatilité soit simple à calculer, il se trouve affligé du défaut de réagir relativement lentement aux variations subites de $\mathbf{\Gamma}_t$. Or, grand nombre de résultats concluent à l'évidence marquée d'*hétéroscédasticité* dans les rendements d'actifs boursiers (Campbell, Lo et MacKinlay 1997). De plus, cet estimateur conduit à des changements brusques de l'estimateur de la variance d'une période à la suivante lorsqu'une observation extrême sort de la fenêtre de calcul, et ce sans que les conditions du marché affectant la volatilité des actifs ne se soient significativement modifiées au cours de cette période.

Pour remédier à ces problèmes, nous présentons la *variance pondérée exponentiellement* sur une fenêtre de M périodes (Cox 1961; Gourieroux et Monfort 1997) :

$$\hat{\sigma}_t^{2\ (\mathsf{E_M})} = \frac{1}{S_M} \sum_{j=0}^{M-1} \lambda^j (r_{t-j} - \bar{r}_{t-j,M})^2, \tag{2.55}$$

avec $S_M = \sum_{j=0}^{M-1} \lambda^j$, et $\bar{r}_t$ défini à l'éq. (2.54). Le facteur λ—appellé *facteur d'oubli*—est une constante qui gouverne la vitesse relative avec laquelle les nouvelles observations sont « absorbées » par l'estimateur. Nous supposons $0 < \lambda < 1$. Nous verrons plus loin les méthodes permettant de choisir ce λ.

Dans la limite où $M \to \infty$, $\hat{\sigma}_t^{2\ (\mathsf{E_M})}$ s'exprime sous une forme récursive

particulièrement simple :

$$\begin{aligned}
\lim_{M\to\infty} \hat{\sigma}_t^{2\ (\mathsf{E_M})} &= \lim_{M\to\infty} \frac{1}{S_M} \sum_{j=0}^{M-1} \lambda^j (r_{t-j} - \bar{r}_{t-j,M})^2 \\
&= (1-\lambda)(\lambda^0 (r_t - \bar{r}_{t,M})^2) + (1-\lambda) \sum_{j=1}^{\infty} \lambda^j (r_{t-j} - \bar{r}_{t-j,M})^2 \\
&= (1-\lambda)(r_t - \bar{r}_{t,M})^2 + \lambda(1-\lambda) \sum_{j=0}^{\infty} \lambda^j (r_{t-1-j} - \bar{r}_{t-1-j,M})^2 \\
&= (1-\lambda)(r_t - \bar{r}_{t,M})^2 + \lambda \lim_{M\to\infty} \hat{\sigma}_{t-1}^{2\ (\mathsf{E_M})}. \qquad (2.56)
\end{aligned}$$

Nous utilisons le fait que $\sum_{j=0}^{\infty} \lambda^j = 1/(1-\lambda)$ pour $0 < \lambda < 1$.

Nous dénotons l'estimateur $\lim_{M\to\infty} \hat{\sigma}_t^{2\ (\mathsf{E_M})}$ par $\hat{\sigma}_t^{2\ (\mathsf{E})}$. Dans toutes nos expériences avec la VàR, nous faisons appel à cet estimateur pour calculer la volatilité.

Fin de la récurrence

L'éq. (2.56) laisse sans définition l'estimateur initial de la variance, $\hat{\sigma}_1^{2\ (\mathsf{E})}$. Cet estimateur peut être choisi de plusieurs façons (Brown 1962) :

- Si on dispose de donnés antérieures au début de la série, on peut les utiliser pour calculer une variance initiale à l'aide de la variance historique simple (2.53).
- Autrement, il faut choisir une valeur initiale plausible à l'aide de connaissances à priori.

Dans toutes nos expériences, nous avons fait un compromis (qui se trouve à fonctionner très bien en pratique) en choisissant $\hat{\sigma}_1^{2\ (\mathsf{E})} = r_1^2$.

Comparaison entre l'estimateur simple et exponentiel

La figure 2.3 compare les estimateurs de volatilité $\hat{\sigma}_{t,60}^{2\ (\mathsf{S})}$ et $\hat{\sigma}_t^{2\ (\mathsf{E})}$. Nous observons que $\hat{\sigma}_t^{2\ (\mathsf{E})}$ répond plus promptement aux périodes de volatilité accrues ; ceci est particulièrement frappant pour le début des années 1980, et le crash de 1987. De plus, nous remarquons un artefact de la fenêtre finie

de $\hat{\sigma}^{2\ (\mathsf{S})}_{t,60}$, qui enregistre une baisse abrupte de la volatilité en octobre 1992, exactement 5 ans après le crash de 1987 (la fenêtre utilisée est de 60 mois). Ce changement brusque n'est motivé par aucune transformation notable de la série des rendements autour de cette période. L'estimateur exponentiel, quant à lui, se comporte de manière beaucoup plus progressive.

Notons que, étant donné que nous utiliserons toujours $\hat{\sigma}^{2\ (\mathsf{E})}_t$ comme estimateur de la volatilité dans les sections suivantes, nous allégeons la notation en nous y référant simplement par $\hat{\sigma}^2_t$. Lorsque nous voulons marquer la dépendance claire de l'estimateur par rapport à un facteur d'oubli λ particulier, nous le dénotons par $\hat{\sigma}^2_t(\lambda)$.

2.6.5 Le cas pour plusieurs actifs

L'estimateur de la matrice de variance–covariance pour plusieurs actifs est une généralisation simple de l'éq. (2.56) :

$$\hat{\mathbf{\Gamma}}_t = \lambda \hat{\mathbf{\Gamma}}_{t-1} + (1-\lambda)(\mathbf{r}_t\ \mathbf{r}'_t), \tag{2.57}$$

où $\mathbf{r}_t$ est le vecteur des rendements des actifs au temps t, et nous avons considéré l'espérance des rendements égale à zéro pour plus de simplicité.

2.6.6 Comment choisir le facteur d'oubli ?

Le facteur d'oubli doit généralement être choisi de façon à produire un estimateur de la variance qui ait la meilleure performance de généralisation (espérée) à travers tous les actifs. L'une des manières de mesurer cette performance est de calculer l'erreur quadratique de l'estimateur par rapport aux rendements carrés $(r_{it} - \bar{r_i})^2$ de l'actif i, pour tout t :

$$\lambda^* = \arg\min_{\lambda} \mathrm{MSE}_i(\lambda) \tag{2.58}$$

$$\mathrm{MSE}_i(\lambda) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{MSE}_{i,t}(\lambda) \tag{2.59}$$

$$\mathrm{MSE}_{i,t}(\lambda) = \mathrm{E}\left[\left(\hat{\sigma}^2_{t|t-1}(\lambda) - (R_{it} - \mathrm{E}[R_{it} \mid \mathcal{I}_{t-1}])^2 \right)^2 \middle| \mathcal{I}_{t-1} \right]. \tag{2.60}$$

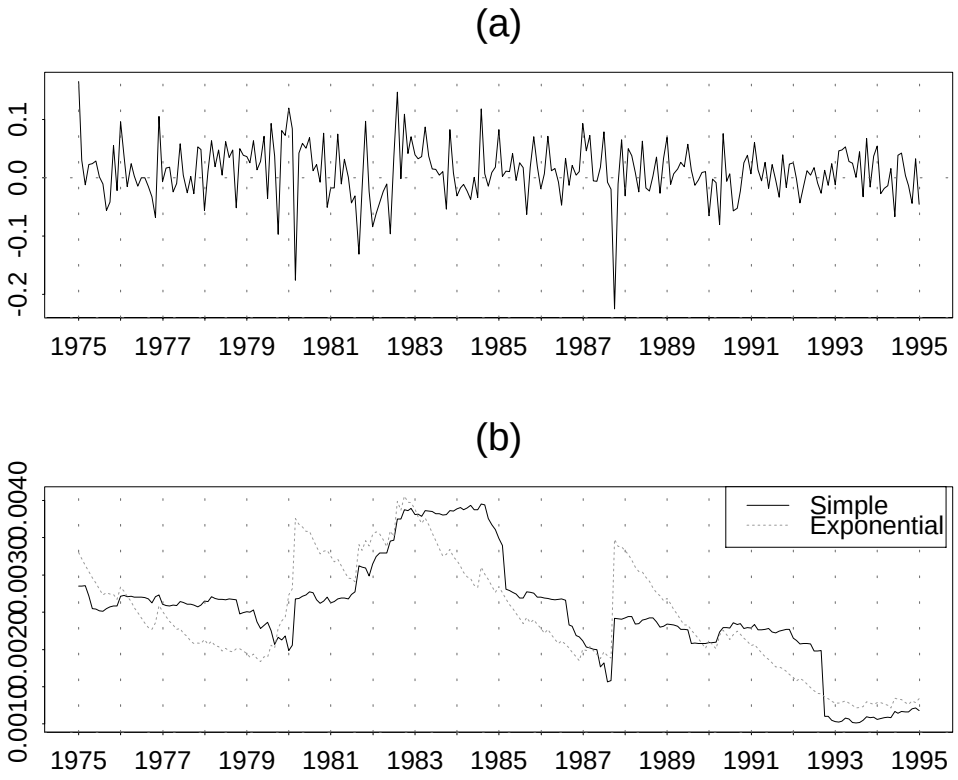


Figure 2.3: *Comparaison entre l'estimateur de variance simple et exponentiel. La partie (a) illustre les rendements mensuels obtenus par le TSE 300 entre 1975 et 1995. La partie (b) compare deux estimateurs de variance découlant de ces rendements. L'estimateur simple* $\hat{\sigma}^{2\ (\mathsf{S})}_{t,60}$ *utilise une fenêtre de 60 observations (5 ans). L'estimateur exponentiel* $\hat{\sigma}^{2\ (\mathsf{E})}_{t}$ *utilise un facteur d'oubli de 0.97.*

Un estimateur de MSE_i est l'erreur quadratique moyenne hors-échantillon :

$$\widehat{\text{MSE}}_i(\lambda) = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{\sigma}^2_{t|t-1}(\lambda) - (r_{it} - \bar{r_i})^2\right)^2. \tag{2.61}$$

Cet estimateur souffre d'un léger biais si r_{it} fait partie du calcul de $\bar{r_i}$, pour certains t ; cependant, nous l'ignorons car il est généralement minuscule par rapport à la variance de l'estimateur.

Nous avons appliqué cet estimateur aux rendements boursiers mensuels de l'indice TSE 300 et obtenons

$$\hat{\lambda}^* = 0.97.$$

La figure 2.4 illustre le comportement de $\widehat{\text{MSE}}_{\text{TSE}}$ pour différents facteurs d'oubli. Dans la suite nous utilisons le même facteur d'oubli $\lambda = 0.97$ dans tous nos calculs de la VàR.

Nos résultats confirment ceux publiés par le groupe RiskMetrics (1996), qui utilise la même procédure pour choisir le facteur d'oubli. Bien que nous n'avons estimé $\hat{\lambda}^*$ que sur la seule série du TSE 300, les résultats de RiskMetrics indiquent que le facteur d'oubli $\lambda = 0.97$ est optimal pour presque tous les marchés boursiers occidentaux. De plus, Gourieroux et Monfort (1997) présentent une une analyse détaillée du comportement de la moyenne exponentielle qui suggère que, pour la plupart des séries rencontrées dans la pratique, l'erreur de prévision est relativement insensible au choix précis du facteur d'oubli, sur une plage étendue de ce dernier.

Il est à noter que la présente procédure d'estimation de $\hat{\lambda}^*$ n'est pas sans défauts. Premièrement, nous ne choisissons qu'un seul facteur d'oubli pour toutes les séquences, alors que différents types d'actifs pourraient bénéficier de facteurs qui leur sont propres.[5] Deuxièmement, notre critère d'erreur ne tient compte que de la prévision de la variance, sans tenir compte de la qualité des estimateurs de covariances entre les actifs. Compte tenu du coût prohibitif

[5]RiskMetrics préconise l'emploi de deux familles très différentes de facteurs, l'une pour les rendements mensuels, et l'autre pour les rendements quotidiens. Pour la majorité des rendements mensuels, le facteur recommandé est 0.97, celui que nous avons utilisé ; cependant, pour les rendements quotidiens, le facteur suggéré est 0.94.

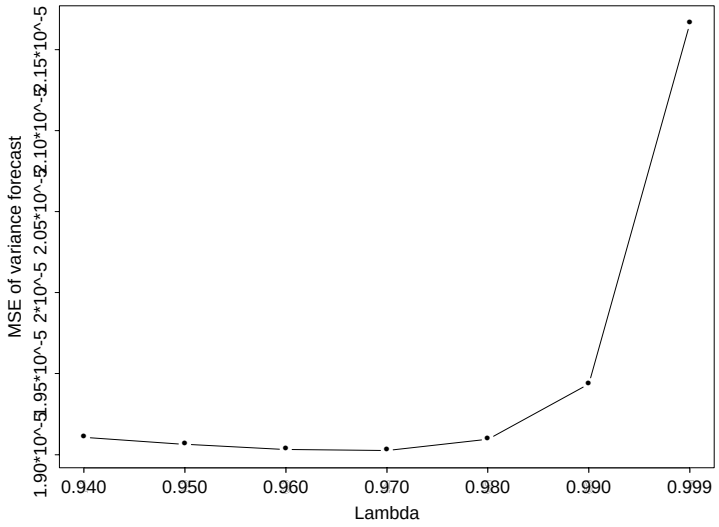


Figure 2.4: *Erreur quadratique moyenne (MSE) de l'estimateur de variance exponentiel des rendements mensuels du TSE 300, en fonction du facteur d'oubli λ.*

d'estimer une *matrice* de facteurs d'oubli (pour utiliser à l'éq. (2.57)) et du peu de données disponibles pour ce faire, nous n'avons pas tenu compte du problème de choisir les facteurs d'oubli appropriés pour les covariances.

Autres méthodes de sélection Il est aussi possible de sélectionner un facteur d'oubli optimal par d'autres principes, en particulier en choisissant le λ qui maximise la vraisemblance conditionnelle des rendements sous un modèle gaussien (Ahlburg 1992; Armstrong et Collopy 1992; Fildes 1992; Magdon-Ismail et Abu-Mostafa 1997). Cependant, cette méthode détermine le meilleur λ « in-sample » (i.e. on n'estime pas la performance de généralisation du modèle de variance), et impose de surcroît des suppositions paramétriques au modèle.

2.6.7 Autres modèles de volatilité

Plusieurs autres modèles de volatilité ont été proposés dans la littérature. Parmi les plus populaires sont les modèles à *hétéroscédasticité autorégressive*

de type ARCH (Engle 1982) et GARCH (Bollerslev 1986).

En supposant les rendements (de moyenne nulle) distribués selon

$$r_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_t^2),$$

un modèle GARCH(1,1)—le type le plus fréquemment utilisé dans la classe des modèles GARCH—pose la relation suivante pour la variance :

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha r_t^2. \tag{2.62}$$

Ce modèle est, clairement, une généralisation du modèle de variance utilisant la moyenne historique exponentielle de l'éq. (2.56). Le groupe RiskMetrics (1996) a publié des résultats démontrant la grande similarité des prévisions de la volatilité utilisant une moyenne historique exponentielle et un modèle GARCH(1,1). Étant donné la difficulté aiguë d'estimation des paramètres pour les modèles de la famille GARCH, nous n'avons pas ressenti la nécessité de remplacer l'estimateur pondéré exponentiellement par l'un d'eux.

2.6.8 À propos de la volatilité implicite

La *volatilité implicite* (VI) (Hull 1999) d'un actif est l'estimé de la volatité découlant du prix des options transigées sur cet actif. Elle est déterminée en « retournant » la direction d'un modèle de valorisation d'options comme celui de Black et Scholes (1973).

Cette volatilité implicite a l'avantage théorique d'incorporer de l'information relative aux attentes futures du marché à propos d'un actif, et non pas seulement les réalisations passées des rendements de l'actif.

Cependant, la VI souffre de plusieurs problèmes. Tout d'abord, elle dépend intimement du modèle de valorisation utilisé, et suppose que le marché utilise le même modèle pour fixer le prix des options. Des résultats récents (Figlewski 1997) suggèrent que la VI n'est pas un meilleur prédicteur de la volatilité future que ne l'est la variance historique. De plus, l'utilisation de la VI suppose qu'un marché d'options bien établi et liquide existe pour tous les actifs considérés dans un portefeuille. Finalement, la VI ne donne aucune

indication de la « *corrélation implicite* » qui existe entre les actifs ; les estimés des corrélations doivent toujours être obtenus séparément.

Pour toutes ces raisons, nous avons choisi d'ignorer la volatilité implicite comme prédicteur de la volatilité des actifs ; nous nous sommes limités à l'utilisation d'une moyenne historique pondérée exponentiellement.

Chapitre 3

Systèmes adaptatifs pour la gestion de portefeuille

Les systèmes adaptatifs peuvent s'incorporer de plusieurs façons à l'intérieur du cadre de gestion de portefeuille utilisant la VàR présenté au chapitre précédent. Dans ce mémoire, nous étudions et comparons deux paradigmes distincts, l'un consistant à prendre une décision basée sur une *prévision des rendements futurs*, l'autre consistant à sauter l'étape de la prévision et à *prendre directement une décision* basée sur des variables explicatives élémentaires.

Nous commençons par présenter les fondements des systèmes adaptatifs basés sur les réseaux de neurones. Nous poursuivons par une comparaison des deux paradigmes de prise de décision, et dérivons finalement leurs propriétés théoriques respectives.

3.1 Un bref survol des réseaux de neurones

Cette section passe en revue les résultats principaux sur la théorie des réseaux de neurones nécessaires à la compréhension de ce mémoire. Elle ne se

veut pas une exposition approfondie du domaine, pour laquelle nous référons le lecteur aux excellents ouvrages parus sur le sujet (Bishop 1995).

3.1.1 Pourquoi ?

Les réseaux de neurones, et plus particulièrement les *perceptrons multi-couches* dont nous traitons ici, ont reçu une attention particulière pour les applications financières dans les années récentes, grâce à leurs nombreux avantages par rapport aux méthodes de modélisation plus conventionnelles :

- Ils sont **flexibles**, en pouvant s'attaquer à grand nombre de problèmes classiques en statistique : *classification*, *régression*, et *estimation de densité*. (Dans ce mémoire, nous utilisons les réseaux de neurones dans un cadre de régression).
- Ils sont **faciles à utiliser** et généralement faciles à entraîner, n'ayant besoin que d'exemples de la tâche à accomplir pour générer un modèle ; dans plusieurs situations, ils peuvent amoindrir le besoin d'une fine expertise (humaine) du problème à modéliser, la remplaçant par des exemples d'entraînement.
- Les perceptrons multi-couches sont, sous certaines conditions, des **approximateurs universels** (Hornik, Stinchcombe et White 1989) : ils peuvent représenter toute fonction continue avec une précision arbitrairement grande. D'un point de vue pratique, ils sont capables de modéliser des relations **non-linéaires** entre des variables, ce qui les rend attrayants par rapport aux modèles de régression linéaire pour un grand nombre de problèmes.

Les applications des réseaux de neurones à la finance sont nombreuses. Ils sont utilisés pour la prévision de séries économiques, où ils devancent généralement les modèles macroéconomiques et les modèles linéaires de séries chronologiques (Moody 1998; Neuneier et Zimmermann 1998). Ils sont aussi employés comme outil général de régression non-linéaire (Campbell, Lo et MacKinlay 1997).

3.1.2 Topologie d'un perceptron multi-couches

Le type particulier de réseaux de neurones que nous retenons pour ce mémoire est le *perceptron multi-couches* (nous utilisons l'acronyme anglais « MLP » pour les dénoter). Le MLP s'utilise pour représenter une fonction (vectorielle) $f : \mathbb{R}^M \mapsto \mathbb{R}^N$. Il possède les caractéristiques suivantes :

- Au cours de l'évaluation d'un fonction, l'information est traitée de façon purement *séquentielle* par le réseau. Le vecteur d'entrée ($\in \mathbb{R}^M$) présenté au réseau[1] est progressivement transformé par une ou plusieurs *couches cachées*, et finalement émerge comme vecteur de résultat ($\in \mathbb{R}^N$) à la *couche de sortie*. La notion de couche est décrite à la section suivante.
- La nature séquentielle du réseau préclut les boucles de rétroaction à l'intérieur du réseau (« feedback »).

Composition des couches

Dans un MLP, une couche représente une fonction vectorielle « simple ». La couche est l'unité fondamentale de traitement du MLP ; la fonction représentée par ce dernier est la composition de la fonction de chacune de ses couches. Soit un MLP $\mathcal{M}$ constitué de κ couches, chacune calculant une fonction $f_i : \mathbb{R}^{H_{i-1}} \mapsto \mathbb{R}^{H_i}$ (la signification de H_i est donnée plus bas). La fonction résultante calculée par le MLP est, par définition,

$$f_{\mathcal{M}} = f_\kappa \circ f_{\kappa-1} \circ \cdots \circ f_1. \tag{3.1}$$

Dans cette équation, f_κ joue le rôle de couche de sortie (f_1 ne *constitue pas* la couche d'entrée ; il s'agit de la première couche cachée). Pour que $f_{\mathcal{M}}$ applique de $\mathbb{R}^M$ vers $\mathbb{R}^N$, nous supposons que f_1 a pour domaine $\mathbb{R}^M$ et que f_κ a pour image $\mathbb{R}^N$.

La dimensionalité des fonctions intermédiaires demeure non spécifiée ; ainsi que nous l'expliquons plus loin, ce sont les paramètres libres les plus importants dans l'application pratique des réseaux de neurones. Nous dénotons la

[1] Certains auteurs font procéder ce vecteur d'une « couche d'entrée », même si cette dernière n'est pas une couche au sens où nous le définissons plus bas.

dimensionalité de *sortie* de la couche i (l'image de f_i) par H_i ; de même, la dimensionalité d'*entrée* de la couche $i+1$ est aussi H_i. Pour simplifier la notation, nous considérons $H_0 = M$.

Types de couches

La fonction f_i calculée par une couche i doit posséder une dérivée première en tout point.[2] Elle est généralement de deux types :

Linéaire Étant donné un vecteur d'entrées $\mathbf{x} \in \mathbb{R}^{H_{i-1}}$, ce type de couche effectue une simple application linéaire :

$$f_i(\mathbf{x}; \mathbf{A}_i, \mathbf{b}_i) = \mathbf{A}_i\mathbf{x} + \mathbf{b}_i \,, \tag{3.2}$$

où $\mathbf{A}_i$ et $\mathbf{b}_i$ sont les *paramètres* de la fonction ; $\mathbf{A}_i$ est la matrice $H_i \times H_{i-1}$ correspondant à l'application linéaire et $\mathbf{b}_i$ est un vecteur de biais. Cette matrice et ce vecteur demeurent constants pendant une évaluation de la fonction (ils sont modifiés pendant la phase d'apprentissage, tel que décrit ci-bas).

Tanh (non-linéaire) Étant donné un vecteur d'entrées $\mathbf{x}$, ce type de couche effectue d'abord une application linéaire, suivie d'une application de la fonction non-linéaire $\tanh(\cdot)$ (nous supposons que cette fonction s'applique élément par élément à son vecteur d'entrées) :[3]

$$f_i(\mathbf{x}; \mathbf{A}_i, \mathbf{b}_i) = \tanh(\mathbf{A}_i\mathbf{x} + \mathbf{b}_i), \tag{3.3}$$

[2] Il est possible de travailler avec des fonctions qui ne sont pas différentiables en certains points, mais nous ne nous préoccupons pas de ces complications et supposons la fonction différentiable partout.

[3] Certains auteurs préfèrent la fonction sigmoide $s(x) \stackrel{\text{def}}{=} 1/(1+\exp(-x))$ à la fonction $\tanh(\cdot)$. L'effet théorique du choix de l'une ou l'autre de ces fonctions sur la capacité d'approximation universelle du réseau est le même. Cependant, nous constatons empiriquement la convergence plus rapide du réseau au moment de l'apprentissage lorsque ce dernier utilise des fonctions $\tanh(\cdot)$; ce phénomène est expliqué par LeCun, Bottou, Orr et Müller (1998).

où comme dans le cas de la couche linéaire, $\mathbf{A}_i$ est une matrice $H_i \times H_{i-1}$ et $\mathbf{b}_i$ est un vecteur de biais, lesquels demeurent constants pendant l'évaluation de la fonction.

Topologie typique d'un MLP

Un MLP « intéressant » (i.e. un permettant d'approximer des fonctions non-linéaires) doit posséder au moins une couche Tanh. (Un réseau ne possédant que des couches linéaires n'effectue rien de plus qu'une multiplication de matrice). La topologie de MLP la plus utilisée en pratique (et celle que nous utilisons dans nos expériences), fixe $\kappa = 2$, ce qui produit la topologie :

Couche cachée f_1 Couche non-linéaire avec fonction Tanh, telle que décrite plus haut. Nous nous référons parfois à la dimension de sortie de cette couche, H_1, par « nombre d'unitées cachées » (ce qui ne constitue pas une ambiguité, car il n'y a qu'une seule couche cachée).

Couche de sortie f_2 Couche linéaire.

Cette topologie produit un réseau qui calcule la fonction suivante, étant donné un vecteur $\mathbf{x}$:

$$\begin{aligned} f_{\mathcal{M}}(\mathbf{x}) &= (f_2 \circ f_1)(\mathbf{x}) \\ &= \mathbf{A}_2 \tanh(\mathbf{A}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2. \end{aligned} \tag{3.4}$$

Les matrices $\mathbf{A}_1, \mathbf{A}_2$ et vecteurs $\mathbf{b}_1, \mathbf{b}_2$ constituent les paramètres ajustables du réseau. Ce sont ces paramètres que le processus d'entraînement modifie en vue de produire un réseau qui approxime « le mieux possible » une fonction cible.

Vecteur de paramètres

Par simplicité de notation, nous dénotons par $\boldsymbol{\theta}^{(i)}$ le vecteur des paramètres ajustables de la couche i d'un réseau. Dans le cas de la première couche du MLP décrit ci-haut, $\boldsymbol{\theta}^{(1)}$ résulte de la concaténation (dans un ordre arbitraire mais fixé) de $\mathbf{A}_1$ et $\mathbf{b}_1$. Il en va de même pour $\boldsymbol{\theta}^{(2)}$. Nous dénotons la fonction calculée par la couche i selon des paramètres $\boldsymbol{\theta}^{(i)}$ par $f_i(\cdot; \boldsymbol{\theta}^{(i)})$.

De plus, nous dénotons par $\boldsymbol{\theta}$ le vecteur résultant de la concaténation de tous les $\boldsymbol{\theta}^{(i)}$; ce vecteur représente l'ensemble de tous les paramétres ajustables du réseau. Nous dénotons la fonction calculée par un MLP (de topologie préétablie) selon des paramètres $\boldsymbol{\theta}$ par $f(\cdot;\boldsymbol{\theta})$.

3.1.3 Entraînement d'un MLP

Le processus d'entraînement vise à ajuster les paramètres $\boldsymbol{\theta}$ du MLP de façon à minimiser une fonction coût donnée. Dénotons par $C(\boldsymbol{\theta})$ la perte associée au réseau $f(\cdot;\boldsymbol{\theta})$. Cette perte peut être, par exemple, l'écart quadratique espéré entre $\mathbf{y}$ et $f(\mathbf{x};\boldsymbol{\theta})$, pour $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^M \times \mathbb{R}^N$ tirés d'une distribution (jointe) fixée (mais pas nécessairement connue) :

$$C(\boldsymbol{\theta}) = \mathrm{E}_{\mathbf{X},\mathbf{Y}}[(f(\mathbf{X};\boldsymbol{\theta}) - \mathbf{Y})^2 \mid \boldsymbol{\theta}]. \tag{3.5}$$

En pratique nous ne disposons que d'un nombre fini d'*exemples d'entraînement* tirés de cette distribution, $\mathcal{D} = \{(\mathbf{x}_\ell, \mathbf{y}_\ell)\}_{\ell=1}^{L}$ (nous supposons les exemples tirés de façon indépendante). Nous pouvons estimer la perte quadratique par rapport à cet ensemble, pour un $\boldsymbol{\theta}$ quelconque, par :[4]

$$\hat{C}(\boldsymbol{\theta}) = \frac{1}{L}\sum_{\ell=1}^{L} Q(\mathbf{x}_\ell, \mathbf{y}_\ell; \boldsymbol{\theta}) \tag{3.6}$$

$$Q(\mathbf{x},\mathbf{y};\boldsymbol{\theta}) = \tfrac{1}{2}\left\|f(\mathbf{x};\boldsymbol{\theta}) - \mathbf{y}\right\|^2, \tag{3.7}$$

où $\|\cdot\|^2$ est la distance euclidienne. La dépendance de $\hat{C}(\boldsymbol{\theta})$ par rapport à l'ensemble $\mathcal{D}$ est généralement laissée implicite par souci de simplifier la notation. Lorsque nous souhaitons spécifier un ensemble explicite, nous le mentionnons en indice, par exemple $\hat{C}_{\mathcal{D}}(\boldsymbol{\theta})$.

Nous effectuons des modifications à cette fonction de coût selon les contextes, que nous précisons en temps opportuns ; cependant, l'esprit général de ces

[4]Cet estimateur est sans biais pour $\boldsymbol{\theta}$ quelconque ; cependant, il est biaisé pour le $\hat{\boldsymbol{\theta}}^*$ « optimal » décrit plus bas. Nous traitons de la délicate question de la *généralisation* à la fin de la présente section.

fonctions demeure le même et nous conservons la perte quadratique ci-haut à titre indicatif.

Objectif de l'entraînement

L'entraînement cherche à trouver le $\boldsymbol{\theta}^*$ qui minimise la fonction de coût :

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} C(\boldsymbol{\theta}). \tag{3.8}$$

Puisque ce but est inaccessible en pratique, nous effectuons la minimisation par rapport à l'estimateur de la fonction de coût :

$$\hat{\boldsymbol{\theta}}^* = \arg\min_{\boldsymbol{\theta}} \hat{C}(\boldsymbol{\theta}). \tag{3.9}$$

Notons que ce problème d'optimisation n'est pas affligé des complications résultant de l'imposition de contraintes sur les paramètres : l'une des caractéristiques des MLP nous utilisons est qu'ils admettent des paramètres $\boldsymbol{\theta}_i \in \mathbb{R}$.

La solution de ce problème peut s'effectuer en deux étapes : (i) d'une part, nous devons calculer d'une façon efficace le gradient de la fonction de coût par rapport à chacun des paramètres $\boldsymbol{\theta}$ du réseau ; ce calcul est possible grâce à l'algorithme de rétropropagation présenté ci-après ; (ii) d'autre part, nous utilisons ce gradient pour guider une recherche itérative dans l'espace des paramètres, dont nous discutons ensuite.

Rétropropagation du gradient

L'algorithme de rétropropagation permet de calculer efficacement et avec précision le gradient de $\hat{C}$ par rapport aux paramètres $\boldsymbol{\theta}$ du réseau, étant donnée une valeur courante $\mathbf{t}$ des paramètres :

$$\left.\frac{\partial \hat{C}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\mathbf{t}}. \tag{3.10}$$

Remarque sur la notation Dans les calculs de rétropropagation, il est important de ne pas confondre les *variables* (par rapport auxquelles nous pouvons dériver) et les points (fixés) auxquels sont évalués les fonctions et les gradients. C'est pourquoi dans ce qui suit, nous dénotons les variables par des lettres grecques ($\boldsymbol{\xi}, \boldsymbol{\theta}$), et les points d'évaluation par des lettres romaines ($\mathbf{x}, \mathbf{t}$).

De plus, lorsque le contexte est clair, nous abrégeons (par exemple)

$$\left.\frac{\partial \hat{C}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\mathbf{t}} \qquad \text{par} \qquad \left.\frac{\partial \hat{C}}{\partial \boldsymbol{\theta}}\right|_{\mathbf{t}}.$$

Gradient total L'éq. (3.6) montre que le gradient complet de $\hat{C}$ est constitué de la somme du gradient de Q pour chaque exemple d'entraînement :

$$\left.\frac{\partial \hat{C}}{\partial \boldsymbol{\theta}}\right|_{\mathbf{t}} = \frac{1}{L}\sum_{\ell=1}^{L} \left.\frac{\partial Q}{\partial \boldsymbol{\theta}}\right|_{\mathbf{x}_\ell,\mathbf{t}}. \tag{3.11}$$

(Formellement, $\frac{\partial Q}{\partial \boldsymbol{\theta}}$ devrait être conditionné sur $\mathbf{y}_\ell$, mais puisque nous ne dérivons jamais par rapport à cette variable et qu'elle n'interagit pas autrement avec les autres, nous omettons de la mentionner pour alléger une notation déjà chargée).

Nous notons de plus que le gradient de Q par rapport au j-ième élément de $\boldsymbol{\theta}$ est :

$$\begin{aligned}\left.\frac{\partial Q}{\partial \boldsymbol{\theta}_j}\right|_{\mathbf{x},\mathbf{t}} &= \sum_{k=1}^{N} \left[\left.\frac{\partial Q}{\partial f(\boldsymbol{\xi};\boldsymbol{\theta})}\right|_{\mathbf{x},\mathbf{t}}\right]_k \left[\left.\frac{\partial f(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right|_{\mathbf{x},\mathbf{t}}\right]_k \\ &= \sum_{k=1}^{N} (f(\mathbf{x};\mathbf{t})_k - (\mathbf{y})_k) \left[\left.\frac{\partial f(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right|_{\mathbf{x},\mathbf{t}}\right]_k. \end{aligned} \tag{3.12}$$

Les sections suivantes se concentrent donc sur le calcul du problématique gradient $\frac{\partial Q}{\partial \boldsymbol{\theta}}$ en un point $\mathbf{x}$ et pour des paramètres $\mathbf{t}$ donnés.

Passe avant La figure 3.1 illustre le calcul (à travers les deux premières couches) de la fonction représentée par le MLP, et la contribution de ce calcul

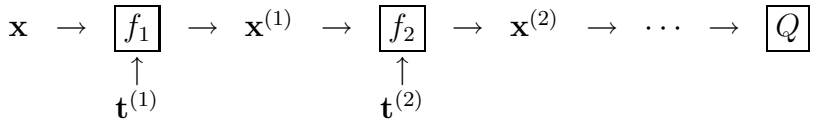


Figure 3.1: *Calcul de la « passe avant » dans un MLP.*

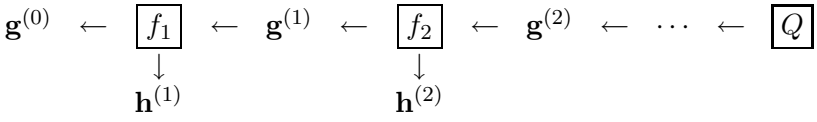


Figure 3.2: *Calcul de la « passe arrière » dans un MLP, qui produit le gradient de $\hat{C}$ par rapport à chacun des paramètres.*

à la fonction de coût $\hat{C}$, pour une entrée $\mathbf{x}$ donnée. Le résultat du calcul de $f_i(\mathbf{x}^{(i-1)}; \boldsymbol{\theta}^{(i)})$ (la fonction calculée par la couche i) est noté $\mathbf{x}^{(i)}$. Nous posons $\mathbf{x}^{(0)} = \mathbf{x}$, l'entrée du réseau. Nous appliquons la même convention aux $\boldsymbol{\xi}$ ($\boldsymbol{\xi}^{(1)}$, etc.) qui sont les variables correspondantes.

Pour des raisons évidentes, ce calcul est appelé « passe avant ».

Passe arrière La passe arrière constitue le coeur de la rétropropagation. C'est un algorithme récursif traversant le graphe du MLP dans le sens inverse des flèches et qui, pour la couche i, calcule

$$\left.\frac{\partial Q}{\partial \boldsymbol{\theta}^{(i)}}\right|_{\mathbf{x}^{(i-1)},\mathbf{t}} \qquad \text{et} \qquad \left.\frac{\partial Q}{\partial \boldsymbol{\xi}^{(i-1)}}\right|_{\mathbf{x}^{(i-1)},\mathbf{t}}$$

à partir des valeurs courantes $\mathbf{t}$ des paramètres, du résultat intermédiaire $\mathbf{x}^{(i-1)}$ calculé par la passe avant, ainsi que du résultat de $\left.\frac{\partial Q}{\partial \boldsymbol{\xi}^{(i)}}\right|_{\mathbf{x}^{(i)},\mathbf{t}}$ (calculé auparavant pour la couche $i+1$).

Définissons $\mathbf{g}^{(i)}$ et $\mathbf{h}^{(i)}$ comme étant le gradient de Q par rapport à, respectivement, $\boldsymbol{\xi}^{(i)}$ et à $\boldsymbol{\theta}^{(i)}$, tous deux évalués au point $\mathbf{x}^{(i)}$ calculé par la passe avant, sous la valeur courante $\mathbf{t}$ du vecteur de paramètres :

$$\mathbf{g}^{(i)} \stackrel{\text{def}}{=} \left.\frac{\partial Q}{\partial \boldsymbol{\xi}^{(i)}}\right|_{\mathbf{x}^{(i)},\mathbf{t}} \qquad \mathbf{h}^{(i)} \stackrel{\text{def}}{=} \left.\frac{\partial Q}{\partial \boldsymbol{\theta}^{(i)}}\right|_{\mathbf{x}^{(i)},\mathbf{t}}. \tag{3.13}$$

La figure 3.2 illustre ces définitions.

Nous rétropropageons à travers la couche i comme suit. Appliquant les règles élémentaires de dérivation en chaîne, nous trouvons, pour le j-ième élément du gradient par rapport à $\boldsymbol{\xi}^{(i-1)}$

$$\frac{\partial Q}{\partial \boldsymbol{\xi}_j^{(i-1)}} = \sum_k \frac{\partial Q}{\partial \boldsymbol{\xi}_k^{(i)}} \frac{\partial \boldsymbol{\xi}_k^{(i)}}{\partial \boldsymbol{\xi}_j^{(i-1)}}, \tag{3.14}$$

où la sommation est effectuée sur tous les éléments du vecteur $\boldsymbol{\xi}^{(i)}$. Évaluant ces dérivées aux points appropriés, nous trouvons pour le j-ième élément de $\mathbf{g}^{(i-1)}$:

$$\mathbf{g}_j^{(i-1)} = \sum_k \mathbf{g}_k^{(i)} \left[\frac{\partial \boldsymbol{\xi}_k^{(i)}}{\partial \boldsymbol{\xi}_j^{(i-1)}} \right]_{\mathbf{x}^{(i-1)}, \mathbf{t}} . \tag{3.15}$$

Nous observons que $\frac{\partial \boldsymbol{\xi}_k^{(i)}}{\partial \boldsymbol{\xi}_j^{(i-1)}}$ n'est rien d'autre que le k-ième élément de la dérivée partielle de f_i par rapport à une de ses variables d'entrée $\boldsymbol{\xi}_j^{(i-1)}$ (le j,k-ième élément du jacobien de la fonction). Nous précisons le calcul de cette dérivée, spécifique à chaque type de couche, un peu plus bas.

De même, le gradient par rapport au j-ième paramètre de la couche i est

$$\frac{\partial Q}{\partial \boldsymbol{\theta}_j^{(i)}} = \sum_k \frac{\partial Q}{\partial \boldsymbol{\xi}_k^{(i)}} \frac{\partial \boldsymbol{\xi}_k^{(i)}}{\partial \boldsymbol{\theta}_j^{(i)}}, \tag{3.16}$$

où, comme plus haut, la sommation est effectuée sur tous les éléments du vecteur $\boldsymbol{\xi}^{(i)}$. Évaluant aux points appropriés, nous trouvons donc pour le j-ième élément de $\mathbf{h}^{(i)}$:

$$\mathbf{h}_j^{(i)} = \sum_k \mathbf{g}_k^{(i)} \left[\frac{\partial \boldsymbol{\xi}_k^{(i)}}{\partial \boldsymbol{\theta}_j^{(i)}} \right]_{\mathbf{x}^{(i-1)}, \mathbf{t}} . \tag{3.17}$$

Une fois complété le calcul du gradient par rapport aux paramètres d'une couche, $\mathbf{h}^{(i)}$, pour toutes les couches $i = \kappa, \ldots, 1$, nous concaténons simplement les vecteurs $\mathbf{h}^{(i)}$ pour obtenir le gradient par rapport à tous les paramètres du réseau, au point de la valeur courante $\mathbf{t}$ des paramètres. (La concaténation est

effectuée dans le même ordre que les $\mathbf{t}^{(i)}$ sont assemblés pour former $\mathbf{t}$).

Comment commencer la récurrence Les équations précédentes (3.15) et (3.17) expliquent comment rétropropager à travers la couche i étant donné les résultats pour la couche $i+1$; cependant, ils supposent l'existence du gradient de Q par rapport à la *sortie* du réseau (qui est la variable $\boldsymbol{\xi}^{(\kappa)}$ pour un réseau à κ couches). Celui-ci est fourni par l'éq. (3.12) :

$$\frac{\partial Q}{\partial \boldsymbol{\xi}^{(\kappa)}} \stackrel{\text{def}}{=} \frac{\partial Q}{\partial f(\boldsymbol{\xi}, \boldsymbol{\theta})} = f(\boldsymbol{\xi}; \theta) - \mathbf{y}. \tag{3.18}$$

Rétropropagation à travers une couche linéaire Il nous reste à calculer les dérivées $\frac{\partial \boldsymbol{\xi}_k^{(i)}}{\partial \boldsymbol{\xi}_j^{(i-1)}}$ et $\frac{\partial \boldsymbol{\xi}_k^{(i)}}{\partial \boldsymbol{\theta}_j^{(i)}}$ laissées en suspens précédemment. Soit la fonction calculée par une couche linéaire

$$\mathbf{z} = f(\mathbf{x}; \mathbf{A}, \mathbf{b}) = \mathbf{A}\mathbf{x} + \mathbf{b}$$

nous calculons la dérivée de la i-ième sortie $\mathbf{z}_i$ par rapport à la j-ième entrée $\mathbf{x}_j$ comme suit :

$$\frac{\partial \mathbf{z}_i}{\partial \mathbf{x}_j} = \mathbf{A}_{i,j}\,. \tag{3.19}$$

De même, les dérivées de la i-ième sortie par rapport aux paramètres $\mathbf{A}_{j,k}$ et $\mathbf{b}_j$ respectivement sont :

$$\frac{\partial \mathbf{z}_i}{\partial \mathbf{A}_{j,k}} = \begin{cases} \mathbf{x}_k & \text{si } i = j \\ 0 & \text{autrement} \end{cases} \tag{3.20}$$

$$\frac{\partial \mathbf{z}_i}{\partial \mathbf{b}_j} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{autrement} \end{cases} \tag{3.21}$$

Rétropropagation à travers une couche Tanh Soit la fonction calculée par une couche Tanh

$$\mathbf{z} = f(\mathbf{x}; \mathbf{A}, \mathbf{b}) = \tanh(\mathbf{A}\mathbf{x} + \mathbf{b})$$

nous calculons la dérivée de la i-ième sortie par rapport à la j-ième entrée $\mathbf{x}_j$ comme suit :

$$\frac{\partial \mathbf{z}_i}{\partial \mathbf{x}_j} = \mathbf{A}_{i,j} \tanh'(\mathbf{A}_i \mathbf{x} + \mathbf{b}_i)\,, \tag{3.22}$$

où $\mathbf{A}_i$ est le vecteur (rangée) correspondant à la i-ième rangée de $\mathbf{A}$.

De même, les dérivées de la i-ième sortie par rapport aux paramètres $\mathbf{A}_{j,k}$ et $\mathbf{b}_j$ respectivement sont :

$$\frac{\partial \mathbf{z}_i}{\partial \mathbf{A}_{j,k}} = \begin{cases} \mathbf{x}_k \tanh'(\mathbf{A}_i \mathbf{x} + \mathbf{b}_i) & \text{si } i = j \\ 0 & \text{autrement} \end{cases} \tag{3.23}$$

$$\frac{\partial \mathbf{z}_i}{\partial \mathbf{b}_j} = \begin{cases} \tanh'(\mathbf{A}_i \mathbf{x} + \mathbf{b}_i) & \text{si } i = j \\ 0 & \text{autrement} \end{cases} \tag{3.24}$$

On peut calculer facilement la dérivée de la fonction tanh en utilisant la relation :

$$\tanh'(x) = 1 - \tanh^2(x). \tag{3.25}$$

Optimisation des paramètres

Le gradient de la fonction de coût $\hat{C}$ par rapport aux paramètres du réseau $\boldsymbol{\theta}$, dont nous expliquons le calcul à la section précédente, peut ensuite être utilisé par un algorithme d'optimisation numérique non-linéaire pour résoudre le problème (3.9).

Il faut noter que la fonction à optimiser n'est pas convexe, donc nous ne pouvons espérer trouver un minimum global en temps raisonnable (car l'optimisation non-linéaire est, de façon générale, un problème NP-complet (Fletcher 1987)).

Les méthodes d'optimisation habituellement utilisées pour les réseaux de neurones sont les suivantes (Bishop 1995) :

Descente de gradient « batch » Cette méthode effectue une descente à partir d'un vecteur initial de paramètres $\mathbf{t}$ en utilisant la règle de mise-à-jour :

$$\mathbf{t}^{+} = \mathbf{t} - \eta \left. \frac{\partial \hat{C}}{\partial \boldsymbol{\theta}} \right|_{\mathbf{t}}$$

où η est le *pas de gradient*, qui doit être choisi avec soin, habituellement par essais-erreurs. Cette méthode a le désavantage d'être très lente et plutôt sensible aux minima locaux.

Descente de gradient stochastique Cette méthode effectue aussi une descente, mais réalise la mise-à-jour des paramètres en utilisant un estimé bruité du gradient qui est calculé pour chaque exemple d'entraînement $\mathbf{x}_\ell$:

$$\mathbf{t}^{+} = \mathbf{t} - \eta \left. \frac{\partial Q}{\partial \boldsymbol{\theta}} \right|_{\mathbf{x}_\ell, \mathbf{t}} .$$

D'un point de vue empirique, cette méthode converge plus rapidement que la précédente et est beaucoup plus robuste au problème des minima locaux.

Gradient conjugué Cette méthode en est aussi une de descente, mais choisit à chaque itération une direction conçue pour ne pas « détruire » le progrès effectué au cours des itérations précédentes (Fletcher 1987). Elle est très performante en pratique pour les réseaux de neurones, mais assez susceptible au problème des minima locaux.

Méthodes quasi–Newton Ces méthodes utilisent de l'information d'ordre supérieur (la matrice hessienne des dérivées secondes de C par rapport à $\boldsymbol{\theta}$) pour approximer localement la fonction de coût par une fonction quadratique, et trouver directement le minimum de cette approximation. Bien qu'elles soient classiques en optimisation non-linéaire, ces méthodes sont coûteuses dans le contexte des réseaux de neurones, car elles exigent le calcul du Hessien, dont la complexité croît selon le carré du nombre de paramètres libres dans le réseau. Bishop (1995) fournit un inventaire exhaustifs des « trucs » permet-

tant d'accélérer ce calcul, parfois au prix de la précision.

Méthode d'optimisation que nous retenons Nous avons utilisé une méthode de gradient conjugué, inspirée de celle présentée par PRESS, FLANNERY, TEUKOLSKY et VETTERLING (1992), et possédant les caractéristiques suivantes :

- Les paramètres initiaux sont choisis de façon aléatoire.
- Nous réoptimisons le réseau plusieurs fois à partir de points de départ différents ; nous retenons la meilleure solution de celles trouvées. Ceci permet de mitiger le problème des minima locaux.

Performance de généralisation

Lorsque l'ensemble $\mathcal{D}$ est utilisé pour trouver les paramètres $\hat{\boldsymbol{\theta}}^*$ qui minimisent la fonction $\hat{C}_{\mathcal{D}}(\boldsymbol{\theta})$, l'estimateur $\hat{C}_{\mathcal{D}}(\hat{\boldsymbol{\theta}}^*)$ en est un *biaisé* de la véritable perte $C(\hat{\boldsymbol{\theta}}^*)$ encourue sous les paramètres $\hat{\boldsymbol{\theta}}^*$ (VAPNIK 1998). Cette perte est souvent appelée « erreur de généralisation », et l'un objectifs des algorithmes d'apprentissage est de la minimiser.

Nous cherchons donc à trouver un estimateur non-biaisé de $C(\hat{\boldsymbol{\theta}}^*)$. Celui-ci est obtenu en tirant un autre ensemble $\mathcal{T}$ issu de la même distribution que $\mathcal{D}$ et indépendant de ce dernier. Sous ces conditions, l'estimateur $\hat{C}_{\mathcal{T}}(\hat{\boldsymbol{\theta}}^*)$ est non-biaisé (VAPNIK 1998).

Les ensembles $\mathcal{D}$ et $\mathcal{T}$ sont appelés, respectivement, *ensemble d'entraînement* et *ensemble de test*.

Nous faisons parfois appel à un troisième ensemble $\mathcal{V}$, l'*ensemble de validation*, indépendant des deux autres, pour estimer l'erreur de généralisation de façon répétée en cours d'entraînement ; cet ensemble peut servir, par exemple, à ajuster des paramètres gouvernant la topologie d'un réseau (comme le nombre d'unités cachées), qu'on appelle *hyper-paramètres*. Nous revenons au chapitre 4 sur ces ensembles ainsi que sur le choix des hyper-paramètres.

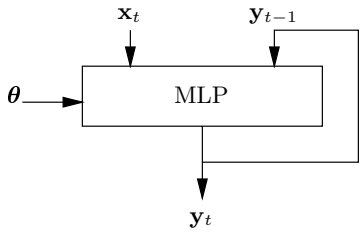


Figure 3.3: *Schéma d'un réseau récurrent, pour lequel les sorties du réseau au temps $t-1$ constituent le vecteur d'état prévalant au temps t.*

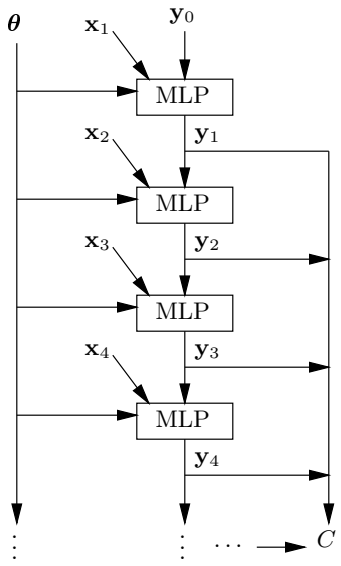


Figure 3.4: *Un réseau récurrent déplié à travers le temps.*

3.1.4 Réseaux récurrents

Les réseaux présentés jusqu'à maintenant sont de nature *statique* : étant donnée une entrée spécifique, ils produisent toujours le même résultat. Ces réseaux ne maintiennent pas d'information d'état conservée entre les évaluations de fonction.

Les *réseaux récurrents*, au contraire, maintiennent de l'information d'état ; ceci permet au réseau de résumer les opérations passées à l'aide d'un *vecteur d'état*. Pour les réseaux que nous utilisons dans ce mémoire, le vecteur d'état au temps t est un sous-ensemble des sorties du réseau au temps $t-1$. Ce vecteur d'état est simplement fourni en entrée au réseau au temps t, en plus des entrées habituelles (voir figure 3.3).

Entraînement

Le calcul du gradient par rapport aux paramètres $\boldsymbol{\theta}$ dans un réseau réseau récurrent procède, tout comme pour un MLP non-récurrent, par rétropropagation (les paramètres sont considérés identiques pour toute la période ;

ils ne changent pas avec le temps). Cependant, la rétropropagation procède à beaucoup plus grande échelle, sur le graphe du réseau déplié à travers le temps pour toute la séquence (voir figure 3.4). Les contributions au gradient par rapport à $\boldsymbol{\theta}$ sont additionnées pour chaque temps, en prenant soin de tenir compte, au temps t, de la contribution induite par le gradient par rapport aux entrées au temps $t+1$.

Nous revenons sur le calcul du gradient d'un réseau récurrent spécifique à la section 3.4.2.

3.1.5 Les réseaux de neurones comme sous-systèmes

La figure 3.5 illustre comment un MLP peut s'intégrer comme sous-système adaptatif à l'intérieur d'un système plus complexe. Étant donnée une entrée $\mathbf{x}$, le MLP calcule une fonction $\mathbf{y} = f(\mathbf{x}; \boldsymbol{\theta})$, dont le résultat est ensuite transformé par une fonction différentiable g (non-adaptative) pour produire une décision $\mathbf{z}$. Cette décision est évaluée par une fonction de coût C différentiable.

Le gradient de C par rapport aux paramètres $\boldsymbol{\theta}$ du MLP procède par simple application des règles de dérivation en chaîne. Partant du gradient $\frac{\partial C}{\partial \mathbf{z}}$, nous trouvons aisément le gradient $\frac{\partial C}{\partial \mathbf{y}}$, dont le i-ième élément est :

$$\frac{\partial C}{\partial \mathbf{y}_i} = \sum_k \frac{\partial C}{\partial \mathbf{z}_k} \frac{\partial \mathbf{z}_k}{\partial \mathbf{y}_i} \tag{3.26}$$

où la sommation est effectuée pour tous les éléments du vecteur $\mathbf{z}$. Le terme $\frac{\partial \mathbf{z}_k}{\partial \mathbf{y}_i}$ ne fait qu'énumérer chacun des éléments de la matrice jacobienne associée à g.

Finalement, nous calculons le gradient par rapport à $\boldsymbol{\theta}$, dont le i-ième élément est :

$$\frac{\partial C}{\partial \boldsymbol{\theta}_i} = \sum_k \frac{\partial C}{\partial \mathbf{y}_k} \frac{\partial \mathbf{y}_k}{\partial \boldsymbol{\theta}_i}. \tag{3.27}$$

Le calcul du terme $\frac{\partial \mathbf{y}_k}{\partial \boldsymbol{\theta}_i}$ est possible par rétropropagation à travers le MLP, tel qu'expliqué précédemment.

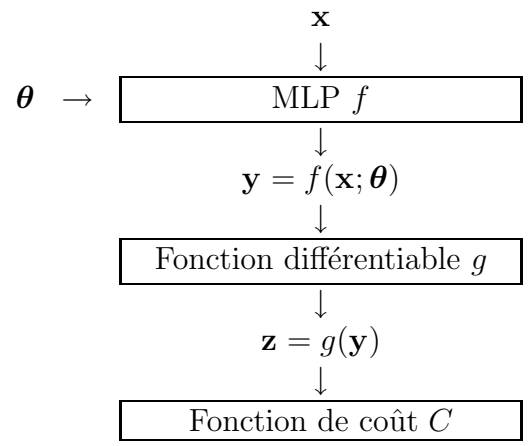


Figure 3.5: *Utilisation d'un MLP comme sous-système adaptatif à l'intérieur d'un système complexe.*

3.2 Deux paradigmes

La figure 3.6 présente un bref aperçu des deux paradigmes de l'utilisation d'un système adaptatif, comme un réseau de neurones, que nous comparons. Ces deux paradigmes se mettent en place dans la partie supérieure de la figure 2.2. La dernière étape illustrée, celle de l'ajustement des positions selon la VàR désirée est telle que décrite à la section 2.3.

Les deux paradigmes diffèrent dans le rôle attribué au système adaptatif :

(a) Dans le premier cas, le système adaptatif est utilisé pour *prédire la distribution jointe du rendement des actifs* à la période suivante, ici notée $\widehat{\Pr}[\mathbf{R}_{t+1}|\mathcal{I}_t]$, ou de moments de celle-ci (tels que la moyenne et la variance). Cette prévision sert ensuite à déduire une allocation « optimale » qui maximise une certaine fonction d'utilité. L'allocation est ensuite rééchelonnée pour atteindre la VàR désirée, de la manière décrite au chapitre précédent.

(b) Dans le second cas, le système adaptatif est utilisé directement pour *produire une décision d'allocation*. Nous notons que l'étape intermédiaire de la prévision est éliminée dans ce second cas. Tout comme dans le cas précédent, l'allocation est ensuite rééchelonnée pour atteindre la VàR désirée.

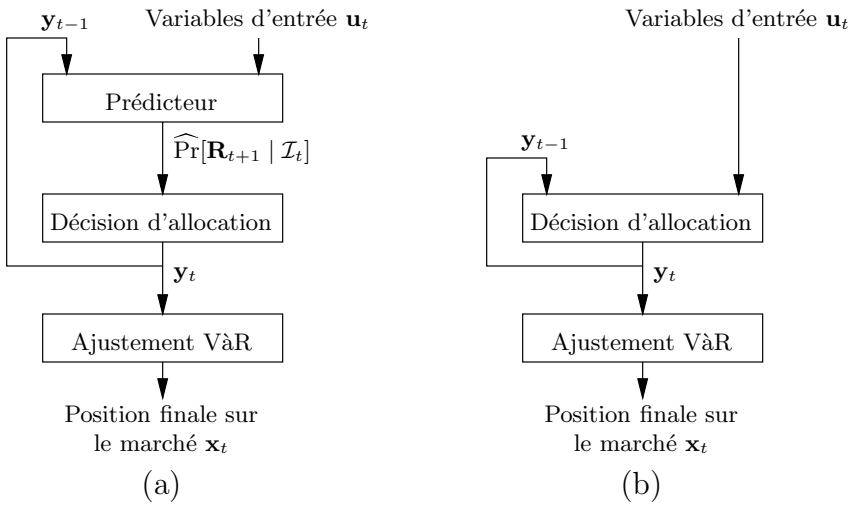


Figure 3.6: *Les deux paradigmes comparés. (a) Utilisation d'un système adaptatif comme prédicteur des rendements, avec un système de décision fixe (§ 3.3) ; (b) Utilisation d'un système adaptatif pour produire directement une décision d'allocation (§ 3.4)*

Dans les deux cas, la boucle de rétroaction qui fournit en entrée au système adaptatif les décisions d'allocation prises au temps précédent est optionnelle.

3.2.1 Intérêt pratique de ces paradigmes

Ces deux paradigmes nous sont d'un intérêt particulier, car ils étudient conjointement deux grands courants d'idées en gestion de portefeuille, et ce, dans un cadre nouveau.

Le premier est d'utiliser des réseaux de neurones pour l'aide à la décision financière ; tel que mentionné précédemment, les réseaux de neurones ont prouvé leur utilité pour la prévision de séries économiques, où ils devancent fréquemment les modèles conventionnels linéaires (Moody 1998). D'autre part, en allocation d'actifs, Bengio (1997) et Moody et Wu (1997) ont obtenu d'excellents résultats en entraînant directement le réseau à rendre des décisions. Toutefois, une comparaison avec les meilleurs systèmes classiques d'allocation (quadratique ou utilisant des fonctions d'utilité plus complexes) n'avait pas été effectuée ; une telle comparaison (limitée au cas de l'allocation

quadratique) est entreprise dans le présent mémoire.

Le second est la longue tradition en finance d'effectuer l'allocation d'actifs à l'intérieur du cadre moyenne–variance de Markowitz (1959). Traditionnellement, à l'intérieur de tels systèmes, les estimés requis de moyennes et variances proviennent de sources *ad hoc* comme les prévisions d'analystes, ou encore sont effectuées à l'aide de modèles classiques, comme un modèle de type ARMA pour prévoir les rendements et de type GARCH pour prévoir la variance. Dans ce mémoire, nous utilisons un réseau de neurones pour prévoir la moyenne des rendements des actifs, étant données des variables explicatives jugées utiles à la tâche de prévision.[5]

Le cadre nouveau est l'allocation sous contrainte de valeur à risque, qui, telle que nous la définissons, n'impose pas la positivité des positions comme elle est normalement exigée dans une allocation classique d'actifs. Cette relaxation permet d'explorer des avenues nouvelles, comme celle de rétropropager à travers une fonction d'allocation quadratique en vue d'entraîner un MLP, comme nous l'expliquons à la section 3.3.4.

3.3 Modèle de prévision

3.3.1 Schéma général

Le modèle de prévision opère dans le cadre général d'une procédure visant à trouver une décision optimale d'allocation des actifs, celle maximisant l'espérance d'une fonction d'utilité fixée à priori, étant donnée une distribution de probabilité des rendements.

Le rôle imputé au système adaptatif dans un tel modèle est fournir des prévisions sur la distribution des rendements futurs des actifs. La prévision sert ensuite de point de départ à une fonction de décision (qui *n'est pas adaptative*) qui produit une allocation des actifs. Nous commençons cette section en dérivant les équations utilisées par la fonction d'allocation quadratique. En-

[5] Nous n'effectuons pas de comparaison avec des systèmes classiques de prévision.

suite, nous expliquons différentes stratégies pour entraîner un MLP à produire de bonnes prévisions.

3.3.2 Maximisation de l'utilité

Nous supposons qu'un investisseur retire une certaine utilité U de la performance de ses placements. Par exemple, l'investisseur uniquement intéressé à maximiser le rendement utilisera comme utilité le rendement du portefeuille réalisé à chaque période :

$$U(\mathbf{R}_{t+1}, \mathbf{w}_t) = \mathbf{R}'_{t+1}\mathbf{w}_t, \tag{3.28}$$

où $\mathbf{w}_t$ représente la proportion (relative) de chaque actif dans le portefeuille ; nous imposons à $\mathbf{w}_t$ la contrainte de somme-à-un : $\sum_i w_{it} = 1$.[6] De plus, nous supposons $\mathbf{w}_t$ connu selon $\mathcal{I}_t$, l'information disponible au temps t.

Le problème de maximisation de l'utilité consiste en choisir, au temps t, la pondération optimale $\mathbf{w}_t^*$ qui maximise l'espérance de l'utilité qui sera obtenue au temps $t+1$:

$$\mathbf{w}_t^* = \arg\max_{\mathbf{w}_t} \mathrm{E}[U(\mathbf{R}_{t+1}, \mathbf{w}_t) \mid \mathcal{I}_t]. \tag{3.29}$$

Ici, $\mathbf{R}_{t+1}$ est la variable aléatoire représentant le rendement des actifs entre t et $t+1$ et l'espérance est conditionnelle à l'information disponible au temps t. La distribution de $\mathbf{R}_{t+1}$ peut être estimée de plusieurs façons ; nous traitons de quelques méthodes d'estimation dans les sections suivantes.

L'espérance de l'éq. (3.29) peut s'exprimer sous la forme d'une intégrale :

$$\mathrm{E}[U(\mathbf{R}_{t+1}, \mathbf{w}_t) \mid \mathcal{I}_t] = \int_{\mathbf{R}_{t+1}} P_{t+1|t}(\mathbf{r})\, U(\mathbf{r}, \mathbf{w}_t)\, d\mathbf{r}, \tag{3.30}$$

où $P_{t+1|t}(\cdot)$ est la fonction de densité de la distribution de $\mathbf{R}_{t+1}$ étant donné l'information disponible au temps t.

[6] Nous imposons la contrainte de somme-à-un uniquement dans la présente section afin de faciliter la dérivation des équations d'allocation ; cette contrainte n'a pas d'impact final sur la taille des portefeuilles effectivement investis après rééchelonnement par la VàR cible.

Pour certaines fonctions d'utilité simples, l'éq. (3.30) se résout de façon analytique, et nous pouvons trouver des solutions exactes au problème d'optimisation (3.29). Le premier cas que nous examinons est l'utilité quadratique, celle principalement utilisée dans ce mémoire. Les fonctions d'utilité plus complexes ne pourront en général s'intégrer qu'avec des méthodes numériques, dont nous traitons brièvement à la section 3.3.5.

3.3.3 Utilité quadratique

La fonction d'utilité simple présentée à l'éq. (3.28) souffre d'un défaut majeur : elle n'encourage que les allocations conduisant à des rendements élevés, sans égard au risque encouru pour produire ces rendements ; elle ne correspond pas au profil de l'investisseur-type, risquophobe, qui n'est pas disposé à obtenir de grands rendements si le risque de grandes pertes est tout aussi probable. Il est donc nécessaire de corriger cette fonction d'utilité par une pénalité qui s'accroîtra dans la mesure des « surprises » produites par les rendements.

Posons les rendements pour la période $t+1$ distribués conditionnellement à $\mathcal{I}_t$ selon une normale,

$$\mathbf{R}_{t+1} \;\sim\; \mathcal{N}(\boldsymbol{\mu}_{t+1}, \boldsymbol{\Gamma}_{t+1}),$$

et où la matrice $\boldsymbol{\Gamma}_{t+1}$ est supposée définie-positive.

La fonction d'*utilité quadratique* prend la forme :

$$U(\mathbf{R}_{t+1}, \mathbf{w}_t) = \mathbf{R}'_{t+1}\mathbf{w}_t - \lambda \left(\mathbf{w}'_t(\mathbf{R}_{t+1} - \boldsymbol{\mu}_{t+1})\right)^2. \tag{3.31}$$

Nous remarquons que cette utilité favorise les bons rendements réalisés (premier terme), mais pénalise la différence–carrée entre le rendement effectivement réalisé ($\mathbf{w}'_t\mathbf{R}_{t+1}$) et celui espéré ($\mathbf{w}'_t\boldsymbol{\mu}_{t+1}$).

Le paramètre $\lambda \in \mathbb{R}^+$ (à ne pas confondre avec le facteur d'oubli d'une moyenne pondérée exponentiellement) détermine l'*aversion au risque* de l'investisseur, c'est-à-dire le degré auquel l'investisseur pénalise les surprises obtenues. Un investisseur hautement risquophobe sera gouverné par un λ élevé ; en contrepartie, un investisseur indifférent au risque préférera un λ plus faible.

Dans tous les cas, λ doit être strictement positif.

Calcul de l'espérance

L'espérance (3.30) de l'utilité quadratique (3.31) peut s'exprimer analytiquement en fonction de $\boldsymbol{\mu}_{t+1}$ et $\boldsymbol{\Gamma}_{t+1}$. Tout d'abord, nous notons que

$$\left(\mathbf{w}_t'(\mathbf{R}_{t+1} - \boldsymbol{\mu}_{t+1})\right)^2 = \mathbf{w}_t'(\mathbf{R}_{t+1} - \boldsymbol{\mu}_{t+1})(\mathbf{R}_{t+1} - \boldsymbol{\mu}_{t+1})'\mathbf{w}_t,$$

ce qui permet de réécrire l'éq. (3.31) comme

$$U(\mathbf{R}_{t+1}, \mathbf{w}_t) = \mathbf{R}_{t+1}'\mathbf{w}_t - \lambda \mathbf{w}_t'(\mathbf{R}_{t+1} - \boldsymbol{\mu}_{t+1})(\mathbf{R}_{t+1} - \boldsymbol{\mu}_{t+1})'\mathbf{w}_t. \tag{3.32}$$

Substituant ce résultat dans l'éq. (3.30), nous obtenons :

$$\begin{aligned}
\mathrm{E}[U(\mathbf{R}_{t+1}, \mathbf{w}_t) \mid \mathcal{I}_t] &= \int_{\mathbf{R}_{t+1}} P_{t+1|t}(\mathbf{r}) \left(\mathbf{r}'\mathbf{w}_t - \lambda \mathbf{w}_t'(\mathbf{r} - \boldsymbol{\mu}_{t+1})(\mathbf{r} - \boldsymbol{\mu}_{t+1})'\mathbf{w}_t\right) d\mathbf{r} \\
&= \left(\int_{\mathbf{R}_{t+1}} P_{t+1|t}(\mathbf{r})\mathbf{r}' d\mathbf{r}\right) \mathbf{w}_t \\
&\quad -\lambda \mathbf{w}_t' \left(\int_{\mathbf{R}_{t+1}} P_{t+1|t}(\mathbf{r})(\mathbf{r} - \boldsymbol{\mu}_{t+1})(\mathbf{r} - \boldsymbol{\mu}_{t+1})' d\mathbf{r}\right) \mathbf{w}_t \\
&= \boldsymbol{\mu}_{t+1}'\mathbf{w}_t - \lambda \mathbf{w}_t' \boldsymbol{\Gamma}_{t+1} \mathbf{w}_t. \qquad (3.33)
\end{aligned}$$

où, comme plus haut, $P_{t+1|t}(\cdot)$ est la fonction de densité conditionnelle des rendements étant donnée l'information disponible au temps t, soit dans le cas présent la densité de $\mathcal{N}(\boldsymbol{\mu}_{t+1}, \boldsymbol{\Gamma}_{t+1})$.

Nous pouvons estimer, au temps t, cette utilité espérée en substituant des estimateurs dans l'éq. (3.33) :

$$\hat{U}_{t+1}(\mathbf{w}_t) = \hat{\boldsymbol{\mu}}_{t+1|t}'\mathbf{w}_t - \lambda \mathbf{w}_t' \hat{\boldsymbol{\Gamma}}_{t+1|t} \mathbf{w}_t, \tag{3.34}$$

où $\hat{\boldsymbol{\mu}}_{t+1|t}$ et $\hat{\boldsymbol{\Gamma}}_{t+1|t}$ dénotent des estimateurs au temps $t+1$ calculé en fonction de $\mathcal{I}_t$. Si ces estimateurs sont non-biaisés, $\hat{U}_{t+1}(\mathbf{w}_t)$ l'est aussi.[7]

[7] Nous abusons légèrement de la notation en dénotant par $\hat{U}$ l'estimateur d'utilité *espérée*.

3.3.4 Équations d'allocation

Partant de l'éq. (3.33), nous pouvons déterminer la pondération $\mathbf{w}_t^*$ qui produira l'utilité espérée maximale, étant donné $\mathcal{I}_t$. Nous imposons d'abord la contrainte que tout l'actif doit être investi, c'est-à-dire

$$\sum_{i=1}^{N} w_{it} = 1 \qquad \text{ou encore} \qquad \mathbf{w}_t' \, \boldsymbol{\iota} = 1. \tag{3.35}$$

Nous incorporons ensuite cette contrainte dans un lagrangien formé à partir de l'éq. (3.33), en remarquant que la maximisation de cette équation revient à minimiser la négative de celle-ci :

$$\mathcal{L}(\mathbf{w}_t, \alpha) = -\boldsymbol{\mu}_{t+1}' \mathbf{w}_t + \lambda \mathbf{w}_t' \boldsymbol{\Gamma}_{t+1} \mathbf{w}_t + \alpha(\mathbf{w}_t' \boldsymbol{\iota} - 1). \tag{3.36}$$

Dérivant par rapport à $\mathbf{w}_t$, et posant ces équations égales à zéro, nous obtenons :

$$\frac{\partial \mathcal{L}(\mathbf{w}_t, \alpha)}{\partial \mathbf{w}_t} = -\boldsymbol{\mu}_{t+1} + \lambda \boldsymbol{\Gamma}_{t+1} \mathbf{w}_t + \alpha \boldsymbol{\iota} = 0 \tag{3.37}$$

$$\frac{\partial \mathcal{L}(\mathbf{w}_t, \alpha)}{\partial \alpha} = \mathbf{w}_t' \boldsymbol{\iota} - 1 = 0, \tag{3.38}$$

d'où

$$\mathbf{w}_t^* = \frac{1}{\lambda} \boldsymbol{\Gamma}_{t+1}^{-1} \left(\boldsymbol{\mu}_{t+1} - \alpha \boldsymbol{\iota} \right). \tag{3.39}$$

L'inverse $\boldsymbol{\Gamma}_{t+1}^{-1}$ existe, car $\boldsymbol{\Gamma}_{t+1}$ est définie-positive par hypothèse. L'équation précédente est valable pour une valeur de α telle que la contrainte à l'éq. (3.35) soit respectée. Pour trouver ce α, nous prémultiplions (3.39) par $\boldsymbol{\iota}'$ et tirons parti du fait que $\boldsymbol{\iota}' \mathbf{w}_t^* = 1$:

$$1 = \boldsymbol{\iota}' \mathbf{w}_t^* = \frac{1}{\lambda} \boldsymbol{\iota}' \boldsymbol{\Gamma}_{t+1}^{-1} \left(\boldsymbol{\mu}_{t+1} - \alpha \, \boldsymbol{\iota} \right).$$

Réarrangeons les termes pour obtenir

$$\lambda - \boldsymbol{\iota}' \boldsymbol{\Gamma}_{t+1}^{-1} \boldsymbol{\mu}_{t+1} = -\alpha \boldsymbol{\iota}' \boldsymbol{\Gamma}_{t+1}^{-1} \boldsymbol{\iota},$$

d'où

$$\alpha = \frac{\boldsymbol{\iota}' \boldsymbol{\Gamma}_{t+1}^{-1} \boldsymbol{\mu}_{t+1} - \lambda}{\boldsymbol{\iota}' \boldsymbol{\Gamma}_{t+1}^{-1} \boldsymbol{\iota}} \,. \tag{3.40}$$

Nous pouvons donc combiner les éq. (3.39) et (3.40) pour obtenir la décision qui maximise l'utilité quadratique espérée au temps $t+1$, étant donné $\mathcal{I}_t$:

$$\mathbf{w}_t^* = \frac{1}{\lambda} \boldsymbol{\Gamma}_{t+1}^{-1} \left(\boldsymbol{\mu}_{t+1} - \frac{\boldsymbol{\iota}' \boldsymbol{\Gamma}_{t+1}^{-1} \boldsymbol{\mu}_{t+1} - \lambda}{\boldsymbol{\iota}' \boldsymbol{\Gamma}_{t+1}^{-1} \boldsymbol{\iota}} \boldsymbol{\iota} \right) \tag{3.41}$$

Puisque la décision optimale résultant d'une fonction d'utilité quadratique n'utilise que les deux premiers moments de la distribution des rendements, on nomme généralement *allocation moyenne–variance* la procédure de gestion de portefeuille qui utilise ces résultats.

Estimation de $\mathbf{w}_t^*$

Nous pouvons estimer $\mathbf{w}_t^*$ en substituant dans l'éq. (3.41) des estimateurs calculés selon $\mathcal{I}_t$:

$$\hat{\mathbf{w}}_t^* = \frac{1}{\lambda} \hat{\boldsymbol{\Gamma}}_{t+1|t}^{-1} \left(\hat{\boldsymbol{\mu}}_{t+1|t} - \frac{\boldsymbol{\iota}' \hat{\boldsymbol{\Gamma}}_{t+1|t}^{-1} \hat{\boldsymbol{\mu}}_{t+1|t} - \lambda}{\boldsymbol{\iota}' \hat{\boldsymbol{\Gamma}}_{t+1|t}^{-1} \boldsymbol{\iota}} \boldsymbol{\iota} \right) . \tag{3.42}$$

À cause de la division par $\boldsymbol{\iota}' \hat{\boldsymbol{\Gamma}}_{t+1|t}^{-1} \boldsymbol{\iota}$, cet estimateur n'est pas sans biais ; cependant, nous ne considérons pas ce problème et utilisons (3.42) sans modification pour rendre des décisions.

Équations de rétropropagation

Comme le montre la figure 3.6, le critère d'utilité quadratique est utilisé à la sortie d'un système adaptatif pour produire une décision d'allocation « optimale » (sous les hypothèses énoncées précédemment). Tel qu'expliqué à la section 3.1.5, nous devons rétropropager à travers cette fonction d'allocation quadratique afin d'être en mesure de calculer le gradient par rapport aux paramètres $\boldsymbol{\theta}$ du système adaptatif (et ainsi procéder à l'entraînement de ce dernier).

Faisant référence à la figure 3.5, les « entrées » de la fonction d'allocation sont simplement les paramètres $\boldsymbol{\mu}_{t+1}$ et $\boldsymbol{\Gamma}_{t+1}$ (ou des estimateurs de ces derniers). Nous trouvons donc le gradient de $\mathbf{w}_t^*$ par rapport à $\boldsymbol{\mu}_{t+1}$. Comme nous l'expliquons plus loin, nous utilisons systématiquement un estimateur de covariance *non-adaptatif* calculé à partir d'une moyenne exponentielle, éq. (2.56) ; nous n'avons donc nul besoin du gradient de $\mathbf{w}_t^*$ par rapport à $\boldsymbol{\Gamma}_{t+1}$.

De l'éq. (3.39), le i-ième élément de $\mathbf{w}_t^*$ est simplement :

$$w_{it}^* = \frac{1}{\lambda}\boldsymbol{\Gamma}^{-1}_{i(t+1)}\,\boldsymbol{\mu}_{t+1} - \frac{\alpha}{\lambda}\boldsymbol{\Gamma}^{-1}_{i(t+1)}\,\boldsymbol{\iota}.$$

Dérivant par rapport au j-ième élément de $\boldsymbol{\mu}_{t+1}$, $\mu_{j(t+1)}$, nous obtenons :

$$\frac{\partial w_{it}^*}{\partial \mu_{j(t+1)}} = \frac{1}{\lambda}\boldsymbol{\Gamma}^{-1}_{ij(t+1)} - \frac{1}{\lambda}\boldsymbol{\Gamma}^{-1}_{i(t+1)}\frac{\partial \alpha}{\partial \mu_{j(t+1)}}. \tag{3.43}$$

De même, selon l'éq. (3.40), nous calculons $\partial\alpha/\partial\mu_{j(t+1)}$ comme étant :

$$\frac{\partial \alpha}{\partial \mu_{j(t+1)}} = \frac{\left(\boldsymbol{\iota}'\boldsymbol{\Gamma}^{-1}_{t+1}\right)_j}{\boldsymbol{\iota}'\boldsymbol{\Gamma}^{-1}_{t+1}\boldsymbol{\iota}}. \tag{3.44}$$

Le numérateur de cette équation calcule la somme de la j-ème colonne de $\boldsymbol{\Gamma}^{-1}_{t+1}$, alors que le dénominateur calcule la somme de tous les éléments de la même matrice.

Finalement, pour compléter le calcul de la rétropropagation, nous aurons en entrée le « gradient venant de la droite », i.e. $\partial C/\partial \mathbf{w}_t^*$, où C est le critère final à optimiser (par exemple, la valeur nette totale, donnée à l'éq. (2.44)), et nous cherchons à connaître le gradient de ce critère par rapport aux entrées $\mu_{j(t+1)}$. Ce gradient est donné par :

$$\frac{\partial C}{\partial \mu_{j(t+1)}} = \sum_{i=1}^{N} \frac{\partial w_{it}^*}{\partial \mu_{j(t+1)}}\frac{\partial C}{\partial w_{it}^*}. \tag{3.45}$$

3.3.5 Au-delà de l'allocation moyenne–variance

L'utilité quadratique (3.31), bien qu'elle soit certes plus satisfaisante que l'utilité simple (3.28), comporte néanmoins ses inconvénients. Son problème majeur est de considérer les surprises comme étant *symétriques* ; en d'autres termes, elle pénalise tout autant les bonnes nouvelles que les mauvaises. Pour remédier à cette situation, nous pouvons considérer des fonctions d'utilité plus générales, qui pourront tenir compte, par exemple, du troisième moment de la distribution conditionnelle de $\mathbf{R}_{t+1}$, ou de différentes statistiques d'ordre jugées propices. De plus, nous pouvons considérer des fonctions d'utilité qui incorporent l'effet des frais de transaction, ou qui tiennent compte du critère final d'évaluation financière, éq. (2.44).

Cependant, avec de telles fonctions, nous perdons généralement la capacité de calculer l'utilité espérée (3.30) de façon analytique, comme il était possible de le faire pour l'utilité quadratique (*cf.* éq. (3.33)). Nous devons recourir à des méthodes d'intégration numériques, comme les méthodes de Monte Carlo qui approximent une intégrale par une sommation (voir, entre autres, Press, Flannery, Teukolsky et Vetterling (1992) et Fishman (1996)) :

$$\mathrm{E}[U(\mathbf{R}_{t+1}, \mathbf{w}) \mid \mathcal{I}_t, \mathbf{w}] = \int_{\mathbf{R}_{t+1}} P_{t+1|t}(\mathbf{r}) U(\mathbf{r}, \mathbf{w}) d\mathbf{r} \approx \frac{1}{M} \sum_{j=1}^{M} U(\mathbf{r}_j, \mathbf{w}), \quad (3.46)$$

où $P_{t+1|t}(\cdot)$ est la fonction de densité conditionnelle des rendements étant donné $\mathcal{I}_t$, et où $\{\mathbf{r}_j\}$ sont tirés i.i.d. de cette distribution.

De façon similaire, la maximisation de l'utilité espérée, telle que requise par l'éq. (3.29) doit se faire numériquement. Si la fonction $U(\mathbf{r}, \mathbf{w})$ est *lisse*, nous notons que le gradient par rapport à la décision se calcule simplement, de par la linéarité des opérateurs différentiels :

$$\begin{aligned} \frac{\partial}{\partial \mathbf{w}} \mathrm{E}[U(\mathbf{R}, \mathbf{w}) \mid \mathbf{w}] &= \frac{\partial}{\partial \mathbf{w}} \int_{\mathbf{R}} P(\mathbf{R}) U(\mathbf{r}, \mathbf{w}) d\mathbf{r} \\ &= \int_{\mathbf{R}} P(\mathbf{R}) \frac{\partial U(\mathbf{r}, \mathbf{w})}{\partial \mathbf{w}} d\mathbf{r}. \end{aligned} \quad (3.47)$$

(Il est permis d'interchanger l'espérance et la dérivée sous des conditions de

régularité appropriées (L'Écuyer 1990; Rubinstein 1989).)

Cette intégrale s'évalue aussi de façon simple par les méthodes de Monte Carlo, du moment que $\frac{\partial U(\mathbf{r},\mathbf{w})}{\partial \mathbf{w}}$ puisse être évalué efficacement.

Détenant les éq. (3.46) et (3.47), calculant respectivement l'utilité espérée d'une décision, et le gradient de cette utilité par rapport à la décision, nous pouvons employer une méthode d'optimisation numérique appropriée, comme le gradient conjugé, pour calculer la décision optimale $\mathbf{w}_t^*$.

3.3.6 Stratégies d'entraînement d'un bon prédicteur

Tel qu'expliqué et justifié à la section 3.2, le rôle imputé au système adaptatif (tel qu'un MLP) dans le modèle de prévision (figure 3.6a) est de fournir des prévisions sur le comportement futur des actifs. Une telle prévision peut prendre deux formes.

D'une part, la prévision peut tenter de décrire toute la distribution conditionnelle (jointe) du rendement futur des actifs. On constate facilement qu'une telle approche devient rapidement insoutenable, en raison de l'accroîssement exponentiel de la taille d'un modèle « exhaustif » de la distribution jointe, en fonction du nombre d'actif.[8]

D'autre part, la prévision peut tenter de décrire quelques uns des moments d'ordre inférieur de la distribution conditionnelle de $\mathbf{R}_{t+1}$, étant donnée de l'information $\mathcal{I}_t$. Nous avons vu précédemment qu'il n'est nécessaire de connaître que les deux premiers moments de cette distribution pour procéder à une allocation moyenne–variance. C'est sur cette dernière approche que nous concentrons nos efforts.

Estimation de la moyenne et des covariances

Nous réduisons l'effort requis du système adaptatif en ne l'utilisant que pour estimer l'espérance conditionnelle des rendements (i.e. le premier moment). Nous utilisons comme estimateur de la matrice de variance–covariance

[8]Par exemple, l'espace requis pour représenter un histogramme en N dimensions croît comme $O(K^N)$, où K est le nombre de divisions comptées dans chaque dimension. Il s'agit un autre exemple de la *malédiction de la dimensionalité* (Bellman 1957).

l'estimateur pondéré exponentiellement décrit précédemment (*cf.* l'éq. (2.56)), le même que celui utilisé pour le rééchelonnement des positions selon la VàR désirée $\tilde{V}_t$.

Dans le modèle de prévision, le système adaptatif est donc utilisé comme suit : étant donné un vecteur d'entrées $\mathbf{u}_t$ (qu'on peut calculer utilisant l'information $\mathcal{I}_t$), il doit produire en sortie un vecteur $\hat{\boldsymbol{\mu}}_{t+1|t}$ qui est un estimateur de l'espérance conditionnelle des rendements.

Types de cibles et entraînement

Le système adaptatif peut être entraîné de deux façons différentes à former un prédicteur d'espérance, la première employant des cibles explicites, et la seconde, des cibles implicites.

Prévisions explicites La méthode classique d'entraînement pour la prévision consiste à fournir une cible explicite que le système adaptatif doit atteindre. Lorsqu'il est entraîné de cette façon (voir la figure 3.7), nous utilisons les rendements réalisés $\mathbf{r}_{t+1}$ comme cibles à atteindre, conjointement à un critère de coût quadratique (que nous calculons à la fin de la séquence, étant donné $\mathcal{I}_T$) :

$$C = \frac{1}{2T} \sum_{t=0}^{T-1} \left\| \hat{\boldsymbol{\mu}}_{t+1|t} - \mathbf{r}_{t+1} \right\|^2,$$

où $\hat{\boldsymbol{\mu}}_{t+1|t}$ est la prévision du système adaptatif (faite étant donné $\mathcal{I}_t$) des rendements au temps $t+1$, et $\|\cdot\|$ est la distance euclidienne. Dans la figure 3.7, QOD_λ dénote une allocation quadratique[9] avec un paramètre d'aversion au risque λ.

Ce critère d'entraînement est utilisé par l'algorithme de rétropropagation (voir section 3.1.3) lors de l'entraînement du système adaptatif. Par exemple, pour un réseau de neurones, nous partons des sorties désirées (notées (**A**) sur la figure), et rétropropageons en (**B**) le gradient par rapport aux sorties

[9] QOD est l'abbréviation de *Quadratic Optimal Decider*, qui est le nom historique donné à ce module.

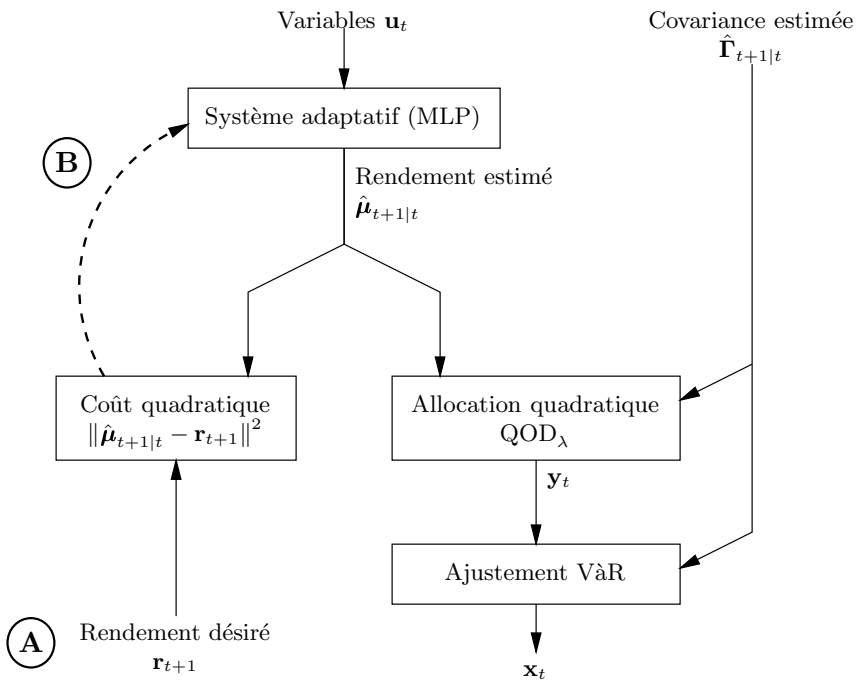


Figure 3.7: *Entraînement d'un système adaptatif à prévoir le rendement espéré, en utilisant explicitement le rendement désiré comme cible d'entraînement.*

$(\partial C/\partial \hat{\boldsymbol{\mu}}_{t+1|t})$ à travers le réseau de neurones, de la façon expliquée à la section 3.1.3.

Prévisions implicites Une autre méthode d'entraînement du réseau n'utilise pas de cibles explicites ; elle utilise plutôt de l'information provenant du critère final à maximiser, le critère de performance financière de l'éq. (2.44). La figure 3.8 illustre cette procédure.

La figure montre que les sorties du réseau sont d'abord combinées avec un *prédicteur naïf*[10] de manière à former le prédicteur du rendement espéré $\hat{\boldsymbol{\mu}}_{t+1|t}$. Nous notons que le réseau n'apprend jamais explicitement les rendements qu'il doit prévoir ; en d'autres termes, il ne reçoit jamais directement d'information sur les rendements effectivement réalisés à chaque période. Cette information

[10]Le prédicteur naïf est, dans notre cas, le rendement historique moyen de chaque actif pour les actifs boursiers, et le rendement à la période précédente pour les actifs « sans risque ».

lui est communiquée indirectement par le gradient qui provient du critère de performance financière.

De façon plus spécifique, nous suivons les étapes suivantes pour rétropropager le gradient (les lettres encerclées font référence à la figure 3.8) :

(A) Le gradient $\partial C/\partial \mathbf{x}_t$ provenant du critère financier est calculé. Le calcul en est expliqué à la section 3.4.2.

(B) Le gradient $\partial C/\partial \mathbf{y}_t$ à travers la normalisation par la VàR désirée est calculé ; voir section 3.4.2.

(C) Finalement, le gradient $\partial C/\partial \hat{\boldsymbol{\mu}}_{t+1|t}$ à travers le module d'allocation quadratique QOD_λ est calculé ; voir section 3.3.4.

Faisant référence à la section 3.1.5, ces trois étapes correspondent à la rétropropagation à travers une fonction non-adaptative lorsqu'un réseau de neurones (par exemple) est utilisé comme sous-système à l'intérieur d'un système plus complexe. Le gradient obtenu obtenu à la suite de ces trois étapes, $\partial C/\partial \hat{\boldsymbol{\mu}}_{t+1|t}$, sert de point de départ pour la rétropropagation à travers les couches du réseau de neurones.

3.4 Modèle de décision

3.4.1 Schéma général

Le modèle de décision (figure 3.6b) se dispense de l'étape de prévision requise précédemment, et conduit le système adaptatif à produire directement une décision d'allocation brute $\mathbf{y}_t$ (étant donné $\mathcal{I}_t$), qui est ensuite normalisée par la VàR désirée selon la procédure habituelle (éq. (2.19)).

Justification du modèle

Avant de considérer les détails de l'entraînement d'un tel système, nous examinons les raisons qui en expliquent l'attrait. Nous notons immédiatement que, pour ce modèle, les étapes suivies par le système adaptatif pour produire

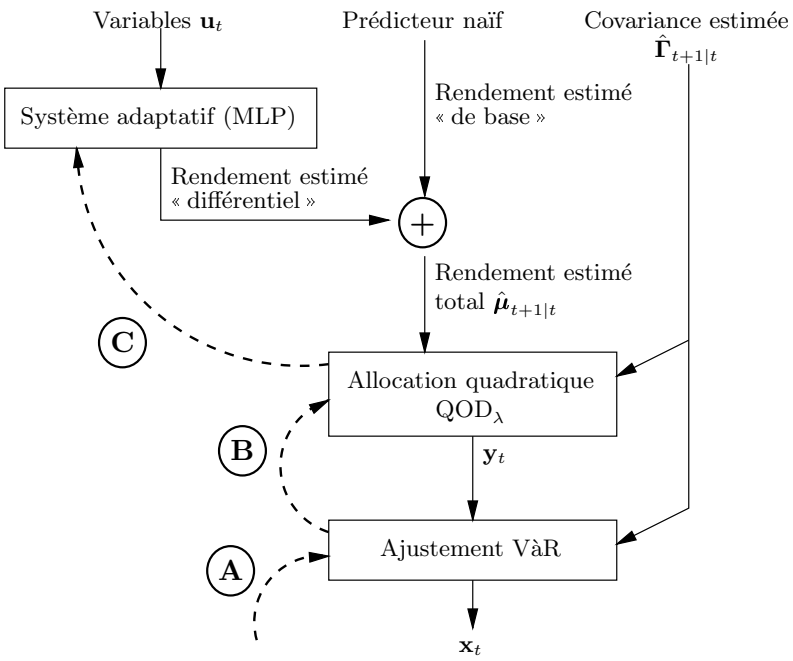


Figure 3.8: *Entraînement d'un système adaptatif à prévoir le rendement espéré, en utilisant le critère final de rendement financier pour conduire l'entraînement.*

une décision sont difficiles à justifier sur un plan théorique : au contraire du modèle de prévision, son rôle ne s'appuie pas sur l'assise théorique attrayante de la maximisation de la fonction d'utilité d'un investisseur, ou encore sur la minimisation d'une erreur de prévision simple et clairement définie. Cependant, nous pouvons croire en l'efficacité pragmatique du modèle de décision en vertu des considérations suivantes :

1. Le problème de l'estimation de densité—qui doit être résolu d'une façon ou d'une autre par le modèle de prévision—est intrinsèquement d'une grande difficulté, particulièrement en haute dimensionalité (rappelons que la dimensionalité est proportionnelle au nombre d'actifs dans le portefeuille). Toute erreur commise dans la prévision aura des répercussions néfastes sur les décisions d'allocation.
2. Le modèle de décision ne trouve pas nécessaire de postuler explicitement une fonction d'utilité permettant un traitement mathématique simple mais qui ne correspond peut-être pas bien aux besoins des investis-

seurs. Le choix de cette fonction est délicat car c'est elle qui conduit aux décisions d'allocation pour le modèle de prévision. Or, dans le cas de l'utilité quadratique, nous savons que cette fonction ne constitue pas la « véritable » utilité de l'investisseur, du fait, comme nous l'avons vu, de sa réaction symétrique aux « surprises » dans les rendements. De plus, cette utilité n'est pas le critère final de rendement financier, éq. (2.44), selon lequel la performance du système est ultimement évaluée.

3. Finalement, il est concevable que le problème de décision (tout spécialement en vue d'optimiser le critère financier mentionné précédemment) soit, d'une façon fondamentale, plus facile à résoudre que le problème de l'estimation de densité.

Entraînement

L'entraînement du système adaptatif suit les grandes lignes présentées précédemment pour l'entraînement du modèle de prévision implicite. La rétropropagation du gradient se fait comme à la figure 3.8, à l'exception que nous tenons compte explicitement de la boucle de rétroaction de la figure 3.6b, qui fournit en entrée au système adaptatif, pour le temps t, les décisions d'allocation $\mathbf{y}_{t-1}$ prises au temps précédent.

Cette boucle de rétroaction introduit une récurrence dans le réseau, laquelle complique sensiblement la dérivation des équations de rétropropagation. Cette boucle est requise pour les raisons suivantes :

– D'une part, les frais de transaction produisent un couplage entre deux temps successifs : la décision rendue au temps t affecte à la fois les frais encourus au temps t et ceux au temps $t+1$. Ce couplage induit à son tour un gradient par rapport aux positions $\mathbf{x}_t$ provenant de $\mathbf{x}_{t+1}$, et cette information peut être pertinente pour guider l'entraînement. Nous approfondissons ces dépendances à la section suivante.

– D'autre part, la connaissance de la décision au temps précédent peut permettre au réseau—en théorie—d'apprendre une stratégie qui minimise les frais de transaction : étant donné un choix entre deux positions

aussi viables au temps t, le réseau peut minimiser les frais en choisissant celle qui est la plus rapprochée de la position au temps $t-1$; pour cette raison, la connaissance de $\mathbf{y}_{t-1}$ est pertinente. Malheureusement, cet idéal ne peut jamais être atteint complètement, car le processus de rééchelonnement des positions pour atteindre la VàR désirée est toujours effectué « inconditionnellement », c'est-à-dire sans égard à la position précédente.

Les équations de rétropropagation tenant compte de la récurrence sont présentées à la section suivante. Il faut noter que ces équations seront, pour un bref moment, légèrement incomplètes : la section d'ensuite présente certaines conditions de régularisation qui assureront l'existence de points localement minima de la fonction de coût.

3.4.2 Équations de rétropropagation

Les équations de rétropropagation sont obtenues de la façon habituelle, en traversant le graphe de flot dans le sens inverse des flèches, et en accumulant toutes les contributions du gradient à un noeud. Il faut porter une attention particulière à deux conséquences de la récurrence dans le réseau :

- Le gradient par rapport aux *entrées* du MLP au temps t fournit une contribution au gradient par rapport aux *sorties* du MLP au temps $t-1$.
- Les frais de transactions au temps t dépendent des décisions prises à la fois au temps t et $t-1$. Donc, il faut tenir compte des frais au temps t pour le calcul du gradient à ces deux temps consécutifs.

La figure 3.9 illustre le graphe de flot de tout le système d'allocation, déplié à travers le temps. Nous calculerons le gradient de *manière récursive*, en remontant dans le temps, partant du dernier instant T jusqu'au premier.

Par simplicité, nous supposons qu'un MLP $\mathcal{M}$ (que nous supposons ici être le système adaptatif) calcule une fonction paramétrisée par $\boldsymbol{\theta}$:

$$\mathbf{y}_t = f_{\mathcal{M}}(\mathbf{y}_{t-1}, \mathbf{u}_t; \boldsymbol{\theta}), \tag{3.48}$$

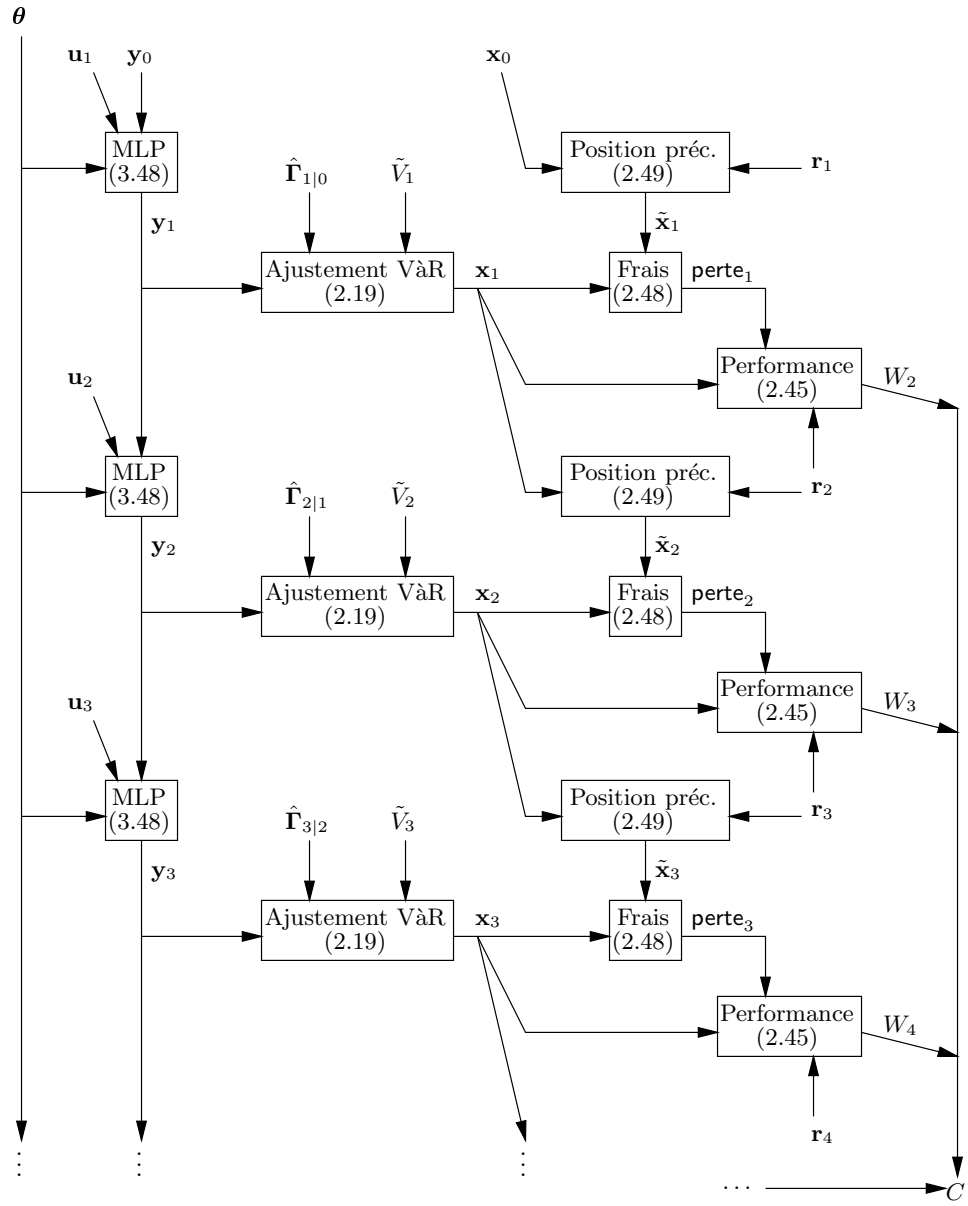


Figure 3.9: *Graphe de flot déplié à travers le temps permettant de dériver les équations de rétropropagation pour le modèle de décision. Les chiffres entre parenthèses font référence équations (dans le texte) utilisées pour calculer chaque valeur.*

où $\mathbf{u}_t$ est un vecteur de variables explicatives jugées utiles à la tâche d'allocation, qu'on peut calculer étant donné $\mathcal{I}_t$.

Le critère C qu'on désire *minimiser* (en bas à droite sur la figure 3.9) est relié au critère de performance financière $\hat{W}$ (éq. (2.46)) :

$$C = -\hat{W}. \tag{3.49}$$

Ce critère ne peut être déterminé qu'à la fin de la séquence, quand on connaît $\mathcal{I}_T$.

Partant de l'éq. (2.44) nous commençons par calculer la contribution apportée au critère total par le gain à chaque temps :

$$\frac{\partial C}{\partial \hat{W}_{t+1}} = -\frac{1}{T}. \tag{3.50}$$

Ensuite, nous nous servons des éq. (2.45), (2.48) et (2.49) pour déterminer la contribution des frais de transaction au gradient :

$$\begin{aligned}
\frac{\partial C}{\partial \mathsf{perte}_t} &= -\frac{1}{T\,\tilde{V}_t} && (3.51)\\
\frac{\partial \mathsf{perte}_t}{\partial x_{it}} &= -c_i\ \mathrm{sign}(x_{it} - \tilde{x}_{it}) && (3.52)\\
\frac{\partial \mathsf{perte}_t}{\partial \tilde{x}_{it}} &= c_i\ \mathrm{sign}(x_{it} - \tilde{x}_{it}) && (3.53)\\
\frac{\partial \mathsf{perte}_{t+1}}{\partial x_{it}} &= c_i\ \mathrm{sign}(x_{i(t+1)} - \tilde{x}_{i(t+1)})\ \frac{\partial \tilde{x}_{i(t+1)}}{\partial x_{it}} \\
&= c_i\ \mathrm{sign}(x_{i(t+1)} - \tilde{x}_{i(t+1)})\ (1 + r_{i(t+1)}). && (3.54)
\end{aligned}$$

Dès lors, utilisant de nouveau l'éq. (2.45), nous calculons la contribution de x_{it} au gradient, qui provient des deux "chemins" par lesquels x_{it} affecte C : une première contribution, directe, par les rendements entre t et $t+1$; ainsi qu'une seconde contribution, indirecte, par les frais de transaction au moment de transiger à $t+1$.

$$\frac{\partial C}{\partial x_{it}} = \frac{\partial C}{\partial \hat{W}_{t+1}} \frac{\partial \hat{W}_{t+1}}{\partial x_{it}} + \frac{\partial C}{\partial \hat{W}_{t+2}} \frac{\partial \hat{W}_{t+2}}{\partial x_{it}}. \tag{3.55}$$

Puisque $\partial C/\partial \hat{W}_{t+1}$ est simplement donné par l'éq. (3.50), nous utilisons l'éq. (2.45) pour calculer

$$\begin{aligned}\frac{\partial C}{\partial \hat{W}_{t+1}}\frac{\partial \hat{W}_{t+1}}{\partial x_{it}} &= -\frac{1}{T\,\tilde{V}_t}\left(r_{i(t+1)} - r_{0t} + \frac{\partial \mathsf{perte}_t}{\partial x_{it}}\right) \\ &= -\frac{1}{T\,\tilde{V}_t}\left(r_{i(t+1)} - r_{0t} - c_i\ \mathrm{sign}(x_{it} - \tilde{x}_{it})\right). \qquad (3.56)\end{aligned}$$

De la même manière, nous calculons la contribution

$$\begin{aligned}\frac{\partial C}{\partial \hat{W}_{t+2}}\frac{\partial \hat{W}_{t+2}}{\partial x_{it}} &= -\frac{1}{T\,\tilde{V}_{t+1}}\frac{\partial \mathsf{perte}_{t+1}}{\partial x_{it}} \\ &= -\frac{1}{T\,\tilde{V}_{t+1}}\left(c_i\ \mathrm{sign}(x_{i(t+1)} - \tilde{x}_{i(t+1)})\,(1 + r_{i(t+1)})\right). \qquad (3.57)\end{aligned}$$

Finalement, nous additionnons ces deux dernières équations pour obtenir :

$$\begin{aligned}\frac{\partial C}{\partial x_{it}} &= -\frac{1}{T\,\tilde{V}_t}\left(r_{i(t+1)} - r_{0t} - c_i\ \mathrm{sign}(x_{it} - \tilde{x}_{it})\right) \\ &\quad -\frac{1}{T\,\tilde{V}_{t+1}}\left(c_i\ \mathrm{sign}(x_{i(t+1)} - \tilde{x}_{i(t+1)})\,(1 + r_{i(t+1)})\right). \qquad (3.58)\end{aligned}$$

Nous sommes maintenant en mesure de calculer le gradient par rapport aux sorties du réseau de neurones. Utilisant les éq. (2.19) et (2.20), nous commençons par évaluer l'effet de y_{it} sur x_{it} :[11]

$$\frac{\partial x_{it}}{\partial y_{it}} = \frac{\tilde{V}_t}{\Phi^{-1}(\alpha)\left(\mathbf{y}_t'\,\hat{\mathbf{\Gamma}}_{t+1|t}\,\mathbf{y}_t\right)^{\frac{1}{2}}} - \frac{y_{it}\,\tilde{V}_t\,\sum_{k=1}^{N}\hat{\gamma}_{ik(t+1)}\,y_{kt}}{\Phi^{-1}(\alpha)\,\left(\mathbf{y}_t'\,\hat{\mathbf{\Gamma}}_{t+1|t}\,\mathbf{y}_t\right)^{\frac{3}{2}}}, \qquad (3.59)$$

et pour $i \neq j$,

$$\frac{\partial x_{it}}{\partial y_{jt}} = -\frac{y_{it}\,\tilde{V}_t\,\sum_{k=1}^{N}\hat{\gamma}_{jk(t+1)}\,y_{kt}}{\Phi^{-1}(\alpha)\,\left(\mathbf{y}_t'\,\hat{\mathbf{\Gamma}}_{t+1|t}\,\mathbf{y}_t\right)^{\frac{3}{2}}}. \qquad (3.60)$$

[11] Pour parvenir à ces équations, il est utile de rappeler que $\mathbf{y}'\,\mathbf{\Gamma}\,\mathbf{y}$ peut être écrit sous la forme $\sum_k \sum_\ell \gamma_{k\ell}\,y_k\,y_\ell$; d'où il vient facilement que $\frac{\partial}{\partial y_i}\mathbf{y}'\,\mathbf{\Gamma}\,\mathbf{y} = 2\sum_k \gamma_{ik}\,y_k$.

(Rappelons que α est le niveau désiré de la VàR et que $\Phi^{-1}(\cdot)$ est la fonction de répartition inverse de la distribution normale centrée réduite).

Le gradient complet est donné par

$$\frac{\partial C}{\partial y_{it}} = \sum_k \frac{\partial C}{\partial x_{kt}} \frac{\partial x_{kt}}{\partial y_{it}} + \frac{\partial C}{\partial f_{\mathcal{M}t+1}}, \tag{3.61}$$

où $\partial C / \partial f_{\mathcal{M}t+1}$ représente le gradient par rapport aux entrées du réseau de neurones au temps $t+1$.

3.4.3 Régularisation de la fonction de coût

La fonction de coût (3.49) correspondant au critère de performance financière (2.44) souffre d'un inconvénient de taille lors de son utilisation pour l'entraînement d'un système adaptatif.

Pour comprendre la nature exacte du problème, il est utile de rappeler la procédure (*cf.* éq. (2.19)) qui rééchelonne les positions « brutes » $\mathbf{y}_t$ pour produire des positions finales $\mathbf{x}_t$ qui respectent la VàR désirée. Soit deux positions brutes $\mathbf{y}_t^{(1)}$ et $\mathbf{y}_t^{(2)}$ qui ne diffèrent que par une constante multiplicative : $\mathbf{y}_t^{(2)} = \delta \mathbf{y}_t^{(1)}$. Les positions finales après rééchelonnement sont (voir éq. (2.19) et (2.20)) :

$$\hat{\beta}_t^{(1)} = \frac{\tilde{V}_t}{\Phi^{-1}(\alpha) \sqrt{\mathbf{y}_t^{(1)\prime} \, \hat{\mathbf{\Gamma}}_{t+1|t} \, \mathbf{y}_t^{(1)}}} \qquad \hat{\beta}_t^{(2)} = \frac{\tilde{V}_t}{\Phi^{-1}(\alpha) \, \delta \sqrt{\mathbf{y}_t^{(2)\prime} \, \hat{\mathbf{\Gamma}}_{t+1|t} \, \mathbf{y}_t^{(2)}}}$$

$$\mathbf{x}_t^{(1)} = \hat{\beta}_t^{(1)} \, \mathbf{y}_t^{(1)} \qquad \boxed{\mathbf{x}_t^{(2)} = \hat{\beta}_t^{(2)} \, \mathbf{y}_t^{(2)} = \frac{\hat{\beta}_t^{(1)}}{\delta} \, \delta \mathbf{y}_t^{(1)} = \mathbf{x}_t^{(1)}}$$

En d'autres termes, des positions brutes qui ne diffèrent que par une constante multiplicative deviennent, après rééchelonnement, *la même position finale.*

Cet effet est illustré à la figure 3.10 qui montre les courbes de niveau d'une coupe d'une fonction de coût pour un problème d'allocation entre trois actifs (un actif sans risque, l'indice TSE 300, et un indice obligataire). Le coût résultant de différentes positions brutes prises sur le TSE et sur les obligations est tracé ; la pondération prise sur l'actif sans risque est fixée à zéro.

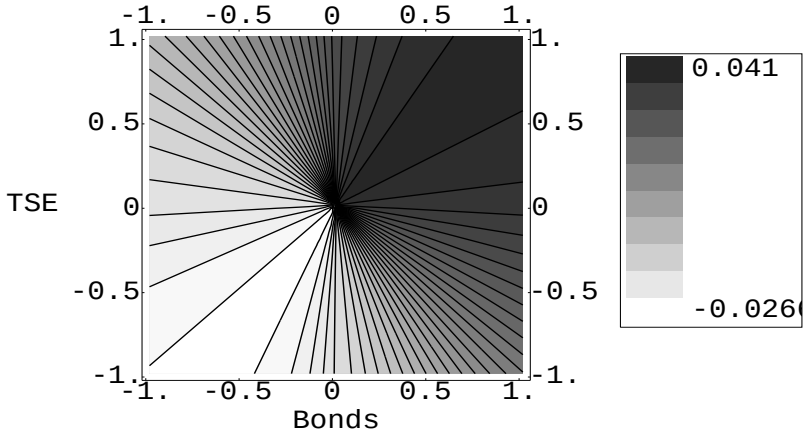


Figure 3.10: *Coupe de la fonction de coût non-régularisée, pour un problème d'allocation de trois actifs. La pondération du 3e actif (l'actif sans risque) est zéro.*

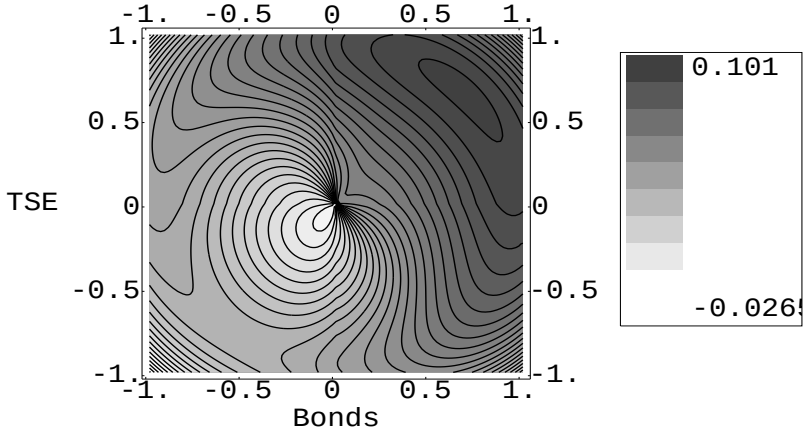


Figure 3.11: *Coupe de la fonction de coût régularisée, pour un problème d'allocation de trois actifs. La pondération du 3e actif (l'actif sans risque) est zéro.*

L'effet d'« équivalence » des positions brutes apparaît clairement sur la figure. Nous constatons que le coût demeure constant pour toutes les positions situées sur un même rayon partant de l'origine, c'est-à-dire sur des positions égales à une constante multiplicative (positive) près.

La conséquence de cette situation est faire en sorte que le problème d'optimisation des paramètres du système adaptatif ne *comporte pas de solution unique* : deux systèmes produisant des solutions égales à une constante près seront jugés comme équivalents par la fonction de coût. Malheureusement, l'algorithme d'optimisation numérique utilisé dans l'entraînement cherche à trouver **la** solution, dans l'espace des paramètres, qui produira le coût minimum. Or, **cette** solution n'existe pas : il y en a une infinité. La fonction de coût, telle qu'elle est spécifiée, peut conduire à de graves divergences des paramètres lors de l'entraînement.

La solution à ce problème est d'introduire des *préférences à priori* sur certaines positions plutôt que d'autres. Nous présentons deux manières de s'acquitter de cette tâche.[12]

Préférence sur la norme des sorties

Comme le montre la figure 3.10, toutes les solutions situées sur un même rayon de l'origine engendrent le même coût. Un premier type de préférence que nous pouvons introduire est de fixer une *longueur cible ρ que devrait avoir une solution*. Selon ce critère, la fonction de coût à minimiser devient :

$$C' = -\hat{W} + C_{\mathsf{norme}} \tag{3.62}$$

[12]Nous préférons ne pas imposer de contraintes « dures » sur l'espace des solutions, car notre problème d'optimisation se pose autrement très bien en se passant de telles contraintes, et l'introduction de ces dernières compliquerait considérablement l'implantation pratique de l'algorithme d'optimisation.

où

$$C_{\mathsf{norme}} = \frac{1}{T}\sum_{t=1}^{T} \mathsf{pén.}_{\mathsf{norme}}(\mathbf{y}_t) \tag{3.63}$$

$$\mathsf{pén.}_{\mathsf{norme}}(\mathbf{y}) = \frac{\phi_{\mathsf{norme}}}{2}\left(\sum_{i=1}^{N} y_i^2 - \rho^2\right)^2. \tag{3.64}$$

Le terme de pénalisation comporte deux paramètres à déterminer : ρ, qui est la norme désirée pour les sorties du système adaptatif,[13] et ϕ_{norme}, qui gouverne l'importance accordée à cette pénalisation dans le coût total.

Le terme de pénalisation entraîne une modification modeste aux équations de rétropropagation. Le seul changement à apporter est à $\partial C/\partial y_{it}$, éq. (3.61), qui devient :

$$\frac{\partial C_{\mathsf{norme}}}{\partial y_{it}} = \frac{\partial C}{\partial y_{it}} + \frac{\phi_{\mathsf{norme}}}{T}(\sum_{j=1}^{N} y_{jt}^2 - \rho^2)(2y_{it}). \tag{3.65}$$

La figure 3.11 montre l'effet d'introduire cette préférence sur la norme des sorties (avec $\rho^2 = 0.9$ et $\phi_{\mathsf{norme}} = 0.1$). La solution optimale demeure dans la même direction qu'auparavant, mais elle maintenant encouragée à avoir une longueur ρ.

Préférence sur un portefeuille de référence

Dans certaines circonstances, nous pouvons connaître à priori des positions brutes qu'on suppose produire de bons résultats, ou qui sont préférables à cause de contraintes réglementaires. Par exemple, un gestionnaire de fonds peut être mandaté de pondérer son portefeuille de telle sorte à ce qu'il contienne « approximativement » 60% d'actions et 40% d'obligations. Cette obligation (qui résulte de politiques sur lesquelles le gestionnaire n'a aucun contrôle) constitue un point de référence, et nous préférons que le système adaptatif ne s'en éloigne pas trop. Le second type de préférence que nous introduisons est donc

[13] Dans la majorité de nos expériences, nous avons fixé arbitrairement $\rho^2 = 0.9$.

par rapport à un portefeuille de référence $\boldsymbol{\psi}_t$. La fonction de coût que nous désirons minimiser devient :

$$C' = -\hat{W} + C_{\mathsf{port.\ ref.}} \tag{3.66}$$

où

$$C_{\mathsf{port.\ ref.}} = \frac{1}{T}\sum_{t=1}^{T} \mathsf{pén.}_{\mathsf{port.\ ref.}}(\mathbf{y}_t) \tag{3.67}$$

$$\mathsf{pén.}_{\mathsf{port.\ ref.}}(\mathbf{y}_t) = \frac{\phi_{\mathsf{port.\ ref.}}}{2}\left\|\mathbf{y}_t - \boldsymbol{\psi}_t\right\|^2, \tag{3.68}$$

avec $\|\cdot\|$ la distance euclidienne.

Les équations de rétropropagation sont aussi simples à modifier que dans le cas précédent : nous ajoutons une contribution à $\partial C/\partial y_{it}$, éq. (3.61), qui devient :

$$\frac{\partial C_{\mathsf{port.\ ref.}}}{\partial y_{it}} = \frac{\partial C}{\partial y_{it}} + \frac{\phi_{\mathsf{port.\ ref.}}}{T}(y_{it} - \psi_{it}). \tag{3.69}$$

Chapitre 4

Cadre expérimental

Ce chapitre décrit les principes et procédures d'entraînement de MLP utilisés au cours des expériences. Après une brève revue des fonctions de coût employées, nous expliquons les méthodes de contrôle de la capacité ainsi que de combinaison de modèles auxquelles nous faisons appel. Nous dressons ensuite un menu expérimental, lequel sera servi au chapitre suivant. Nous terminons par une description des ensembles de données servant aux expériences, ainsi que des formes de prétraitement appliquées.

4.1 Sommaire des deux paradigmes

Cette section présente un résumé des deux paradigmes d'allocation d'actifs retenus pour ce mémoire, et explicite la topologie des MLP utilisés pour les expériences, les processus de prise de décision, ainsi que les fonctions de coût utilisées pour l'entraînement.

4.1.1 Topologie

Tous les systèmes adaptatifs que nous utilisons sont des perceptrons multi-couches (MLP), tels que décrits à la section 3.1. Nous retenons la topologie standard, qui comporte une couche cachée non-linéaire (Tanh), avec un nombre variable d'unités cachées, et une couche de sortie linéaire. Nous dénotons la fonction calculée par un MLP de cette topologie, étant donné un vecteur de paramètres $\boldsymbol{\theta}$, par $f_{\mathcal{M}}(\cdot;\boldsymbol{\theta})$.

Nous incorporons de plus dans la topologie $\mathcal{M}$ d'un réseau la spécification des éléments suivants :

- Nombre d'unités cachées (i.e. la dimensionalité de la couche cachée).
- Les valeurs des hyperparamètres ϕ_{ID} et ϕ_{ID} (tels que décrits à la section 4.3) utilisés dans la fonction de coût servant à entraîner le MLP.

Les entrées du MLP—les variables explicatives au temps t—sont données par le vecteur $\mathbf{u}_t$. Ce vecteur peut être calculé étant donné l'information $\mathcal{I}_t$. (Le détail des variables utilisées pour les expériences est fourni à la section 4.6).

Le MLP comporte N sorties pour un problème d'allocation entre N actifs ; la signification de ces sorties dépend du rôle du MLP dans le système :

- Pour le **modèle de décision** (section 3.4), les sorties représent directement les recommandations d'allocation à prendre au temps t.
- Pour le **modèle de prévision** (section 3.3), les sorties représentent la prévision faite au temps t du rendement de chaque actif pour la période $t+1$.

4.1.2 Prise de décisions

Les étapes suivies par les deux paradigmes pour prendre, à chaque période t, une décision d'allocation $\mathbf{x}_t$ sont résumées au tableau 4.1.

Remarque Il est parfois nécessaire de noter explicitement la dépendance d'une recommandation ou d'une décision par rapport au vecteur de paramètres

du MLP utilisé ; nous indiquons une telle dépendance par $\mathbf{y}_t^S(\boldsymbol{\theta})$ ou $\mathbf{x}_t^S(\boldsymbol{\theta})$ selon le cas, où $S \in \{\mathsf{D}, \mathsf{P}\}$.

4.1.3 Entraînement

Nous résumons la forme finale des fonctions de coût utilisées pour l'entraînement des MLP (introduites au chapitre précédent), en les présentant dans le contexte d'ensembles d'entraînement spécifiques au problème d'allocation d'actifs.

Il est important de distinguer les *critères d'optimisation* des réseaux de neurones, que nous présentons maintenant, du *critère d'évaluation* de la performance du modèle. Ce dernier critère demeure, en toutes circonstances, le rendement financier normalisé par la valeur à risque, éq. (2.44). C'est lui qui permet, en analyse finale, de comparer la performance de différents systèmes. Nous fournissons les détails d'une méthodologie non-biaisée d'évaluation de la performance à la section 4.2.

Définitions préliminaires

Soit un ensemble de variables explicatives $\mathcal{U} = \{\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{T-1}\}$ et un ensemble de rendements réalisés $\mathcal{R} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_T\}$, où $\mathbf{u}_t$ et $\mathbf{r}_t$ sont calculables étant donnée l'information $\mathcal{I}_t$.

Définition 4.1 *Une* **séquence temporelle** S *entre* t_1 *et* t_2, $0 \leq t_1 < t_2 \leq T-1$, *est l'ensemble de tous les indices temporels compris entre* t_1 *et* t_2 *(inclusivement) :*

$$S = \{t : t_1 \leq t \leq t_2,\ t \text{ entier}\}. \tag{4.7}$$

Définition 4.2 *Un* **ensemble d'entraînement** $\mathcal{D}$ *défini par rapport à la séquence temporelle* $S_\mathcal{D}$ *est l'ensemble des paires entrées–sorties associées à la séquence :*

$$\mathcal{D} = \{\langle \mathbf{u}_j, \mathbf{r}_{j+1} \rangle : j \in S_\mathcal{D}\}. \tag{4.8}$$

Définition 4.3 *Un MLP de topologie* $\mathcal{M}$ *est dit* **entraîné** *sur un ensemble* $\mathcal{D}$ *par rapport à une fonction de coût* $\hat{C}_\mathcal{D}(\boldsymbol{\theta}')$ *si ses paramètres* $\boldsymbol{\theta}_\mathcal{D}$ *sont choisis*

Tableau 4.1: *Résumé du processus de prise de décisions*

Modèle de prévision	**Modèle de décision**
1 Calculer, à l'aide du MLP, l'estimateur des rendements des actifs à la période $t+1$, étant données les variables explicatives $\mathbf{u}_t$: $$\hat{\boldsymbol{\mu}}_{t+1\vert t} = f_{\mathcal{M}}(\mathbf{u}_t; \boldsymbol{\theta})$$	1 Calculer, à l'aide du MLP, une recommandation d'allocation $\mathbf{y}_t^{\mathsf{D}}$, étant donnée la recommandation au temps précédent et les variables explicatives $\mathbf{u}_t$: $$\mathbf{y}_t^{\mathsf{D}} = f_{\mathcal{M}}(\mathbf{u}_t, \mathbf{y}_{t-1}^{\mathsf{D}}; \boldsymbol{\theta}) \quad (4.1)$$
2 Calculer la répartition « optimale » des actifs, $\hat{\mathbf{w}}_t^*$, étant donnés $\hat{\boldsymbol{\mu}}_{t+1\vert t}$ (calculé à l'étape précédente), $\hat{\boldsymbol{\Gamma}}_{t+1\vert t}$ (calculé indépendamment par un estimateur pondéré exponentiellement), et le paramètre d'aversion au risque de l'investisseur λ (*cf.* éq. (3.42)) ; cette répartition est aussi la « recommandation d'allocation » $\mathbf{y}_t^{\mathsf{P}}$ (*cf.* section 2.3) : $$\mathbf{y}_t^{\mathsf{P}} = \frac{1}{\lambda} \hat{\boldsymbol{\Gamma}}_{t+1\vert t}^{-1} \left(\hat{\boldsymbol{\mu}}_{t+1\vert t} - \frac{\boldsymbol{\iota}' \hat{\boldsymbol{\Gamma}}_{t+1\vert t}^{-1} \hat{\boldsymbol{\mu}}_{t+1\vert t} - \lambda}{\boldsymbol{\iota}' \hat{\boldsymbol{\Gamma}}_{t+1\vert t}^{-1} \boldsymbol{\iota}} \boldsymbol{\iota} \right) \quad (4.2)$$	
3 Rééchelonner la recommandation $\mathbf{y}_t^{\mathsf{P}}$ pour obtenir un portefeuille ayant la VàR cible $\tilde{V}_{t+1}$ avec probabilité α. Ce portefeuille est noté $\mathbf{x}_t^{\mathsf{P}}$: $$\mathbf{x}_t^{\mathsf{P}} = \hat{\beta}_t^{\mathsf{P}} \mathbf{y}_t^{\mathsf{P}} \quad (4.3)$$ $$\hat{\beta}_t^{\mathsf{P}} = \frac{\tilde{V}_{t+1}}{\Phi^{-1}(\alpha) \sqrt{\mathbf{y}_t^{\mathsf{P}\prime} \hat{\boldsymbol{\Gamma}}_{t+1\vert t} \mathbf{y}_t^{\mathsf{P}}}} \quad (4.4)$$	2 Rééchelonner la recommandation $\mathbf{y}_t^{\mathsf{D}}$ pour obtenir un portefeuille ayant la VàR cible $\tilde{V}_{t+1}$ avec probabilité α. Ce portefeuille est noté $\mathbf{x}_t^{\mathsf{D}}$: $$\mathbf{x}_t^{\mathsf{D}} = \hat{\beta}_t^{\mathsf{D}} \mathbf{y}_t^{\mathsf{D}} \quad (4.5)$$ $$\hat{\beta}_t^{\mathsf{D}} = \frac{\tilde{V}_{t+1}}{\Phi^{-1}(\alpha) \sqrt{\mathbf{y}_t^{\mathsf{D}\prime} \hat{\boldsymbol{\Gamma}}_{t+1\vert t} \mathbf{y}_t^{\mathsf{D}}}} \quad (4.6)$$

comme

$$\boldsymbol{\theta}_{\mathcal{D}} = \arg\min_{\boldsymbol{\theta}'} \hat{C}_{\mathcal{D}}(\boldsymbol{\theta}').$$

La fonction calculée par un tel réseau est dénotée $f_{\mathcal{M}}(\cdot; \boldsymbol{\theta}_{\mathcal{D}})$.

En pratique, il n'est possible que de trouver un minimum local de la fonction de coût ; nous qualifions néanmoins toujours d'« entraînement » le processus du choix de $\boldsymbol{\theta}$.

Modèle de prévision

Soit un ensemble d'entraînement $\mathcal{D}$, défini par rapport à la séquence temporelle $S_{\mathcal{D}}$. Le critère d'entraînement du MLP de topologie $\mathcal{M}$ et de paramètres $\boldsymbol{\theta}$ sous-tendant le modèle de prévision est l'erreur quadratique de prévision des rendements :

$$\hat{C}^{\mathsf{P}}_{\mathcal{D}}(\boldsymbol{\theta}) = \sum_{t \in S_{\mathcal{D}}} \| f_{\mathcal{M}}(\mathbf{u}_t; \boldsymbol{\theta}) - \mathbf{r}_{t+1} \|^2 + C_{\mathsf{WD}}(\boldsymbol{\theta}) + C_{\mathsf{ID}}(\boldsymbol{\theta}) \tag{4.9}$$

Les termes C_{WD} et C_{ID} (pour *weight decay* et *input decay* respectivement) sont des termes de régularisation destinés à réfléter des préférences à priori sur le vecteur de paramètres $\boldsymbol{\theta}$ du MLP. La nature de ces termes est explicitée à la section 4.3.

Modèle de décision

Conservant les mêmes définitions que précédemment, le critère servant à entraîner le MLP sous-tendant le modèle de décision est le critère de perfor-

mance financière $\hat{W}$ régularisé par un portefeuille de référence :

$$\hat{C}^{\mathsf{D}}_{\mathcal{D}}(\boldsymbol{\theta}) = -\frac{1}{|S_{\mathcal{D}}|}\sum_{t\in S_{\mathcal{D}}} \hat{W}^{\mathsf{D}}_{t+1} + C_{\mathsf{port.\,ref.}} + C_{\mathsf{WD}}(\boldsymbol{\theta}) + C_{\mathsf{ID}}(\boldsymbol{\theta}) \quad (4.10)$$

$$\hat{W}^{\mathsf{D}}_{t+1} = \frac{(\mathbf{r}_{t+1} - \boldsymbol{\iota}\, r_{0t})'\mathbf{x}^{\mathsf{D}}_t(\boldsymbol{\theta}) + \mathsf{perte}_t}{\tilde{V}_{t+1}} \quad (4.11)$$

$$C_{\mathsf{port.\,ref.}} = \frac{1}{|S_{\mathcal{D}}|}\sum_{t\in S_{\mathcal{D}}} \mathsf{pén.}_{\mathsf{port.\,ref.}}(\mathbf{y}^{\mathsf{D}}_t) \quad (4.12)$$

$$\mathsf{pén.}_{\mathsf{port.\,ref.}}(\mathbf{y}_t) = \frac{\phi_{\mathsf{port.\,ref.}}}{2}\left\|\mathbf{y}_t - \boldsymbol{\psi}_t\right\|^2, \quad (4.13)$$

où $\mathbf{x}^{\mathsf{D}}_t$ et $\mathbf{y}^{\mathsf{D}}_t$ sont donnés respectivement par l'éq. (4.5) et (4.1).

Optimisation

La minimisation des fonctions de coût précédentes s'effectue en optimisant les paramètres du réseau à l'aide de l'algorithme du *gradient conjugué* (Fletcher 1987; Press, Flannery, Teukolsky et Vetterling 1992; Bishop 1995). Pour chaque ronde d'entraînement nous effectuons cinq reprises aléatoires à partir de points de départ différents, ce qui permet de réduire l'effet indésirable des minima locaux qui accablent parfois le gradient conjugué. Nous n'employons pas « d'arrêt prématuré » (*early stopping*) après un certain nombre d'époques d'entraînement ; nous cherchons le meilleur minimum de la fonction de coût dans tous les cas, et nous fions à d'autres techniques pour contrôler la capacité (voir sections 4.3 et 4.4).

4.2 Estimation de la performance

Définition 4.4 *Un* **ensemble de test** $\mathcal{T}$ *associé à l'ensemble d'entraînement* $\mathcal{D}$ *ayant une séquence temporelle* $S_{\mathcal{D}}$ *est un ensemble de paires n'appartenant pas à* $\mathcal{D}$ *:*

$$\mathcal{T} = \{\langle \mathbf{u}_j, \mathbf{r}_{j+1}\rangle : 0 \le j \le T-1, j \notin S_{\mathcal{D}}, j+1 \notin S_{\mathcal{D}}\}. \quad (4.14)$$

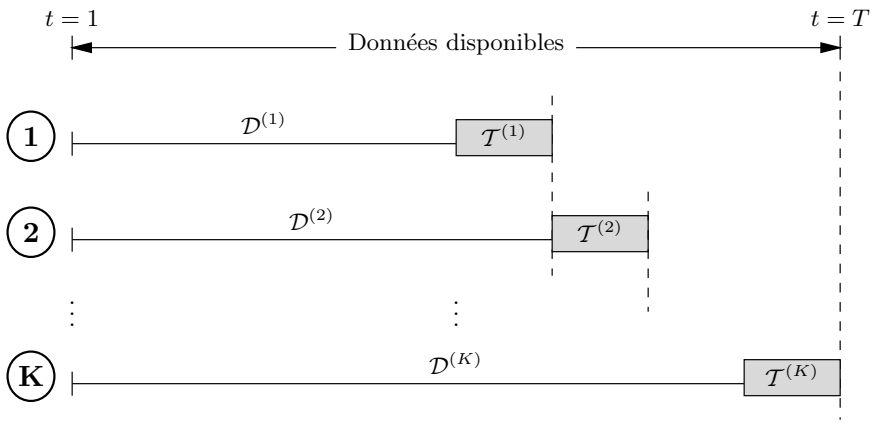


Figure 4.1: *Illustration de l'entraînement et de l'évaluation de la performance par* validation séquentielle.

(Notons que l'ensemble $\mathcal{T}$ n'est pas nécessairement l'ensemble de toutes les paires n'appartenant pas à $\mathcal{D}$).

Définition 4.5 *Une séquence d'ensembles d'entraînement* $\{\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K)}\}$ *est* **emboîtée** *si et seulement si* $\mathcal{D}^{(k)} \subset \mathcal{D}^{(k+1)}$, $1 \leq k \leq K-1$.

À chaque ensemble d'entraînement $\mathcal{D}^{(k)}$ d'une séquence emboîtée nous associons un ensemble de test $\mathcal{T}^{(k)}$. Il découle de la définition d'un ensemble de test que $\mathcal{T}^{(k)} \cap \mathcal{D}^{(j)} = \emptyset, \forall j \leq k$.

La figure 4.1 illustre une configuration particulière possible de ces ensembles, pour laquelle chaque ensemble de test suit immédiatement (dans le temps) l'ensemble d'entraînement qui lui est associé, et est d'une taille fixe P. De plus, l'ensemble $\mathcal{D}^{(k+1)}$ « absorbe » les éléments de l'ensemble de test précédent :

$$\mathcal{D}^{(k+1)} = \mathcal{D}^{(k)} \cup \mathcal{T}^{(k)}.$$

4.2.1 Validation séquentielle

Soit $\{\mathcal{D}^{(k)}\}$ une séquence emboîtée d'ensembles d'entraînement, et $\{\mathcal{T}^{(k)}\}$ une séquence associée d'ensembles de test. Nous choisissons une « stratégie » $S \in \{\mathsf{D}, \mathsf{P}\}$ (modèle de décision ou de prévision) à tester. Nous fixons de plus

la topologie $\mathcal{M}$ du MLP.

La validation séquentielle procède comme suit : pour chaque ensemble $\mathcal{D}^{(k)}$, nous entraînons un MLP de topologie $\mathcal{M}$ sur cet ensemble par rapport au critère $\hat{C}^S_{\mathcal{D}^{(k)}}(\boldsymbol{\theta})$ (ces critères sont résumés à la section précédente), et obtenons un vecteur de paramètres résultant de l'entraînement, $\boldsymbol{\theta}_{\mathcal{D}^{(k)}}$.

Afin d'uniformiser la notation des décisions prises par les modèles de décision ou de prévision, nous dénotons par $\mathbf{x}^S_t(\boldsymbol{\theta})$ les décisions issues de la stratégie S à tester utilisant $\boldsymbol{\theta}$ comme vecteur de paramètres du MLP (en d'autres termes, la méthodologie de test ne diffère pas selon qu'on évalue le modèle de décision ou celui de prévision).

Nous obtenons un estimateur de la performance du système qui utilise le k-ième MLP en calculant le critère de performance financière $\hat{W}$ (éq. (2.44) et (2.45)) sur l'ensemble de test $\mathcal{T}^{(k)}$ ayant pour séquence temporelle $S_{\mathcal{T}^{(k)}}$:

$$\hat{W}^{(k)}_{\mathcal{M}} = \frac{1}{|S_{\mathcal{T}^{(k)}}|} \sum_{t \in S_{\mathcal{T}^{(k)}}} \hat{W}^S_{\mathcal{M}}(t+1; \boldsymbol{\theta}_{\mathcal{D}^{(k)}}) \tag{4.15}$$

$$\hat{W}^S_{\mathcal{M}}(t+1; \boldsymbol{\theta}_{\mathcal{D}^{(k)}}) = \frac{(\mathbf{r}_{t+1} - \boldsymbol{\iota}\, r_{0t})' \, \mathbf{x}^S_t(\boldsymbol{\theta}_{\mathcal{D}^{(k)}}) + \mathsf{perte}_t}{\tilde{V}_{t+1}} \tag{4.16}$$

Sous l'hypothèse de stationarité de la distribution conditionnelle de $\mathbf{r}_t$ étant donné $\mathcal{I}_{t-1}$, l'estimateur $\hat{W}^{(k)}_{\mathcal{M}}$ en est un non-biaisé de la performance financière du système utilisant le k-ième MLP, car il est calculé à partir d'observations n'ayant pas servi à l'entraînement (voir section 3.1.3).

Nous pouvons obtenir un estimateur de la performance financière moyenne d'un système utilisant un MLP de topologie $\mathcal{M}$ en calculant la moyenne des estimateurs $\hat{W}^{(k)}_{\mathcal{M}}$ individuels :

$$\hat{W}_{\mathcal{M}} = \frac{1}{K} \sum_{k=1}^{K} \hat{W}^{(k)}_{\mathcal{M}}. \tag{4.17}$$

Cet estimateur est non-biaisé (sous la même hypothèse de stationarité que précédemment) car il est une simple combinaison linéaire d'estimateurs eux-mêmes non-biaisés.

4.2.2 Pourquoi une validation *séquentielle* ?

La validation séquentielle tire son nom de l'emboîtement des ensembles d'entraînement, et du fait que dans son application pratique (voir figure 4.1) les ensembles de test suivent immédiatement ceux d'entraînement.

Cette procédure est dérivée de la *validation croisée* bien connue en statistiques (Stone 1974). La validation croisée est utilisée dans le domaine des algorithmes d'apprentissage pour obtenir un estimateur non-biaisé de la performance de généralisation dans le cas ou les données disponibles pour l'entraînement et le test sont véritablement i.i.d. (et où ces données peuvent être librement interchangées).

La validation séquentielle s'applique, au contraire, à des données pour lesquelles l'ordonnancement est important, comme celles dotées d'une structure temporelle. Pour ces données, l'hypothèse d'indépendance que pose la validation croisée peut s'avérer trop forte, mais cette hypothèse n'est pas nécessaire au bon fonctionnement de la validation séquentielle. La procédure, dans les faits, pose à répétition la question suivante :

> *Étant donné tout le passé, quelle est la performance attendue d'un système de décision dans un proche*[1] *avenir ?*

La moyenne de ces réponses constitue l'estimateur de performance calculé par la validation séquentielle.

4.2.3 Taille minimale d'entraînement

La description précédente de la validation séquentielle laisse en suspens deux paramètres importants gouvernant la procédure : quelle devrait être la taille P de chaque ensemble de test (et de la progression entre les ensembles d'entraînement emboîtés), et quelle est la taille du *premier* ensemble d'entraînement.

La réponse à la première question est simple : pour des séries financières (qui, par conséquent, comportent vraisemblablement des non-stationarités),

[1] « *Proche* » *étant donné par la taille P des ensembles de test.*

P doit être choisi le plus petit possible, en fonction des capacités de calcul disponibles. Si le changement dans la distribution des données est progressif, l'emploi d'une petite valeur de P assure que le MLP n'est jamais testé sur des données qui deviennent très différentes de celles utilisées pour l'entraînement.[2] Dans presque toutes nos expériences, nous utilisons $P = 12$, ce qui correspond à ré-entraîner le MLP à toutes les années pour des données mensuelles.

La réponse à seconde question, la taille du premier segment ou la taille minimale d'entraînement, est plus délicate : deux facteurs contradictoires entrent en considération. D'une part, l'utilisation de trop peu de *données d'entraînement* conduit à une moins bonne performance en généralisation (biais dans la solution). D'autre part, l'emploi de trop peu de *données de test* ne fournit pas d'estimé fiable de l'erreur (variance dans l'estimateur de performance).

Ce problème est illustré de façon probante à la figure 4.2. Le problème considéré est l'allocation entre trois actifs : un indice boursier, un indice obligataire, et un actif sans risque. Le modèle utilisé est un modèle constant, d'une extrême simplicité : nous cherchons à trouver la pondération constante du portefeuille qui maximise le rendement (normalisé par la VàR) sur l'ensemble de test. Pour la période considérée, la solution optimale est d'allouer 100% à l'indice boursier.

La figure montre la performance obtenue par le modèle (qui apprend les constantes), celle d'une stratégie qui investit tout le capital dans le marché boursier, et la différence entre les deux.[3] La variable indépendante est la fraction des données réservée à *l'ensemble de test*.

On constate que lorsque trop peu de données sont réservées au test (de l'ordre de 10 à 20%), les estimés de la performance sont hautement variables, tel que démontré par l'amplitude des barres d'erreur. A contrario, lorsqu'une

[2]Évidemment, en présence de non-stationnarités, l'ensemble d'entraînement doit être revu avec discernement ; il peut être nécessaire de pondérer les données pour attribuer plus d'importance aux données récentes, et en contrepartie, moins aux données anciennes. Bengio et Dugas (1999) discute d'un moyen automatique de trouver de bons paramètres de pondération.

[3]Nous avons utilisé un t-test apparié (*paired t-test*) pour calculer les intervalles de confiance sur les différences.

trop grande quantité de données est utilisée pour le test (donc trop peu pour l'entraînement), la performance du modèle décroît par rapport au marché. Finalement, pour un bon équilibre entre les données d'entraînement et de test, le modèle appris dans cet exemple converge exactement vers celui qui pondère l'indice boursier à 100%, comme en font foi les différences de zéro (et les intervalle de confiance pratiquement invisibles sur ces différences qui sont représentées par des marqueurs triangulaires sur la figure) pour la région entre 30% et 60%.

Bien qu'il soit nécessaire de se garder d'une interprétation trop stricte de ces résultats–jouets, il nous a paru raisonnable d'appliquer la procédure de validation séquentielle décrite précédemment en utilisant 40% de la séquence disponible comme taille minimale du premier entraînement.

4.2.4 Ensemble de validation

La procédure usuelle d'entraînement d'un réseau de neurones consiste à réserver une partie des données—un *ensemble de validation*—distinctes des ensembles d'entraînement et de test. Cet ensemble permet de contrôler la capacité du réseau, en choisissant, par exemple un bon point pour stopper l'entraînement (arrêt prématuré) (Bishop 1995; Prechelt 1998). Dans la procédure de validation séquentielle, l'ensemble de validation pour chaque segment est naturellement situé immédiatement avant chaque ensemble de test.

Cependant, la nature des données financières utilisées limitent l'utilité pratique d'un tel ensemble de validation. Premièrement, l'échantillonnage des données est effectué à une relativement basse fréquence (sur une base mensuelle) ; par conséquent, pour éviter de biaiser l'entraînement, nous devons nous contenter d'un tout petit ensemble de validation, mais comme nous l'avons vu précédemment, les mesures prises sur ce petit ensemble seront très bruitées. Deuxièmement, les séries financières sont souvent non-stationnaires ; si l'ensemble de test est séparé de celui d'entraînement par un ensemble de validation significatif, nous courons le risque de soumettre au test un réseau entraîné sur des données provenant d'une distribution significativement

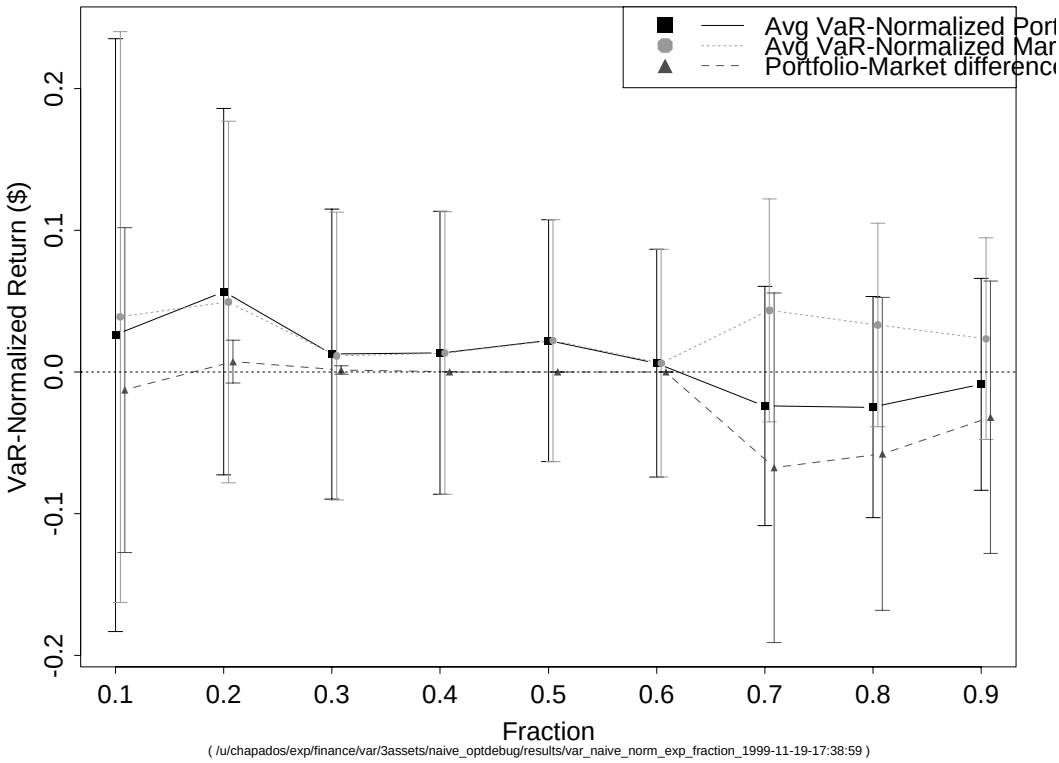


Figure 4.2: *Performance d'un prédicteur naïf sur une tâche d'allocation de trois actifs, en fonction de la fraction des données utilisées pour l'ensemble de test (N total = 320). La ligne continue représente le modèle appris ; le pointillé est le modèle de marché boursier ; le hachuré est la différence entre les deux. Les barres d'erreur représentent des intervalles à 95%.*

différente.

Pour ces deux raisons, nous ne faisons pas appel à un ensemble de validation au cours de nos expériences.

4.3 Contrôle de la capacité

Puisque nous n'utilisons pas d'ensemble de validation séparé de l'ensemble de test pour contrôler l'arrêt prématuré, il est nécessaire d'employer d'autres méthodes de contrôle de la capacité des réseaux de neurones. Autrement, les réseaux courent le risque de « sur-apprendre » les données d'entraînement, et de se comporter de piètre façon en généralisation. Les méthodes que nous avons choisies à cet effet, la *pénalisation sur la norme des poids*[4] et la *pénalisation sur la norme des entrées*[5] introduisent des termes de régularisation dans la fonction de coût à optimiser par le réseau.

4.3.1 Pénalisation sur la norme des poids

La pénalisation sur la norme des poids est une méthode classique (Hinton 1987; Bishop 1995) qui consiste à imposer une pénalité à tous les poids du réseau pour pénaliser les poids trop gros. Le terme suivant est ajouté à la fonction de coût :

$$C_{\mathsf{WD}}(\boldsymbol{\theta}) = \frac{\phi_{\mathsf{WD}}}{2} \sum_k \boldsymbol{\theta}_k^2, \tag{4.18}$$

où la somme se fait sur tous les poids du réseau à l'exception des biais (les vecteurs $\mathbf{b}_i$ des éq. (3.2) et (3.3)), et ϕ_{WD} est un hyperparamètre (à déterminer expérimentalement) qui contrôle l'importance relative de C_{WD} dans le coût total.

L'effet de cette pénalisation est d'encourager les poids du réseau à devenir plus petits en magnitude, et elle réduit la capacité du réseau. Empiriquement,

[4] En anglais, *weight decay*. Nous utiliserons parfois l'acronyme *WD*.

[5] En anglais, *input decay*, bien que cette terminologie ne soit pas standard. Nous utiliserons parfois l'acronyme *ID*.

elle conduit souvent à une meilleure généralisation quand le nombre d'exemples est « petit » (quelques centaines) (Hinton 1987). Son défaut est de ne pas tenir compte de la fonction à apprendre : cette pénalisation s'applique sans discrimination à tous les poids, qu'elle cherche à réduire.

4.3.2 Pénalisation sur la norme des entrées

La pénalisation sur la norme des entrées est une méthode « douce » de sélection de variables. Au contraire des méthodes combinatoires comme les procédures de *branch and bound* (Narendra et Fukunaga 1977) et la sélection avant ou arrière (Bishop 1995), le modélisateur ne cherche pas un « bon » sous-ensemble des entrées à fournir au réseau,— il les fournit toutes ; la méthode pénalisera automatiquement les connections dans le réseau provenant des entrées qui ne s'avèrent pas importantes.

La pénalisation sur la norme des entrées est similaire à celle sur la norme des poids présentée ci-haut en ce qu'elle ajoute un terme à la fonction de coût. Elle pénalise les poids qui relient une entrée particulière du réseau à toutes les unités cachées, en agissant comme suit : soit $\boldsymbol{\theta}_{jh}^{(1)}$ le poids (situé sur la première couche du MLP) unissant l'entrée j à l'unité cachée h, la pénalisation attribuée à l'entrée j est :

$$C_{\mathsf{ID}}^{(j)}(\boldsymbol{\theta}) = \sum_{h=1}^{H} \left(\boldsymbol{\theta}_{jh}^{(1)}\right)^2, \tag{4.19}$$

où H est le nombre d'unités dans la première couche cachée. La figure 4.3 illustre les poids qui font partie de $C_{\mathsf{ID}}^{(j)}(\boldsymbol{\theta})$.

Pour déterminer la contribution complète $C_{\mathsf{ID}}(\boldsymbol{\theta})$ à la fonction de coût, nous faisons la somme des contributions de toutes les entrées j, comme suit :

$$C_{\mathsf{ID}}(\boldsymbol{\theta}) = \phi_{\mathsf{ID}} \sum_{j} \frac{C_{\mathsf{ID}}^{(j)}}{\eta + C_{\mathsf{ID}}^{(j)}(\boldsymbol{\theta})}, \tag{4.20}$$

et ϕ_{ID}, déterminé expérimentalement, contrôle l'importance relative de la pénalisation sur la norme des poids.

La figure 4.4 illustre l'allure de la fonction $x^2/(\eta + x^2)$, qui module la

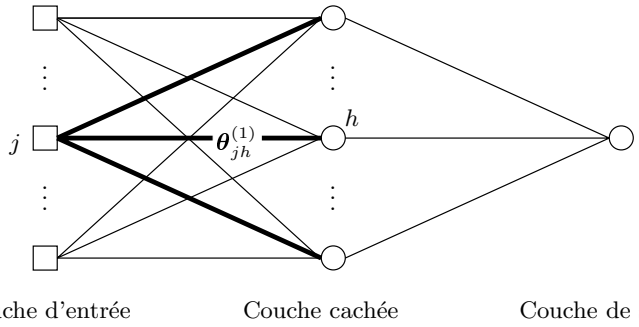


Figure 4.3: *Schéma des connections affectées par la pénalisation sur la norme des entrées pour une entrée j dans un perceptron multi–couches.*

contribution de chaque entrée. On constate que pour de petites valeurs de la norme des poids, la fonction se comporte de manière quadratique, mais elle sature rapidement dès que cette norme dépasse un seuil déterminé par le paramètre η.

Nous pouvons comprendre l'effet de cette fonction comme suit : initialement lors de l'entraînement, lorsque les poids émergeant d'une entrée j sont encore petits, le réseau doit subir un coût marginal élevé pour en accroître la grandeur ; si l'entrée ne s'avère utile que pour peu d'unités cachées, il est préférable de laisser ces poids petits, et donc d'éliminer l'impact pratique de cette entrée sur le réseau. En contrepartie, dès le moment où l'entrée prouve son utilité (avec des poids suffisamment grands), la contribution de ces poids à la fonction de coût devient constante, indépendamment de leurs valeurs.

Cette méthode est similaire à celle de *weight elimination* proposée par Weigend, Rumelhart et Huberman (1991) en ce qu'elle utilise une fonction de coût de forme similaire à l'éq. (4.20). Cependant, le *weight elimination* n'est pas habituellement employé pour faire de la sélection de variables ; il s'applique à tous les poids du réseau.

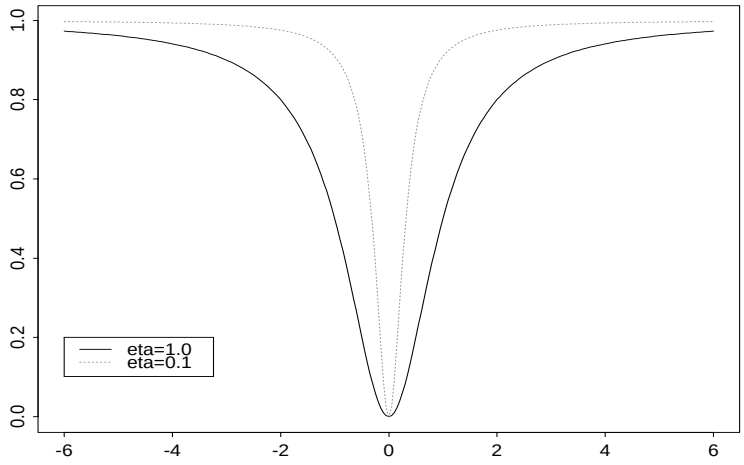


Figure 4.4: *Fonction de coût $x^2/(\eta + x^2)$ appliquée par la pénalisation sur la norme des entrées à la norme des poids reliant une entrée donnée du réseau à toutes les unités cachées. Deux valeurs du paramètre de seuil η sont illustrées.*

4.4 Combinaisons de modèles

Les méthodes de contrôle de la capacité présentées plus haut n'expliquent pas comment choisir les hyperparamètres ϕ_{WD} et ϕ_{ID}. Le choix de ces paramètres est important, car il détermine en bonne partie la capacité d'apprentissage d'un MLP, et donc sa performance de généralisation. De plus, leurs valeurs optimales dépendent intimement de la fonction de coût considérée ainsi que de la topologie du MLP ; pour la majorité des problèmes, nous ne connaissons pas à priori quelles valeurs des paramètres seront les meilleures.

La technique habituelle du choix des hyperparamètres consiste à effectuer une recherche plus ou moins exhaustive de l'espace, et de choisir le modèle qui produit la meilleure performance sur un ensemble de validation. Des méthodes plus récentes consistent à optimiser les hyperparamètres comme s'il s'agissait de paramètres ordinaires en effectuant une descente de gradient par rapport à l'erreur de validation (Larsen, Svarer, Andersen et Hansen 1998; Bengio et Dugas 1999).

Malheureusement, toutes ces méthodes supposent l'utilisation d'un en-

semble de validation spécifique, distinct de l'ensemble de test, si on souhaite conserver la possibilité d'utiliser l'ensemble de test pour obtenir un estimé non-biaisé de l'erreur de généralisation. Or, à cause des raisons du peu de données et de non-stationnarité expliquées plus haut, nous désirons éliminer complètement la nécessité d'un ensemble de validation. La question demeure : sans ensemble de validation, comment choisir de bonnes valeurs pour les hyperparamètres ? La réponse triviale est de comparer les modèles sur l'ensemble de test et de retenir le meilleur. Toutefois, ce faisant, le modèle que nous choisissons ne constitue plus un estimé non-biaisé de la performance de généralisation, car il est choisi en fonction des spécificités de l'ensemble de test. Il est donc exclus de choisir entre plusieurs modèles de cette façon.

La solution que nous employons est de *ne pas avoir à choisir* entre plusieurs modèles concurrents, mais de les *combiner* de différentes manières pour former un « super-modèle » (que nous appelons *comité*) dont nous évaluerons la performance sur l'ensemble de test.

4.4.1 Comité : définition

Nous considérons un ensemble de G fonctions *sous-jacentes*, $\mathcal{G} = \{g_j : \mathbb{R}^M \mapsto \mathbb{R}^N\}_{j=1}^G$. Chaque fonction peut être issue, par exemple, d'un MLP avec sa topologie et son vecteur de paramètres propres (lesquels nous considérons implicitement partie de g_j).

Définition 4.6 *Un* **comité** *défini par rapport à un ensemble de fonctions $\mathcal{G}$ est une fonction $g^{\mathsf{com}} : \mathbb{R}^M \mapsto \mathbb{R}^N$ résultant d'une combinaison convexe des éléments de $\mathcal{G}$,*

$$g^{\mathsf{com}}(\mathbf{x}) = \sum_{j=1}^{G} w_j \, g_j(\mathbf{x}), \tag{4.21}$$

avec $w_j \geq 0, \forall j$, et $\sum_j w_j = 1$.

Nous précisons plus loin l'application de cette définition au contexte de gestion de portefeuille étudié dans ce mémoire ; auparavant, nous dérivons quelques propriétes des comités.

4.4.2 Performance de généralisation d'un comité

Nous démontrons dans cette section une décomposition en deux parties de l'erreur de généralisation d'un comité : une première attribuable aux erreurs des membres individuels du comité, et une seconde résultant de la discordance ou de l'*ambiguité* entre les membres. Cette décomposition est similaire à celle classique entre le biais et la variance d'un algorithme d'apprentissage (German, Bienenstock et Doursat 1992). La présente dérivation est inspirée de (Krogh et Vedelsby 1995).

Par simplicité, dans cette section, nous considérons uniquement des fonctions réelles $g_i : \mathbb{R} \mapsto \mathbb{R}$; la généralisation en plusieurs dimensions s'effectue naturellement.

Définitions

Nous nous limitons au problème de la régression ; nous dénotons la « véritable » fonction à apprendre par $h(x)$. Chacune des fonctions $g_j(x)$ peut s'écrire comme la somme de $h(x)$ et d'une composante d'erreur :

$$g_j(x) = h(x) + \epsilon_j(x). \tag{4.22}$$

L'erreur quadratique d'un membre du comité est simplement

$$\epsilon_j^2(x) = (g_j(x) - h(x))^2, \tag{4.23}$$

et l'erreur moyenne des membres du comité est

$$\begin{aligned} \bar{\epsilon}^2(x) &= \sum_j w_j \, \epsilon_j^2(x) \\ &= \sum_j w_j (g_j(x) - h(x))^2. \end{aligned} \tag{4.24}$$

L'erreur du comité entier est :

$$e(x) = (g^{\mathsf{com}}(x) - h(x))^2 \tag{4.25}$$

Nous définissons l'**ambiguité** $a_j(x)$ d'un membre du comité comme étant l'écart quadratique entre ce membre et le comité :

$$a_j(x) = (g_j(x) - g^{\mathsf{com}}(x))^2. \tag{4.26}$$

L'ambiguité de l'ensemble, $\bar{a}(x)$, est simplement la moyenne pondérée des ambiguités individuelles, que nous pouvons voir comme la variance dans les décisions rendues par les membres du comité :

$$\begin{aligned} \bar{a}(x) &= \sum_j w_j \, a_j(x) \\ &= \sum_j w_j \, (g_j(x) - g^{\mathsf{com}}(x))^2 \end{aligned} \tag{4.27}$$

Relation entre l'ambiguité et l'erreur

Additionnant et soustrayant $h(x)$ dans l'éq. (4.27), nous obtenons :

$$\begin{aligned} \bar{a}(x) &= \sum_j w_j \, ((g_j(x) - h(x)) + (h(x) - g^{\mathsf{com}}(x)))^2 \\ &= \sum_j w_j (g_j(x) - h(x))^2 \quad - \quad (h(x) - g^{\mathsf{com}}(x))^2 \\ &\qquad + \quad 2 \, (h(x) - g^{\mathsf{com}}(x)) \sum_j w_j \, (g_j(x) - h(x)) \\ &= \sum_j w_j (g_j(x) - h(x))^2 - e(x) \\ &= \bar{\epsilon}^2(x) - e(x) \end{aligned}$$

d'où nous trouvons

$$e(x) = \bar{\epsilon}^2(x) - \bar{a}(x). \tag{4.28}$$

Erreur de généralisation

L'erreur de généralisation du comité est l'espérance de l'erreur quadratique, $E^{\mathsf{com}} = \mathrm{E}[e(X)]$. Prenant l'espérance à gauche et à droite de l'éq. (4.28),

l'erreur de généralisation s'écrit

$$E^{\mathsf{com}} \stackrel{\text{def}}{=} \mathrm{E}[e(X)] \;=\; \mathrm{E}[\bar{\epsilon}^2(X)] - \mathrm{E}[\bar{a}(X)]. \tag{4.29}$$

L'aspect remarquable de cette expression est qu'elle sépare l'erreur de généralisation du comité en un premier terme qui dépend des erreurs espérées des membres individuels, et un second (l'ambiguité espérée) qui incorpore l'effet de *toutes les corrélations* entre les membres du comité. Si les membres du comité rendent des décisions presque identiques, l'erreur du comité sera comparable à l'erreur moyenne de ses membres. En contrepartie, si les membres sont suffisamment différents pour être fréquemment en désaccord, la grande ambiguité espérée causera une *réduction* de l'erreur du comité, pour une erreur moyenne des membres demeurant constante.[6]

Comment trouver de bonnes pondérations

La dérivation précédente laisse sans réponse la question de trouver de bonnes pondérations w_j. La réponse usuelle à ce problème (Krogh et Vedelsby 1995; Bishop 1995), qui ne tient pas compte de la structure temporelle que nous devrons considérer ensuite, est d'écrire l'erreur de généralisation du comité comme

$$\begin{aligned} E^{\mathsf{com}} &= \mathrm{E}[(g^{\mathsf{com}}(X) - h(X))^2] \\ &= \mathrm{E}\left[\left(\sum_j w_j \left(h(X) + \epsilon_j(X)\right) - h(X)\right)^2\right] \\ &= \mathrm{E}\left[\left(\sum_j w_j\, \epsilon_j(X)\right)\left(\sum_k w_k\, \epsilon_k(X)\right)\right], \end{aligned}$$

[6] En pratique, il est fréquent d'observer une augmentation de l'erreur moyenne des membres du comité allant de pair avec une augmentation de l'ambiguité (Krogh et Vedelsby 1995) ; il faut faire preuve de discernement dans l'application des présents résultats.

ce qui peut s'écrire comme

$$E^{\mathsf{com}} = \sum_j \sum_k \mathbf{\Sigma}_{jk}\, w_j\, w_k, \tag{4.30}$$

où $\mathbf{\Sigma}$ est la matrice de covariance entre les erreurs des membres du comité,

$$\mathbf{\Sigma}_{jk} = \mathrm{E}[\epsilon_j(X)\, \epsilon_k(X)]. \tag{4.31}$$

Un ensemble de pondérations $\hat{\mathbf{w}}^* = (\hat{w}_1^*, \ldots, \hat{w}_G^*)'$ peut être déterminé en estimant d'abord la matrice de covariance, et en résolvant le programme quadratique

$$\hat{\mathbf{w}}^* = \min_{\mathbf{w}} \mathbf{w}'\hat{\mathbf{\Sigma}}\mathbf{w}, \tag{4.32}$$

sujet aux contraintes $\mathbf{w} \geq 0$ et $\mathbf{w}'\boldsymbol{\iota} = 1$.

Malheureusement, dans le contexte de gestion de portefeuille que nous étudions dans ce mémoire, cette approche de sélection des pondérations est peu applicable, car le problème n'en est pas un de régression, mais de maximisation d'un critère de performance financière : nous ne connaissons pas, à priori, la « véritable » fonction vers laquelle les membres du comité (et le comité lui-même) doivent tendre.

Nous décrivons nos méthodes de sélection de poids dans la section suivante, après avoir établi le cadre dans lequel nous appliquons les comités à la gestion de portefeuille.

4.4.3 Application au contexte de gestion de portefeuille

Le contexte dans lequel nous appliquons les comités est d'entraîner, pour un type de modèles (prévision ou décision) et un nombre d'unités cachées fixés, un certain nombre de MLPs variant dans leurs hyperparamètres ϕ_{ID} et ϕ_{WD}. Les fonctions g_j qui sont regroupées dans un comité sont celles qui donnent les *recommandations* d'allocation à chaque temps (*cf.* éq. (4.1) et (4.2)).[7]

[7]Pour le modèle de décision, ceci est équivalent à regrouper directement les différents MLP ; cependant, pour le modèle de prévision, nous regroupons en comités les résultats de l'allocation quadratique.

Pour une stratégie $S \in \{\mathsf{D}, \mathsf{P}\}$, la recommandation d'allocation du comité formé d'un ensemble de G fonctions est donnée par :

$$\mathbf{y}_t^{\mathsf{com}} = \sum_{j=1}^{G} \mathbf{y}_t^S(\boldsymbol{\theta}_{j,\mathcal{D}^{(k(t))}}) \tag{4.33}$$

où le vecteur de paramètres du j-ième élément du comité, $\boldsymbol{\theta}_{j,\mathcal{D}^{(k(t))}}$ est choisi comme celui résultant de l'entraînement sur l'ensemble $\mathcal{D}^{(k(t))}$ qui s'est terminé le plus « récemment », c'est-à-dire celui associé à l'ensemble de test qui compte t dans sa séquence temporelle $S_{\mathcal{T}^{(k)}}$,

$$k(t) = \min\ k : t \in S_{\mathcal{T}^{(k)}}. \tag{4.34}$$

La décision d'allocation $\mathbf{x}_t^{\mathsf{com}}$ résulte de la normalisation de $\mathbf{y}_t^{\mathsf{com}}$ par la VàR désirée $\tilde{V}_{t+1}$ (*cf.* éq. (4.3) ou (4.5)).

4.4.4 Sélection des poids

Les trois méthodes que nous considérons pour la sélection des pondérations w_j des membres du comité tentent de s'acquitter de la tâche intuitive suivante : accorder plus d'importance aux membres qui ont « mieux » fonctionné dans le passé. Nous souhaitons aussi que la méthode réagisse sans délai aux changements abrupts dans la performance d'un ou de plusieurs membres du comité.

Nous mettons l'accent sur la nécessité d'évaluer la performance de chaque membre du comité de façon *non-biaisée*. Le moyen que nous employons est de considérer les décisions de généralisation *passées*, c'est-à-dire, pour une pondération qu'on doit choisir au temps t, celles effectuées sur les ensembles de test $\mathcal{T}^{(j)}$, pour $j \leq k(t)$, avec $k(t)$ déterminé par l'éq. (4.34).[8]

I—Hardmax

La méthode la plus simple de combinaison est de choisir, au temps t, le modèle qui a fourni la meilleure performance de généralisation moyenne pour

[8] Il faut faire attention au cas $j = k(t)$ pour ne pas choisir d'exemples dans le futur.

tous les temps précédents. Soit $\hat{W}_j^S(\tau+1;\boldsymbol{\theta}_{j,\mathcal{D}^{(k(t))}})$ le rendement financier dégagé au temps $\tau+1$ par le j-ième membre du comité (*cf.* éq. (4.16)). Soit t_0 le premier instant auquel peut être pris une décision de généralisation, c'est-à-dire le plus petit élément du premier ensemble de test $\mathcal{T}^{(1)}$:

$$t_0 = \min\ S_{\mathcal{T}^{(1)}}.$$

Soit j_t^* le « meilleur modèle » jusqu'au temps $t-1$:

$$j_t^* = \arg\max_j \frac{1}{t-t_0}\sum_{\tau=t_0}^{t-1} \hat{W}_j^S(\tau+1;\boldsymbol{\theta}_{j,\mathcal{D}^{(k(t))}}). \tag{4.35}$$

La pondération accordée au j-ième membre du comité par la procédure de combinaison « hardmax » est :

$$w_{jt}^{\mathsf{hardmax}} = \begin{cases} 1 & \text{si } j = j_t^*, \\ 0 & \text{autrement.} \end{cases} \tag{4.36}$$

On constate aisément que pour une distribution stationnaire des rendements, cette méthode de pondération converge asymptotiquement (et en pratique rapidement) vers un poids de 1 pour le meilleur membre du comité, et de 0 pour les autres.

II—Softmax

La seconde méthode est une légère modification de la précédente. Elle consiste à combiner les performances moyennes obtenues jusqu'à présent à l'aide de la fonction softmax. Soit $\bar{W}_{jt}$ la performance financière moyenne obtenue par le j-ième élément du comité jusqu'au temps $t-1$:

$$\bar{W}_{jt} = \frac{1}{t-t_0}\sum_{\tau=t_0}^{t-1} \hat{W}_j^S(\tau+1;\boldsymbol{\theta}_{j,\mathcal{D}^{(k(t))}}). \tag{4.37}$$

La pondérations accordée au j-ième membre du comité par la procédure

de combinaison « softmax » est :

$$w_{jt}^{\mathsf{softmax}} = \frac{\exp(\bar{W}_{jt})}{\sum_{k=1}^{M} \exp(\bar{W}_{kt})}. \tag{4.38}$$

III—Gradient exponentiel

Finalement, la dernière méthode de combinaison de modèles que nous considérons est la version à *tranche fixe* (Herbster et Warmuth 1998) de l'algorithme du gradient exponentiel de Kivinen et Warmuth (1997). Cette méthode emploie une mise-à-jour exponentielle des poids, suivie d'une étape de « redistribution » qui ne permet pas à aucun poids de devenir trop important.

Tout d'abord, les poids « bruts » sont calculés à partir de la perte encourue au temps précédent :

$$w'_{jt} = w_{j(t-1)}^{\mathsf{exp.\,grad.}}\, e^{\eta \hat{W}_j^S(t;\boldsymbol{\theta}_{j,\mathcal{D}^{(k(t))}})}. \tag{4.39}$$

Ensuite, un « impôt » proportionnel est prélevé sur tous les poids, et redistribué de manière uniforme, produisant la nouvelle valeur des poids :

$$\begin{aligned} \mathsf{pool}_t &= \sum_{k=1}^{G} w'_{kt} \\ w_{jt}^{\mathsf{exp.\,grad.}} &= (1-\alpha) w'_{jt} + \frac{1}{G-1}(\mathsf{pool}_t - \alpha w'_{jt}). \end{aligned} \tag{4.40}$$

Les paramètres η et α contrôlent respectivement la vitesse de convergence de la moyenne exponentielle, et le seuil sous lequel aucun poids ne peut descendre.

Herbster et Warmuth (1998) présentent une preuve détaillée de la convergence de cet algorithme et de bornes sur l'erreur de généralisation.

4.5 Schéma des expériences

Nous sommes maintenant en mesure de dresser un portrait des expériences dont les résultats sont présentés et analysés au chapitre suivant.

Le problème de gestion de portefeuille que nous considérons pour ce mémoire est l'allocation entre les 14 secteurs de l'indice boursier canadien TSE 300. Chaque secteur correspond à un volet important de l'économie canadienne. À chaque secteur est associée une pondération w_{it}, dérivée de la capitalisation boursière des titres qui le constitue (cette pondération varie donc dans le temps), et le rendement pondéré de chaque secteur produit, par définition, celui de l'indice :

$$r_{(\text{TSE300})t} \stackrel{\text{def}}{=} \sum_{i=1}^{14} w_{i(t-1)} r_{it}.$$

Il est à noter que le rendement de l'indice pour une période est défini par rapport aux pondérations qui prévalent *au début de la période*, et non à la fin de celle-ci. C'est pourquoi nous utilisons $w_{i(t-1)}$ pour les rendements de la période t.

Modèle de marché Nous comparons fréquemment la performance d'un modèle par rapport « au marché ». Pour nos expériences, le rendement du marché est toujours défini comme étant celui du TSE 300.

Portefeuille de référence La fonction de coût que nous utilisons pour certains modèles fait appel à un portefeuille de référence (*cf.* éq. (3.67)). Dans tous les cas où c'est nécessaire, nous utilisons les pondérations de chaque secteur comme référence :

$$\boldsymbol{\psi}_t = (w_{0t}, w_{1t}, \ldots, w_{14t})'.$$

Cette référence est adéquate pour une comparaison par rapport au TSE 300, car elle favorise « par défaut » des rendements comparables à ceux du marché.

4.5.1 Comparaison entre les types de modèles

La première série d'expériences vise à comprendre l'effet du type de modèle (et donc de la fonction de coût), de la topologie du réseau, et des paramètres de contrôle de la capacité sur la performance obtenue. Nous considérons :

Type de modèle Nous comparons (i) le modèle de décision sans récurrence, (ii) le modèle de décision avec récurrence, (iii) le modèle de prévision sans récurrence, utilisant une allocation QOD_{15}, c'est-à-dire de type moyenne–variance avec un paramètre d'aversion au risque $\lambda = 15$.

Topologie du réseau Pour chaque type de modèle, nous évaluons l'impact du nombre d'unités cachées sur ce modèle. Nous considérons les cas pour 2, 5, et 10 unités cachées.

Contrôle de la capacité Nous évaluons l'impact des pénalisation sur la norme des poids et des entrées sur chacun des cas ci-haut. Puisque nous ne connaissons pas de bons estimés à priori, nous évaluons la performance pour les combinaisons $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\} \times \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$ des valeurs des paramètres.

4.5.2 Comparaison entre les méthodes de combinaison

Le deuxième série d'expériences vise à comparer l'efficacité des méthodes de combinaison de modèles décrites précédemment (hardmax, softmax, gradient exponentiel). Nous évaluons l'efficacité relative de ces méthodes, de même que la performance d'un comité par rapport aux modèles sous-jacents qui le composent.

4.6 Ensembles de données

4.6.1 Description générale

Les séries financières brutes que nous utilisons pour les expériences sont les suivantes :

- 14 secteurs du TSE 300. Données mensuelles. Données disponibles de 1956 à fin-1996.
- Taux sans risque : obligations à court-terme (90 jours) du gouvernement canadien.
- Données « fondamentales » et macro-économiques : nous avons accès aux données de profit par action (*earning per share*), moyennées sur chaque secteur du TSE. Nous avons accès à un ensemble de 24 variables macro-économiques (dont la nature ne peut être révélée, et qui sont donc utilisées ensembles).

Après élimination des valeurs manquantes, ces séries sont continues sur l'intervalle allant de 1971/01/30 à 1996/07/30, pour un total de 307 observations mensuelles.

4.6.2 Variables explicatives et prétraitements

Les variables explicatives ($\mathbf{u}_t$) fournies en entrées sont le résultat de plusieurs prétraitements. Nous construisons tout d'abord un vecteur de base (non-normalisé) de 75 entrées, que nous dénotons $\tilde{\mathbf{u}}_t$, comme suit :

3 séries de rendements En plus des rendements des 14 secteurs au temps courant $\mathbf{r}_t$, nous fournissons une moyenne mobile des rendements passés sur une courte et une moyenne échéance. La moyenne mobile est calculée suivant une pondération exponentielle (Gourieroux et Monfort 1997) :

$$\mathbf{r}_t^{(\lambda)} = \lambda \mathbf{r}_{t-1}^{(\lambda)} + (1-\lambda)\mathbf{r}_t, \tag{4.41}$$

où λ est le facteur d'oubli. Nous utilisons $\lambda = 0.80$ et $\lambda = 0.94$, respectivement pour la moyenne mobile à courte et à moyenne échéance.[9]

2 séries de volatilités Nous fournissons le log-écart-type du rendement des secteurs, calculé par l'estimateur de variance pondéré exponentiellement, éq. (2.56). (Nous n'incluons pas les corrélations entre les actifs). Nous

[9] Ces valeurs correspondent à une demie-vie de 3.11 mois pour $\lambda = 0.80$, et de 11.20 mois pour $\lambda = 0.94$.

employons les facteurs d'oubli $\lambda = 0.97$ et $\lambda = 0.80$, pour des indicateurs de volatilité à long et court terme.

Sommes des séries ci-haut Pour chacune des cinq séries précédentes, nous calculons la somme (au temps t) de ses 14 éléments, et nous incluons le résultat comme entrée supplémentaire. Nous justifions cette opération sur la base suivante : la normalisation des entrées (décrite plus bas) transformera cette somme en une sorte de moyenne, et cette dernière pourra, à peu de coût pour le réseau, être comparée à l'entrée correspondant à un secteur individuel ; le résultat de cette comparaison peut ensuite servir à sur– ou sous–pondérer un secteur si ce dernier est attrayant ou non par rapport à l'ensemble du marché.

Normalisation

Ensuite, le vecteur $\tilde{\mathbf{u}}_t$ est normalisé avant d'être fourni au réseau. Cette opération en ramène la distribution à une moyenne de 0 et un écart-type de 1, en utilisant les données historiques disponibles jusqu'au temps t :

$$u_{it} = \frac{\tilde{u}_{it} - \hat{\mu}_{it}}{\hat{\sigma}_{it}}, \tag{4.42}$$

où

$$\hat{\mu}_{it} = \frac{1}{t}\sum_{\tau=1}^{t} \tilde{u}_{i\tau} \qquad \text{et} \qquad \hat{\sigma}_{it} = \frac{1}{t-1}\sum_{\tau=1}^{t} (\tilde{u}_{i\tau} - \hat{\mu}_{it})^2. \tag{4.43}$$

Chapitre 5

Résultats et analyse

Ce chapitre présente les résultats expérimentaux et analyse leur signification. Nous commençons par passer en revue quelques principes statistiques qui seront nécessaires aux analyses. Ensuite, nous présentons les résultats comparant les différents types de modèles. Finalement, nous présentons les résultats reliés aux méthodes de combinaison de modèles.

5.1 Brèves remarques statistiques

5.1.1 Comparaison de séries de rendements

Nous devons fréquemment comparer les séries des rendements (normalisés par la valeur à risque) générés par deux modèles différents, et tester l'hypothèse « est-ce que le modèle 1 produit un rendement *moyen* significativement supérieur au modèle 2 ? » Bien que la réponse habituelle est d'utiliser un t-test standard, nous devons procéder avec prudence car la présence éventuelle de corrélations dans les séries risquerait d'invalider les résultats.

La figure 5.1 montre le corrélogramme de deux séries de rendements tirées

au hasard parmi tous les résultats (toutes les autres séries examinées ont un comportement similaire). Nous ne remarquons pas d'autocorrélation significative à aucun délai, mais une corrélation croisée instantanée très importante (au délai 0). Les autres corrélations croisées ne sont pas significatives.

Cette structure de corrélation dans les séries justifie l'utilisation d'un t-test apparié (*paired t-test*) pour comparer la différence de leurs moyennes. Pour des séries r_1 et r_2 de T observations, nous calculons la statistique :

$$t = \frac{\hat{\mu}}{\hat{\sigma}},$$

où

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} r_{1t} - r_{2t} \qquad \text{et} \qquad \hat{\sigma}^2 = \frac{1}{T-1}\sum_{t=1}^{T}((r_{1t} - r_{2t}) - \hat{\mu})^2$$

et testons l'hypothèse nulle $t = 0$, sachant que t est distribué, sous l'hypothèse nulle, selon une distribution de Student avec $T-1$ degrés de liberté (Stuart, Ord et Arnold 1998).

5.1.2 Analyse de la variance

L'analyse de la variance (Scheffé 1959; Box, Hunter et Hunter 1978), ou ANOVA, est une méthode utilisée en statistique pour tester l'effet de plusieurs facteurs expérimentaux (chaque facteur pouvant prendre des niveaux discrets) sur une quantité mesurée. La procédure décompose la variance observée dans un ensemble d'échantillons en plusieurs parties qui sont attribuables aux différents facteurs.

Considérons le cas simple d'un seul facteur. Supposons que nous avons M échantillons, chacun ayant subi un « traitement » différent. Chaque échantillon comporte un nombre T d'observations. Nous dénoterons les observations individuelles r_{mt}.[1] Nous voulons tester si les *moyennes* des échantillons diffèrent les unes des autres. Nous faisons l'hypothèse que la variance de tous les

[1]Cette notation diffère légèrement de celle habituellement utilisée dans la littérature statistique ; cependant c'est celle qui est la plus appropriée pour notre application immédiate de comparer des séries de rendements générés par différents modèles.

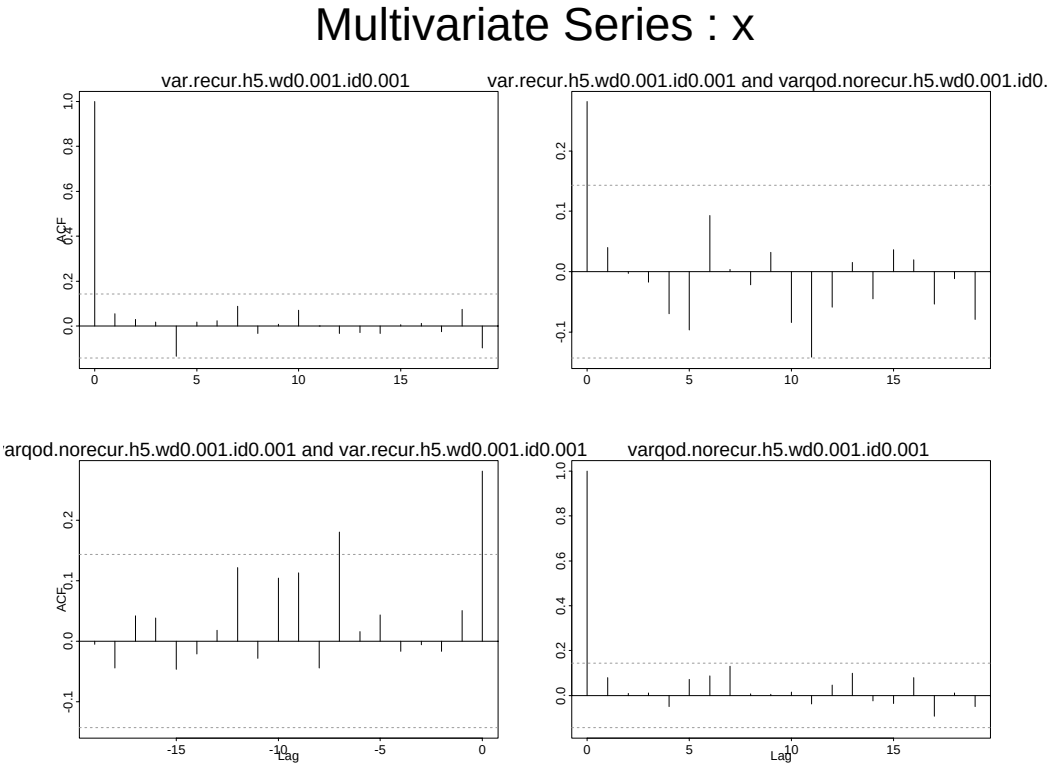


Figure 5.1: *Autocorrélation des rendements générés par le* **modèle de décision avec récurrence** *(haut gauche), et par le* **modèle de prévision** QOD_{15} **sans récurrence** *(bas droite), les deux utilisant un MLP à 5 unités cachées avec* $ID = 10^{-3}$ *et* $WD = 10^{-3}$*. Les corrélations croisées entre les deux séries pour les délais positifs et négatifs sont en haut droite, et en bas à gauche respectivement. Les droites horizontales pointillées représentent les seuils de significativité à 5%.*

échantillons demeure la même. Nous posons le modèle suivant pour expliquer une observation r_{mt} :

$$r_{mt} = \mu + \alpha_m + \epsilon_{mt}, \tag{5.1}$$

où pour tout m et tout t, $\epsilon_{mt} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, et $\sum_m \alpha_m = 0$. Ce modèle signifie qu'une observation peut être décomposée en trois parties : une moyenne globale, dénotée μ ; un effet spécifique à l'échantillon (causé par le « traitement » subi), dénoté α_m ; et une erreur expérimentale ϵ_{mt}.

L'essence de l'analyse de la variance repose sur l'observation que, sous l'hypothèse nulle où tous les α_m sont zéro, la variance à l'intérieur de chaque échantillon devrait être comparable à celle entre les échantillons. Nous testons formellement cette hypothèse ainsi : nous commençons par former la somme des carrés internes et externes :

$$S_I = \sum_{m=1}^{M} \sum_{t=1}^{T} (r_{mt} - \bar{r}_m)^2 \qquad \text{et} \qquad S_E = \sum_{m=1}^{M} T(\bar{r}_m - \bar{r})^2, \tag{5.2}$$

où $\bar{r}_m = \frac{1}{T} \sum_t r_{mt}$ est la moyenne de l'échantillon m, et $\bar{r} = \frac{1}{M} \sum_m \bar{r}_m$ est la moyenne globale.

Sous l'hypothèse nulle d'égalité des moyennes des échantillons, la statistique

$$F = \frac{S_E/(M-1)}{S_I/(T-M)} \tag{5.3}$$

suit une distribution F avec $M-1$ et $T-M$ degrés de liberté. Nous pouvons rejeter l'hypothèse nulle si $\Pr(F) \leq 0.05$, pour un niveau de test à 95%.

5.1.3 Généralisation à plusieurs facteurs

Lorsque plusieurs facteurs expérimentaux entrent en jeu simultanément, nous devons considérer un modèle plus complexe, qui tient compte des interactions entre ces facteurs. Par exemple, pour deux facteurs :

$$r_{mnt} = \mu + \alpha_m + \beta_n + \gamma_{mn} + \epsilon_{mnt}. \tag{5.4}$$

De même, pour trois facteurs, nous aurons trois termes caractérisant les effets principaux, trois termes caractérisant les interactions du second ordre, et un terme pour l'interaction du troisième ordre. Nous n'avons jamais recours à une ANOVA de plus de trois facteurs.

5.1.4 Vérification des hypothèses

Le test présenté ci-haut repose sur deux hypothèses : celle de l'homogénéité de la variance, et celle de la normalité de la distribution des rendements. Bien que nous ne présentons pas de test formel confirmant ces hypothèses, nous constatons qu'elles semblent raisonnablement respectées, par les figures 5.2 et 5.3, qui illustrent respectivement un « boxplot » de la distribution marginale de chaque facteur de tous les résultats, et un « QQ-plot » de la distribution des rendements illustrés à la figure 5.1.

5.2 Comparaison entre les types de modèles

5.2.1 Résultats bruts

Les résultats bruts de la première série d'expériences, comparant les types de modèles, apparaîssent aux tableaux 5.1, 5.2, et 5.3, respectivement donnant les résultats du **modèle de décision sans récurrence**, du **modèle de décision avec récurrence**, et du **modèle de prévision** QOD_{15} **sans récurrence**.

Chaque tableau donne le résultat de 48 expériences, chacune résultant de l'une des combinaisons $NC \in \{2, 5, 10\}$ (nombre d'unités cachées), et $WD \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$, $ID \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$ (pénalisation sur la norme des poids et des entrées respectivement).

Dans tous les cas, la mesure de performance est le rendement mensuel moyen net (après frais de transaction, qui sont fixés à 0.1%), normalisé par

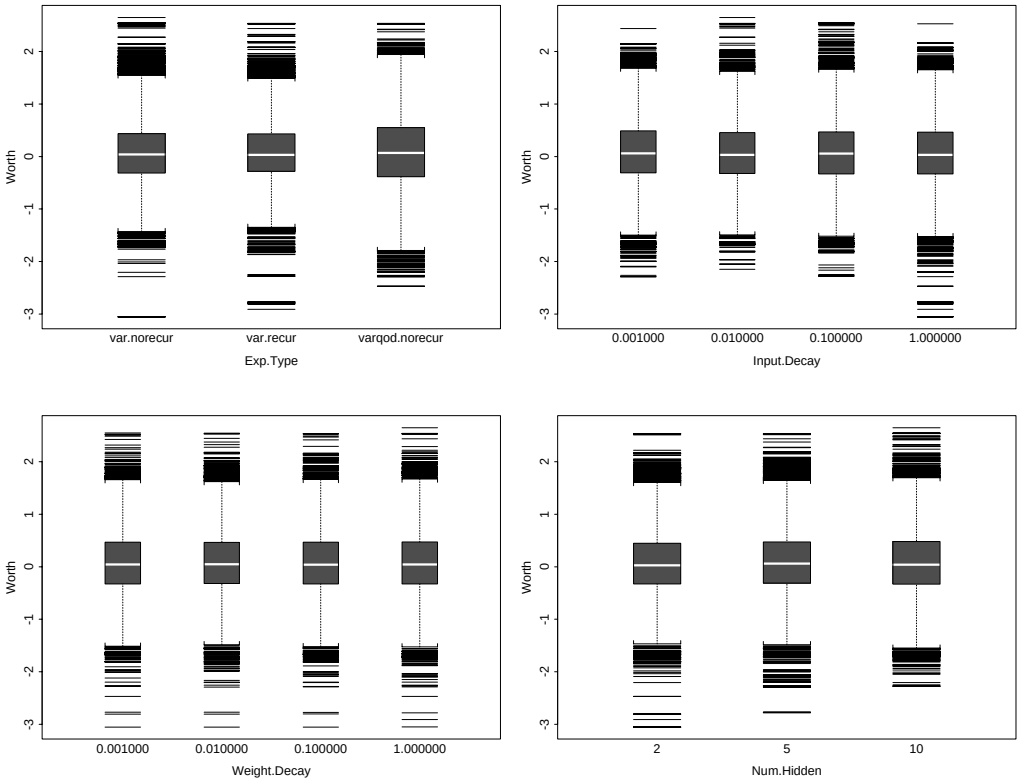


Figure 5.2: « *Boxplots* » *du rendement marginal pour chaque facteur de toutes les expériences, ce qui confirme visuellement l'homogénéité de la variance pour les différents facteurs.*

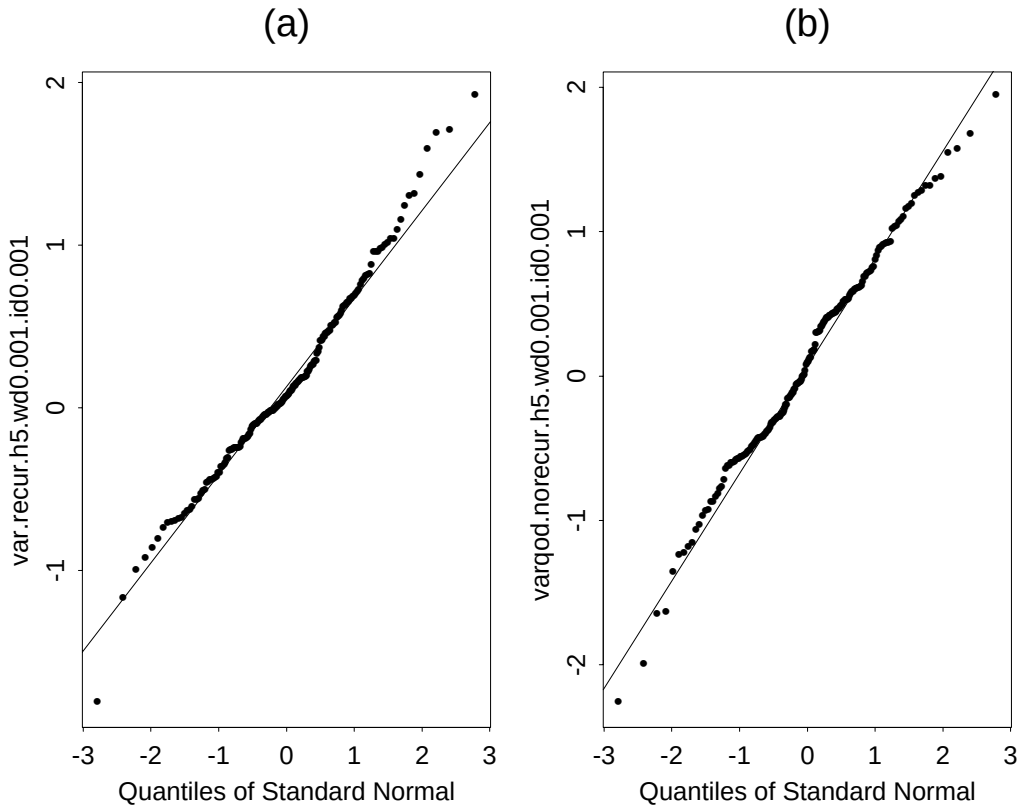


Figure 5.3: *Graphiques quantile-quantile comparant la distribution normale aux distributions empiriques des rendements obtenus par (a) le* **modèle de décision avec récurrence**, *et (b) le* **modèle de prévision** QOD_{15} **sans récurrence**, *les deux utilisant un MLP à 5 unités cachées avec* $ID = 10^{-3}$ *et* $WD = 10^{-3}$.

la valeur-à-risque.[2] L'erreur–type de l'estimé de chaque moyenne est donnée entre parenthèse, et est calculée sous l'hypothèse que les rendements sont non-autocorrélés et suivent une distribution normale.

Tous les résultats sont comparés aux rendements du marché sur la même période, et les différences statistiquement significatives à 5% (selon le t-test apparié décrit à la section 5.1.1) sont notées en gras.

À titre d'exemple, les décisions prises par l'un des modèles, de même que les rendements mensuels et cumulatifs de ces décisions sont illustrés à la figure 5.4.

[2]Par simplicité, la VàR est fixée à 1$, mais cette valeur est immatérielle.

Tableau 5.1: *Résultats pour les* **modèles de décision sans récurrence** *(MLP ; 2, 5, et 10 unités cachées), en fonction des pénalisations sur la norme des poids (notée WD) et des entrées (notée ID). Le rendement moyen est en dollars de profit net sur une base mensuelle. Les erreurs–types des estimés des moyennes sont entre parenthèses. Les résultats significativement différents de zéro ($p = 0.95$) sont en gras. La différence par rapport au marché est notée ; le rendement mensuel du marché est, dans tous les cas, de 0.007, avec une erreur–type de 0.042.*

		2 unités cachées				5 unités cachées				10 unités cachées			
WD	ID	Rend. moy.		Diff. marché		Rend. moy.		Diff. marché		Rend. moy.		Diff. marché	
10^{-3}	10^{-3}	0.034	(0.042)	0.027	(0.052)	**0.094**	(**0.044**) ⋆	0.087	(0.065)	0.082	(0.045)	0.075	(0.064)
10^{-3}	10^{-2}	0.021	(0.042)	0.014	(0.053)	0.084	(0.043)	0.077	(0.061)	0.060	(0.046)	0.053	(0.062)
10^{-3}	10^{-1}	0.046	(0.045)	0.039	(0.058)	**0.104**	(**0.042**) ⋆	0.097	(0.053)	**0.104**	(**0.046**) ⋆	0.097	(0.057)
10^{-3}	10^{0}	0.035	(0.045)	0.028	(0.022)	0.031	(0.045)	0.024	(0.043)	0.033	(0.044)	0.027	(0.057)
10^{-2}	10^{-3}	0.035	(0.041)	0.028	(0.051)	**0.116**	(**0.042**) ⋆	0.109	(0.065)	0.078	(0.043)	0.071	(0.062)
10^{-2}	10^{-2}	0.007	(0.041)	0.000	(0.052)	**0.096**	(**0.043**) ⋆	0.089	(0.059)	0.065	(0.046)	0.058	(0.062)
10^{-2}	10^{-1}	0.046	(0.045)	0.039	(0.057)	0.072	(0.042)	0.065	(0.055)	**0.102**	(**0.045**) ⋆	0.095	(0.057)
10^{-2}	10^{0}	0.022	(0.045)	0.015	(0.022)	0.022	(0.045)	0.015	(0.046)	0.042	(0.044)	0.036	(0.058)
10^{-1}	10^{-3}	0.049	(0.042)	0.042	(0.052)	**0.104**	(**0.042**) ⋆	0.097	(0.061)	0.078	(0.044)	0.071	(0.064)
10^{-1}	10^{-2}	0.021	(0.042)	0.014	(0.052)	**0.088**	(**0.042**) ⋆	0.081	(0.058)	0.077	(0.047)	0.070	(0.062)
10^{-1}	10^{-1}	0.054	(0.044)	0.047	(0.058)	**0.091**	(**0.044**) ⋆	0.084	(0.055)	**0.101**	(**0.046**) ⋆	0.094	(0.057)
10^{-1}	10^{0}	0.002	(0.045)	−0.005	(0.022)	0.029	(0.044)	0.022	(0.048)	0.052	(0.044)	0.045	(0.057)
10^{0}	10^{-3}	0.033	(0.041)	0.026	(0.050)	**0.140**	(**0.042**) ⋆	**0.133**	(**0.062**) ⋆	0.072	(0.046)	0.065	(0.064)
10^{0}	10^{-2}	0.034	(0.042)	0.027	(0.053)	0.079	(0.042)	0.072	(0.060)	0.051	(0.047)	0.044	(0.063)
10^{0}	10^{-1}	0.052	(0.044)	0.045	(0.057)	0.079	(0.044)	0.072	(0.056)	0.074	(0.046)	0.067	(0.057)
10^{0}	10^{0}	0.035	(0.045)	0.029	(0.020)	0.014	(0.047)	0.007	(0.036)	0.054	(0.044)	0.047	(0.056)

Tableau 5.2: *Résultats pour les* **modèles de décision avec récurrence** *(MLP ; 2, 5, et 10 unités cachées), en fonction des pénalisations sur la norme de poids (notée WD) et des entrées (notée ID). Les mêmes notes qu'au tableau 5.1 s'appliquent.*

		2 unités cachées				5 unités cachées				10 unités cachées			
WD	ID	Rend. moy.		Diff. marché		Rend. moy.		Diff. marché		Rend. moy.		Diff. marché	
10^{-3}	10^{-3}	0.076	(0.041)	0.070	(0.056)	**0.129**	(**0.042**) ⋆	**0.122**	(**0.059**) ⋆	**0.117**	(**0.042**) ⋆	0.110	(0.057)
10^{-3}	10^{-2}	0.048	(0.039)	0.041	(0.055)	**0.141**	(**0.041**) ⋆	**0.134**	(**0.054**) ⋆	0.080	(0.043)	0.073	(0.057)
10^{-3}	10^{-1}	0.009	(0.041)	0.002	(0.049)	0.058	(0.041)	0.051	(0.053)	0.048	(0.044)	0.041	(0.056)
10^{-3}	10^{0}	0.037	(0.043)	0.030	(0.026)	0.032	(0.042)	0.025	(0.030)	0.041	(0.046)	0.035	(0.053)
10^{-2}	10^{-3}	0.075	(0.041)	0.068	(0.055)	**0.125**	(**0.041**) ⋆	**0.118**	(**0.058**) ⋆	0.079	(0.042)	0.072	(0.058)
10^{-2}	10^{-2}	**0.078**	(**0.038**) ⋆	0.071	(0.051)	**0.139**	(**0.042**) ⋆	**0.132**	(**0.055**) ⋆	0.082	(0.043)	0.075	(0.057)
10^{-2}	10^{-1}	−0.021	(0.042)	−0.028	(0.050)	0.044	(0.042)	0.037	(0.053)	0.032	(0.043)	0.025	(0.056)
10^{-2}	10^{0}	0.037	(0.043)	0.030	(0.026)	0.038	(0.042)	0.031	(0.032)	0.036	(0.046)	0.029	(0.054)
10^{-1}	10^{-3}	0.037	(0.040)	0.030	(0.054)	**0.122**	(**0.041**) ⋆	**0.116**	(**0.058**) ⋆	**0.088**	(**0.042**) ⋆	0.081	(0.057)
10^{-1}	10^{-2}	**0.081**	(**0.039**) ⋆	0.074	(0.052)	**0.141**	(**0.042**) ⋆	**0.135**	(**0.055**) ⋆	0.064	(0.042)	0.057	(0.057)
10^{-1}	10^{-1}	−0.006	(0.041)	−0.013	(0.051)	0.054	(0.042)	0.047	(0.054)	0.037	(0.044)	0.030	(0.056)
10^{-1}	10^{0}	0.041	(0.043)	0.034	(0.027)	0.047	(0.041)	0.041	(0.032)	0.028	(0.045)	0.021	(0.052)
10^{0}	10^{-3}	**0.094**	(**0.040**) ⋆	0.087	(0.056)	**0.114**	(**0.043**) ⋆	0.107	(0.056)	0.077	(0.043)	0.070	(0.057)
10^{0}	10^{-2}	0.066	(0.039)	0.059	(0.056)	**0.123**	(**0.043**) ⋆	**0.116**	(**0.055**) ⋆	0.057	(0.041)	0.050	(0.056)
10^{0}	10^{-1}	0.011	(0.041)	0.004	(0.049)	0.057	(0.041)	0.050	(0.053)	0.035	(0.044)	0.028	(0.056)
10^{0}	10^{0}	0.031	(0.043)	0.024	(0.027)	0.026	(0.043)	0.019	(0.033)	0.009	(0.044)	0.002	(0.052)

Tableau 5.3: *Résultats pour les* **modèles de prévision explicite** QOD_{15} **sans récurrence** *(MLP ; 2, 5, et 10 unités cachées), en fonction des pénalisations sur la norme des poids (notée WD) et des entrées (notée ID). Les mêmes notes qu'au tableau 5.1 s'appliquent.*

		2 unités cachées				5 unités cachées				10 unités cachées			
WD	ID	Rend. moy.		Diff. marché		Rend. moy.		Diff. marché		Rend. moy.		Diff. marché	
10^{-3}	10^{-3}	**0.111**	(**0.050**) ⋆	0.104	(0.054)	0.076	(0.053)	0.069	(0.057)	**0.099**	(**0.047**) ⋆	0.092	(0.060)
10^{-3}	10^{-2}	0.082	(0.050)	0.075	(0.051)	0.021	(0.051)	0.014	(0.055)	0.017	(0.046)	0.011	(0.060)
10^{-3}	10^{-1}	0.075	(0.051)	0.068	(0.050)	**0.151**	(**0.050**) ⋆	**0.144**	(**0.055**) ⋆	**0.125**	(**0.057**) ⋆	0.118	(0.061)
10^{-3}	10^{0}	0.103	(0.057)	0.096	(0.049)	0.079	(0.053)	0.073	(0.051)	0.092	(0.057)	0.085	(0.060)
10^{-2}	10^{-3}	**0.121**	(**0.051**) ⋆	**0.114**	(**0.052**) ⋆	0.078	(0.053)	0.071	(0.056)	**0.097**	(**0.047**) ⋆	0.091	(0.058)
10^{-2}	10^{-2}	**0.112**	(**0.052**) ⋆	**0.106**	(**0.052**) ⋆	0.002	(0.051)	−0.005	(0.054)	0.041	(0.048)	0.034	(0.059)
10^{-2}	10^{-1}	0.098	(0.052)	0.092	(0.053)	**0.163**	(**0.053**) ⋆	**0.156**	(**0.056**) ⋆	0.087	(0.052)	0.080	(0.056)
10^{-2}	10^{0}	0.081	(0.057)	0.075	(0.049)	**0.107**	(**0.052**) ⋆	**0.100**	(**0.050**) ⋆	0.068	(0.057)	0.062	(0.059)
10^{-1}	10^{-3}	**0.100**	(**0.050**) ⋆	0.093	(0.052)	0.076	(0.056)	0.069	(0.060)	0.087	(0.048)	0.080	(0.059)
10^{-1}	10^{-2}	0.077	(0.051)	0.070	(0.052)	0.029	(0.052)	0.023	(0.055)	0.030	(0.048)	0.023	(0.059)
10^{-1}	10^{-1}	0.098	(0.051)	0.092	(0.051)	**0.145**	(**0.050**) ⋆	**0.138**	(**0.054**) ⋆	**0.120**	(**0.057**) ⋆	0.113	(0.062)
10^{-1}	10^{0}	0.071	(0.058)	0.064	(0.051)	**0.106**	(**0.053**) ⋆	**0.099**	(**0.049**) ⋆	0.081	(0.057)	0.074	(0.059)
10^{0}	10^{-3}	**0.105**	(**0.050**) ⋆	0.098	(0.052)	0.063	(0.054)	0.056	(0.058)	**0.110**	(**0.048**) ⋆	0.103	(0.055)
10^{0}	10^{-2}	0.088	(0.049)	0.081	(0.053)	0.069	(0.054)	0.062	(0.064)	0.019	(0.050)	0.013	(0.058)
10^{0}	10^{-1}	0.086	(0.052)	0.079	(0.052)	**0.148**	(**0.052**) ⋆	**0.141**	(**0.056**) ⋆	**0.115**	(**0.055**) ⋆	0.108	(0.060)
10^{0}	10^{0}	0.109	(0.058)	**0.102**	(**0.051**) ⋆	**0.118**	(**0.052**) ⋆	**0.111**	(**0.046**) ⋆	0.076	(0.056)	0.069	(0.056)

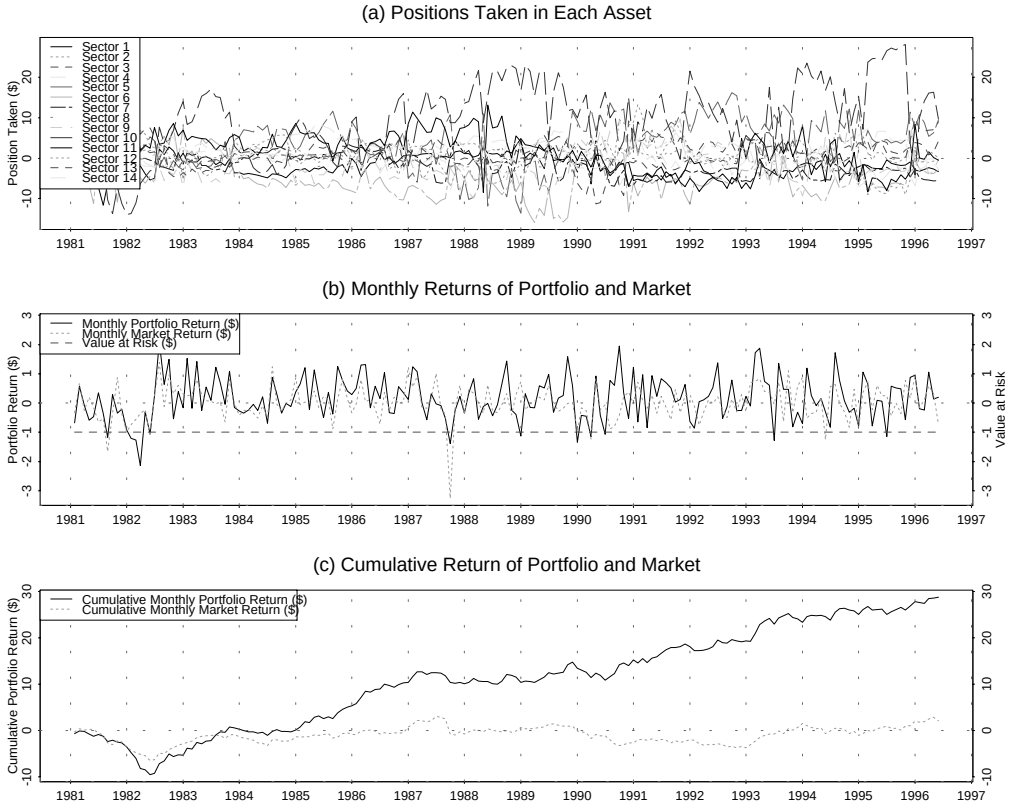


Figure 5.4: *Exemple détaillé de la gestion de portefeuille pour le comité formé des modèles de prévision explicite à l'aide de la méthode du gradient exponentiel. La partie du haut illustre les positions prises (en $) dans chacun des 14 secteurs du TSE 300. La partie centrale illustre les rendements mensuels, et compare aux rendements du marché ; la VàR de 1$ est aussi notée, laquelle ne devrait pas (en principe) être traversée pour 95% des périodes. La partie du bas trace les rendements cumulatifs.*

5.2.2 Analyse

Inspection préliminaire

Une inspection superficielle des tableaux précédents permet de dégager les grandes lignes suivantes :

- La pénalisation sur la norme des entrées a un effet important pour tous les types de modèles.
- Le nombre d'unités cachées joue aussi un certain rôle, surtout pour les modèles de décision.
- L'effet de la pénalisation sur la norme des poids est plus mitigé.
- Le type de modèle semble avoir une importance, mais elle est modulée par les autres paramètres ; globalement, le modèle de prévision semble avoir un léger avantage.

Ces impressions sont confirmées visuellement à la figure 5.5, qui illustre, pour tous les types de modèles, le rendement moyen obtenu pour chaque niveau de chaque facteur (marginalisant par rapport aux autres facteurs). Par exemple, la partie (a) de la figure montre, pour le modèle de décision sans récurrence, les effets sur la performance financière du nombre d'unités cachées, et des pénalisations sur la norme des poids et des entrées. Le graphique s'interprète comme suit : pour chaque niveau d'un facteur donné (par exemple, le nombre d'unités cachées prend les niveaux 2, 5, et 10), la performance financière moyenne à ce niveau du facteur est tracée, *faisant la moyenne sur tous les autres facteurs* (en l'occurrence les pénalisations sur la norme des poids et des entrées).

Les parties (b) et (c) de la figure donnent la même information, respectivement, pour les modèles de décision avec récurrence et de prévision sans récurrence.

Cette figure confirme la très faible variance dans la performance financière attribuable à la pénalisation sur la norme des poids, et l'effet beaucoup plus important du nombre d'unités cachées et de la pénalisation sur la norme des entrées.

Les effets des pénalisations sur la norme des poids et des entrées sont illustrés de façon plus spécifique à la figure 5.6, en prenant une coupe des résultats pour le modèle de prévision, fixant le nombre d'unités cachées à 5.

Résultats de l'anova

Les tableaux 5.4, 5.5, et 5.6 présentent les résultats de l'ANOVA effectuée sur chaque type de modèle. Ces ANOVAs permettent de dégager quels facteurs ont une importance statistiquement significative.

Les résultats obtenus sont clairs et confirment les inspections graphiques :

- Pour tous les types de modèles, la pénalisation sur la norme des entrées a un impact très significatif.
- Le nombre d'unités cachées est significatif pour les modèles de décision (avec et sans récurrence), mais ne l'est pas pour le modèle de prévision.
- La pénalisation sur la norme des poids n'a jamais d'effet significatif ; ne sont non plus significatives aucune des interactions d'ordre supérieur entre ces facteurs.

Comparaison entre les modèles

Nous nous penchons maintenant sur la comparaison entre les différents types de modèles. Les résultats de l'ANOVA sur les type de modèles (tableau 5.7) rejette clairement l'hypothèse nulle qu'ils sont tous équivalents.

Pour comprendre leurs différences, nous avons comparé les modèles paire par paire. Le tableau 5.8 montre la différence de performance entre toutes les paires de modèles (en moyennant sur tous les autres facteurs, i.e. le nombre d'unités cachées, pénalisations sur la norme des poids et des entrées). Nous observons que, au plus haut niveau, le modèle de prévision est significativement meilleur que les deux modèles de décision ; de plus, il n'y a pas de différence entre ces deux derniers.

Cependant, ces conclusions tirées d'après une moyenne à grande échelle, ne tiennent pas lorsqu'on considère des sous-ensembles des paramètres. Par exemple le tableau 5.9 montre la différence de performance entre les paires de

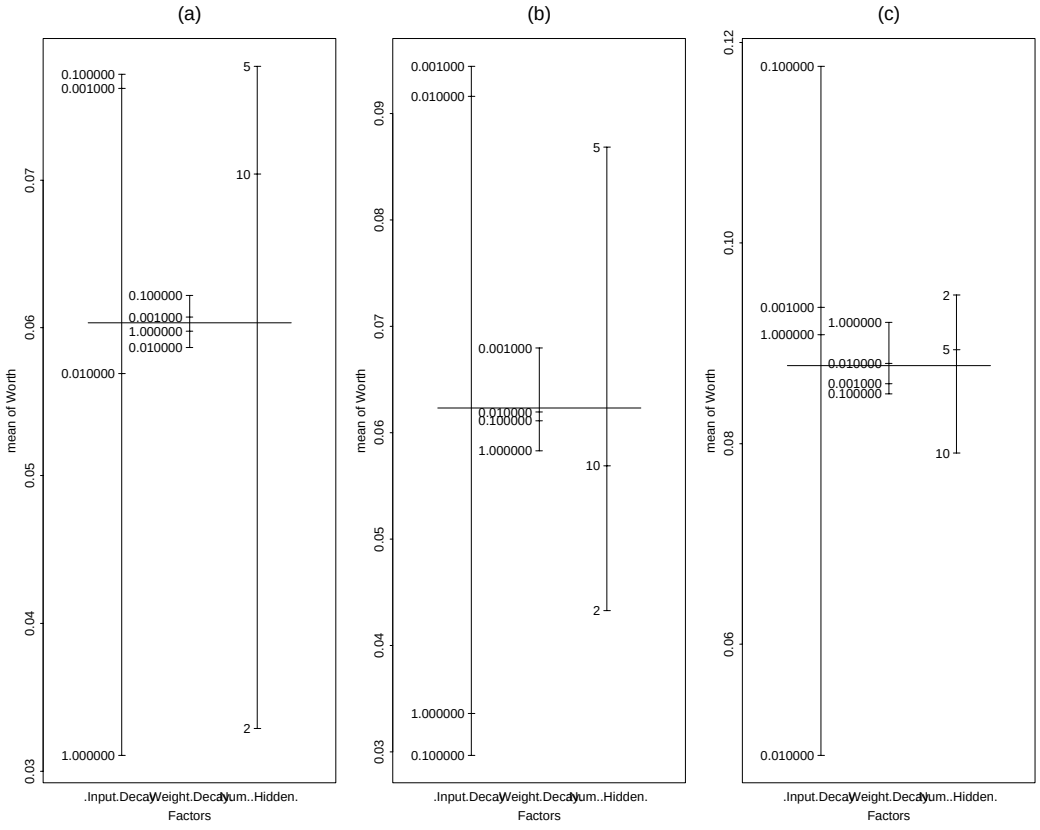
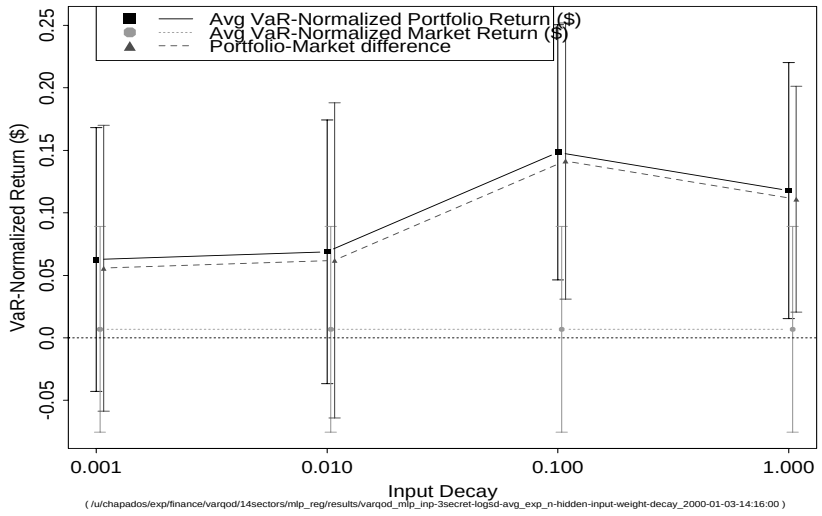
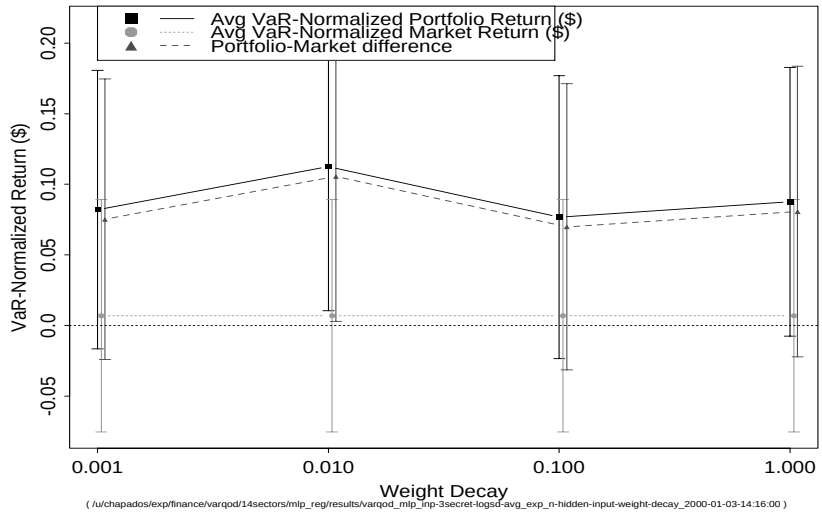


Figure 5.5: *Effet de chaque facteur sur le rendement mensuel moyen : pour chaque facteur (*weight decay, input decay, *etc.), la performance financière moyenne pour tous les niveaux de ce facteur est tracée, en moyennant par rapport à tous les autres facteurs. (a) Modèle de décision sans récurrence (*cf. *tableau 5.1). (b) Modèle de décision avec récurrence (*cf. *tableau 5.2). (c) Modèle de prévision* QOD_{15} *sans récurrence (*cf. *tableau 5.3).*



(a)



(b)

Figure 5.6: *Effets de la pénalisation sur la norme des entrées (a) et de la pénalisation sur la norme des poids (b) sur la performance, et comparaison par rapport au marché. Dans les deux cas, le modèle employé est un modèle de prévision* QOD_{15}*, utilisant un MLP à 5 unités cachées. En (a), la pénalisation sur la norme des poids (WD) est fixée à 1.0. En (b), pénalisation sur la norme des entrées (ID) est fixée à 0.01. Les barres d'erreur représentent des intervalles de confiance à 95% de la moyenne des rendements.*

Tableau 5.4: *Résultats de l'*ANOVA *pour le* **modèle de décision sans récurrence**, *montrant l'effet des facteurs du nombre d'unités cachées (NC), des pénalisations sur la norme des poids (WD) et des entrées (ID), ainsi que des interactions du second et du troisième ordre entre ces facteurs.*

	Degrés de liberté	Somme des carrés	Moy. des carrés	F-valeur	Pr(F)	
ID	3	3.146	1.048831	2.936924	**0.0319713**	$\star$
WD	3	0.015	0.005050	0.014140	0.9977056	
NC	2	3.458	1.728920	4.841302	**0.0079175**	$\star$
$ID:WD$	9	0.158	0.017587	0.049248	0.9999819	
$ID:NC$	6	1.483	0.247151	0.692068	0.6560646	
$WD:NC$	6	0.114	0.019039	0.053312	0.9993946	
$ID:WD:NC$	18	0.589	0.032716	0.091610	0.9999998	
Résiduelles	8928	3188.357	0.357119			

Tableau 5.5: *Résultats de l'*ANOVA *pour le* **modèle de décision avec récurrence**.

	Degrés de liberté	Somme des carrés	Moy. des carrés	F-valeur	Pr(F)	
ID	3	8.482	2.827232	8.649223	**0.0000100**	$\star$
WD	3	0.111	0.037133	0.113599	0.9521794	
NC	2	2.969	1.484744	4.542210	**0.0106745**	$\star$
$ID:WD$	9	0.392	0.043592	0.133360	0.9988186	
$ID:NC$	6	1.505	0.250776	0.767188	0.5956445	
$WD:NC$	6	0.286	0.047722	0.145995	0.9898756	
$ID:WD:NC$	18	0.336	0.018677	0.057136	1.0000000	
Résiduelles	8928	2918.357	0.326877			

Tableau 5.6: *Résultats de l'*ANOVA *pour le* **modèle de prévision explicite** QOD_{15} **sans récurrence**.

	Degrés de liberté	Somme des carrés	Moy. des carrés	F-valeur	Pr(F)	
ID	3	5.483	1.827653	3.617164	**0.0125913**	$\star$
WD	3	0.068	0.022501	0.044532	0.9875203	
NC	2	0.384	0.191980	0.379954	0.6839038	
$ID:WD$	9	0.172	0.019150	0.037901	0.9999942	
$ID:NC$	6	3.684	0.613925	1.215037	0.2949560	
$WD:NC$	6	0.207	0.034479	0.068238	0.9987720	
$ID:WD:NC$	18	0.977	0.054266	0.107399	0.9999991	
Résiduelles	8928	4511.072	0.505272			

modèles entraînés avec une valeur de la pénalisation sur la norme des entrées $ID \in \{0.001, 0.01\}$ (et en moyennant sur les autres facteurs). Pour ce sous-ensemble de modèles, le modèle de décision avec récurrence est significativement meilleur que les deux autres.

Sur la base de ces résultats, nous ne pouvons tirer de conclusions définitives sur la valeur relative des types de modèles : la performance de chaque modèle dépend intimement des valeurs des hyperparamètres, et ces dernières, optimales pour un type de modèle seront inacceptables pour l'autre.

5.3 Combinaisons de modèles

5.3.1 Résultats bruts

Les résultats bruts de la seconde série d'expériences, comparant les méthodes de combinaison de modèles, apparaîssent aux tableaux 5.10, 5.11, 5.12, respectivement donnant les résultats des comités obtenus pour le **modèle de décision sans récurrence**, le **modèle de décision avec récurrence**, et le **modèle de prévision** QOD_{15} **sans récurrence**.

Chaque tableau donne le résultat des comités formés en combinant les modèles pour toutes les valeurs des pénalisations sur la norme des poids et des entrées, pour le nombre d'unités cachées (NC) et la méthode de combinaison indiqués. Les mêmes remarques que celles données à la section 5.2.1, concernant la mesure de performance et la comparaison par rapport au marché, s'appliquent.

Dans le cas de la méthode de combinaison par gradient exponentiel, les paramètres utilisés sont $\eta = 0.3, \alpha = 0.01$. Ces valeurs ont été choisies plus ou moins arbitrairement ; certaines expériences préalables ont suggéré une relative insensibilité des résultats à la valeur de ces paramètres.

Un graphique de ces résultats apparaît à la figure 5.7.

À titre indicatif, la figure 5.8 montre l'évolution des pondérations accordées aux $M = 16$ modèles sous-jacents pour l'exemple du modèle de décision sans

Tableau 5.7: *Résultats de l'*ANOVA *sur les types de modèles.*

	Degrés de liberté	Somme des carrés	Moy. des carrés	F-valeur	Pr(F)	
Type de modèle	2	4.19	2.095976	5.298083	**0.0050**	$\star$
Résiduelles	26925	10651.81	0.395610			

Tableau 5.8: *Comparaisons pairées entre les rendements des différents modèles, en moyennant la performance sur tous les facteurs.*

Modèle x	Modèle y	$x-y$ échant.	t-valeur	Degr. de liberté	Pr(t)	
Déc./récur.	Déc./sans récur	0.0020	0.3714	8975	0.7103	
Prév./sans récur	Déc./sans récur	0.0274	3.4569	8975	**5e-04**	$\star$
Prév./sans récur	Déc./récur	0.0254	3.2752	8975	**0.0011**	$\star$

Tableau 5.9: *Comparaisons pairées entre les rendements des différents modèles, pour $ID \in \{0.001, 0.01\}$ et en moyennant la performance sur les autres facteurs.*

Modèle x	Modèle y	$x-y$ échant.	t-valeur	Degr. de liberté	Pr(t)	
Déc./récur.	Déc./sans récur	0.0264	3.1639	4487	**0.0016**	$\star$
Prév./sans récur	Déc./sans récur	0.0047	0.4187	4487	0.6755	
Prév./sans récur	Déc./récur	−0.0217	−2.0006	4487	**0.0455**	$\star$

récurrence utilisant un MLP avec 5 unités cachées. On remarque aisément une grande différence qualitative dans l'évolution des pondérations :

- Le gradient exponentiel fait converger la majorité des poids après un certain temps vers une valeur aux environs de $1/M$, à l'exception de quelques uns qui demeurent plus importants tout continuant à varier.
- Les poids sous softmax convergent rapidement aux environs de $1/M$.
- Hardmax hésite initialement quelque peu dans son choix du meilleur modèle, puis se fixe définitivement ; les autres modèles obtiennent, bien sûr, une pondération de zéro.

Tableau 5.10: *Résultats de trois méthodes de combinaison de modèles, appliquées aux* **modèles de décision sans récurrence**, *dont les résultats bruts apparaissent au tableau 5.1. NC fait référence au nombre d'unités cachées. Le rendement du marché pour la période observée, qui diffère légèrement des résultats bruts des tableaux précédents, est de 0.009 avec une erreur–type de 0.042.*

	Gradient exponentiel		Softmax		Hardmax	
NC	Rend. moy.	Diff. marché	Rend. moy.	Diff. marché	Rend. moy.	Diff. marché
2	0.023 (0.044)	0.014 (0.035)	0.043 (0.043)	0.034 (0.033)	0.012 (0.045)	0.003 (0.059)
5	**0.107** (**0.041**) ⋆	0.099 (0.056)	**0.099** (**0.041**) ⋆	0.090 (0.053)	**0.126** (**0.041**) ⋆	0.117 (0.061)
10	**0.090** (**0.045**) ⋆	0.081 (0.060)	**0.089** (**0.045**) ⋆	0.080 (0.060)	0.083 (0.046)	0.074 (0.057)

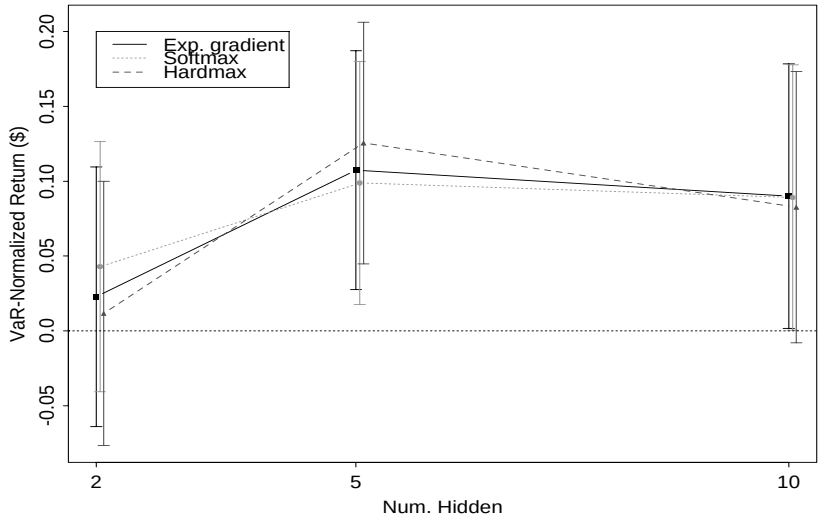
Tableau 5.11: *Résultats de trois méthodes de combinaison de modèles, appliquées aux* **modèles de décision avec récurrence**, *dont les résultats bruts apparaissent au tableau 5.2. Les mêmes remarques qu'au tableau 5.10 s'appliquent.*

	Gradient exponentiel		Softmax		Hardmax	
NC	Rend. moy.	Diff. marché	Rend. moy.	Diff. marché	Rend. moy.	Diff. marché
2	0.072 (0.043)	0.063 (0.038)	**0.087** (**0.042**) ⋆	**0.078** (**0.036**) ⋆	0.050 (0.039)	0.041 (0.052)
5	**0.138** (**0.041**) ⋆	**0.129** (**0.051**) ⋆	**0.132** (**0.041**) ⋆	**0.123** (**0.047**) ⋆	**0.124** (**0.041**) ⋆	**0.116** (**0.054**) ⋆
10	**0.090** (**0.042**) ⋆	0.081 (0.056)	**0.084** (**0.043**) ⋆	0.076 (0.056)	**0.106** (**0.042**) ⋆	0.097 (0.057)

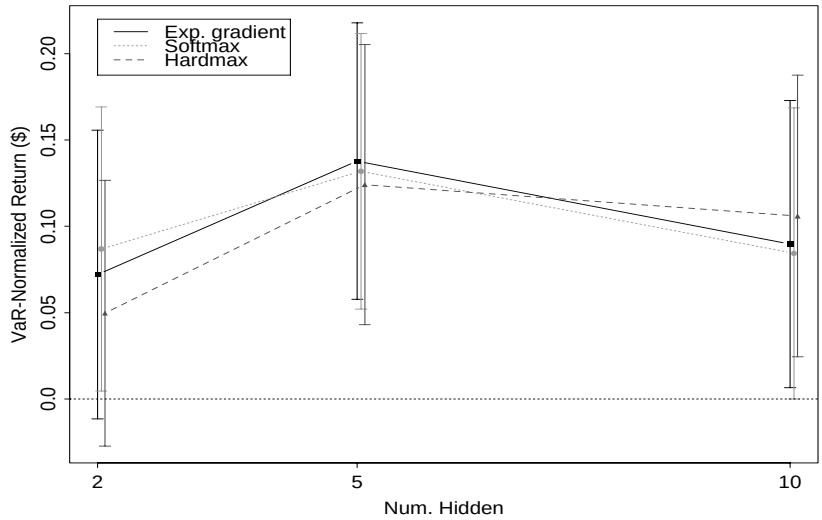
Tableau 5.12: *Résultats de trois méthodes de combinaison de modèles, appliquées aux* **modèles de prévision explicite** QOD_{15} **sans récurrence**, *dont les résultats bruts apparaissent au tableau 5.3. Les mêmes remarques qu'au tableau 5.10 s'appliquent.*

	Gradient exponentiel		Softmax		Hardmax	
NC	Rend. moy.	Diff. marché	Rend. moy.	Diff. marché	Rend. moy.	Diff. marché
2	**0.127** (**0.052**) ⋆	**0.119** (**0.048**) ⋆	**0.137** (**0.052**) ⋆	**0.128** (**0.049**) ⋆	0.031 (0.050)	0.022 (0.048)
5	**0.156** (**0.055**) ⋆	**0.147** (**0.053**) ⋆	**0.138** (**0.053**) ⋆	**0.129** (**0.054**) ⋆	**0.130** (**0.054**) ⋆	**0.121** (**0.050**) ⋆
10	**0.113** (**0.052**) ⋆	0.104 (0.058)	**0.120** (**0.051**) ⋆	0.111 (0.057)	0.040 (0.052)	0.032 (0.058)

(a)



(b)



(c)

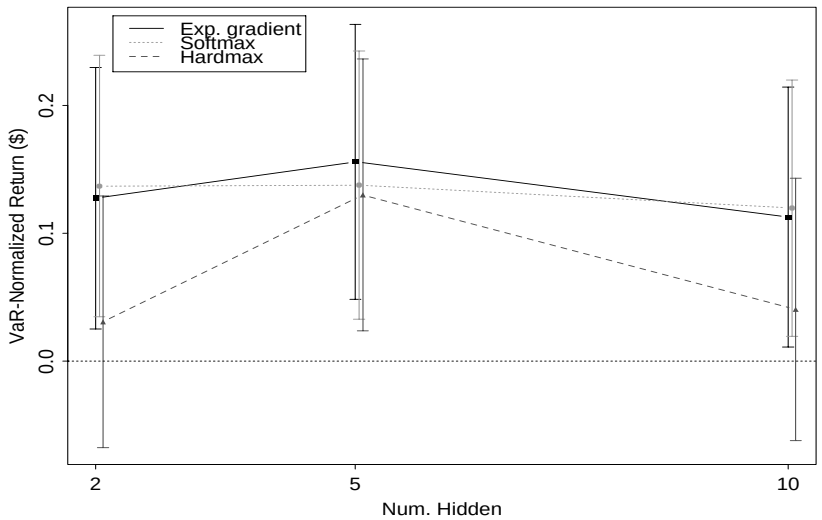


Figure 5.7: *Résultats de trois méthodes de combinaison de modèles, appliqué (a) au modèle de décision sans récurrence, (b) au modèle de décision avec récurrence, (c) au modèle de prévision explicite* QOD_{15} *sans récurrence.*

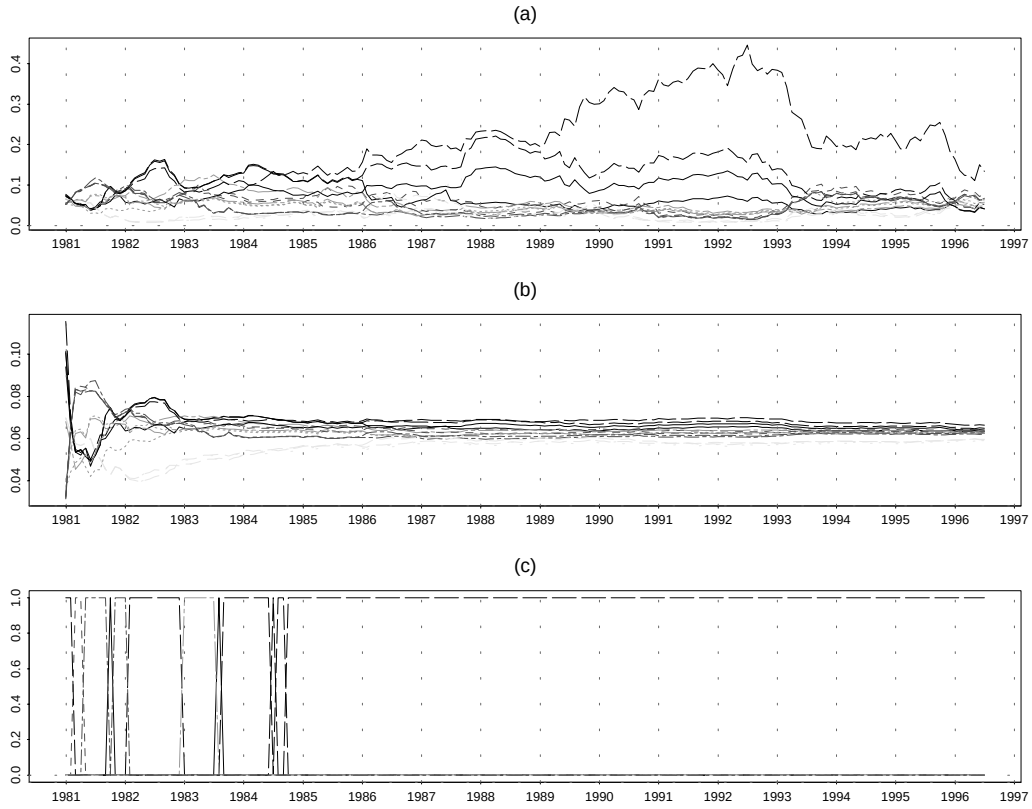


Figure 5.8: *Évolution des pondérations attribuées à chacun des 16 modèles sous-jacents pour le* **modèle de décision sans récurrence** *utilisant un MLP avec 5 unités cachées ; voir les résultats bruts au tableau 5.1. (a) La fonction de combinaison utilise le gradient exponentiel,* $\eta = 0.3, \alpha = 0.01$. *(b) La fonction de combinaison est* softmax. *(c) La fonction de combinaison est* hardmax. *Noter la grande différence des échelles verticales.*

Tableau 5.13: *Résultats de l'*ANOVA *pour le comité combiné selon la méthode du gradient exponentiel, en fonction des facteurs de type de modèle (noté M), du nombre d'unités cachées (noté NC), et de l'interaction entre les deux.*

	Degrés de liberté	Somme des carrés	Moy. des carrés	F-valeur	Pr(F)
M	2	0.9592	0.4796018	1.215027	0.2969649
NC	2	1.0055	0.5027491	1.273669	0.2800756
$M : NC$	4	0.3460	0.0864934	0.219123	0.9278671
Résiduelles	1665	657.2173	0.3947251		

Tableau 5.14: *Résultats de l'*ANOVA *comparant la méthode de combinaison de modèle (*softmax, hardmax, gradient exponentiel*; noté C), le type de modèle (décision avec ou sans récurrence, prévision sans récurrence; noté M), le nombre d'unités cachées (noté NC), et les interactions de second et troisième ordre entre ces facteurs.*

	Degrés de liberté	Somme des carrés	Moy. des carrés	F-valeur	Pr(F)	
C	2	0.673	0.336727	0.862298	0.4222538	
M	2	1.094	0.547144	1.401138	0.2464134	
NC	2	3.365	1.682725	4.309161	**0.0134948**	$\star$
$C : M$	4	0.905	0.226128	0.579073	0.6778177	
$C : NC$	4	0.546	0.136523	0.349611	0.8444505	
$M : NC$	4	0.674	0.168479	0.431446	0.7860204	
$C : M : NC$	8	0.302	0.037735	0.096633	0.9993146	
Résiduelles	4995	1950.545	0.390499			

5.3.2 Analyse

Les tableaux 5.13 et 5.14 analysent formellement l'effet des méthodes de combinaison de modèles.

D'une part, nous ne notons (tableau 5.13) l'impact significatif d'*aucun facteur* (type de modèle ou nombre d'unités cachées) sur la performance de la combinaison par gradient exponentiel.

D'autre part, le contraste de toutes les méthodes de combinaison (tableau 5.14) révèle l'impact général significatif du nombre d'unités cachées, mais pas des autres facteurs. Comme le laisse deviner la figure 5.7, cet effet du nombre d'unités cachées semble être la manifestation de la relative faiblesse de la méthode de combinaison hardmax, même si aucune évidence statistique directe ne peut confirmer cette conjecture. Les deux autres méthodes

de combinaison—softmax et gradient exponentiel—ont une performance statistiquement équivalente.

5.3.3 Gradient exponentiel contre sous-jacents

Nous comparons maintenant les modèles formés par le comité (utilisant la méthode du gradient exponentiel) avec la performance des *meilleurs* modèles sous-jacents, et la *performance moyenne* des modèles sous-jacents, pour tous les types de modèle et nombre d'unités cachées.

Le tableau 5.15 indique quel modèle sous-jacent respectif a obtenu la meilleure performance (à posteriori) pour chaque comité, et présente la différence moyenne entre la performance du comité obtenu par le gradient exponentiel (notée x) et la performance de ce meilleur sous-jacent (notée y). Même si le comité souffre, en général, d'une légère dégradation de la performance par rapport au meilleur sous-jacent, cette différence n'est, pour aucun cas, statistiquement significative. (Nous notons, par ailleurs, que ce meilleur sous-jacent ne peut jamais être utilisé par lui-même, car sa performance n'est pas un estimé non-biaisé de l'erreur de généralisation espérée.)

Le tableau 5.16 donne les résultats de la moyenne des performances des sous-jacents (notée y) et la compare à la performance des comités respectifs (notée x). Nous notons que la performance du comité est statistiquement significativement meilleure dans quatre cas sur neuf, et « presque significativement » meilleure dans deux autres cas.[3] Nous observons que de comparer un comité à la performance moyenne de ses sous-jacents équivaut à le comparer à l'un de ses sous-jacents choisi au hasard.

Nous concluons de ces résultats que, contrairement à leurs pendants humains, les comités de modèles peuvent être significativement plus intelligents que l'un de leurs membres tiré au hasard, et qu'ils ne sont jamais (selon nos expériences) significativement pires que le *meilleur* de leurs membres.

[3]Les comparaisons multiples peuvent biaiser légèrement ces résultats, sans affecter, nous croyons, les conclusions.

Tableau 5.15: *Pour chaque type de modèle, comparaison entre la performance du comité obtenu par la méthode du gradient exponentiel (notée x) et la performance du meilleur sous-jacent participant au comité (notée y). « DDL » veut dire* degrés de liberté.

Modèle		Meilleur sous-jacent	$x-y$ échant.	t-valeur	DDL	Pr(t)
Décision sans récur.	NC=2	WD=10^{-1}, ID=10^{-1}	−0.0336	−0.77	185	0.43
	NC=5	WD=10^{0} , ID=10^{-3}	−0.0326	−1.65	185	0.10
	NC=10	WD=10^{-3}, ID=10^{-1}	−0.0186	−0.82	185	0.41
Décision récur.	NC=2	WD=10^{0} , ID=10^{-3}	−0.0242	−0.71	185	0.47
	NC=5	WD=10^{-3}, ID=10^{-2}	−0.0009	−0.05	185	0.95
	NC=10	WD=10^{-3}, ID=10^{-3}	−0.0235	−1.59	185	0.11
Prévision sans récur.	NC=2	WD=10^{-2}, ID=10^{-3}	0.0054	0.16	185	0.86
	NC=5	WD=10^{-2}, ID=10^{-1}	−0.0067	−0.22	185	0.82
	NC=10	WD=10^{-3}, ID=10^{-1}	−0.0154	−0.46	185	0.64

Tableau 5.16: *Pour chaque type de modèle, comparaison entre la performance du comité obtenu par la méthode du gradient exponentiel (notée x) et la moyenne arithmétique des performances des sous-jacents participant au comité (notée y).*

Modèle		Moyenne des sous-jacents	$x-y$ échant.	t-valeur	DDL	Pr(t)	
Décision sans récur.	NC=2	0.0329 (0.0329)	−0.0124	−0.539	185	0.5906	
	NC=5	0.0777 (0.0342)	0.0258	1.595	185	0.1123	
	NC=10	0.0704 (0.0395)	0.0163	1.834	185	0.0682	
Décision récur.	NC=2	0.0433 (0.0295)	0.0253	1.106	185	0.2698	
	NC=5	0.0868 (0.0321)	0.0490	3.156	185	**0.0019**	⋆
	NC=10	0.0569 (0.0378)	0.0320	2.653	185	**0.0087**	⋆
Prévision sans récur.	NC=2	0.0948 (0.0423)	0.0295	2.008	185	**0.0461**	⋆
	NC=5	0.0893 (0.0400)	0.0647	3.471	185	**0.0006**	⋆
	NC=10	0.0790 (0.0400)	0.0307	1.902	185	0.0587	

Chapitre 6

Conclusion

6.1 Contributions théoriques

L'optimisation directe d'un algorithme d'apprentissage par rapport à un critère financier est une voie de recherches dont la popularité est croissante depuis quelques années (voir entre autres Bengio (1997), Choey et Weigend (1997), et Moody et Wu (1997)). Dans ce mémoire, nous avons poussé plus loin l'étude de cette question, dans les directions suivantes.

Nous avons commencé par présenter le cadre d'investissement utilisant la **valeur à risque**, qui prend deux formes complémentaires : d'une part, l'utilisation de la VàR pour contrôler activement le risque auquel un portefeuille est exposé (§ 2.3) ; d'autre part, son emploi dans un critère d'évaluation de la performance financière (§ 2.5). Nous avons introduit un estimateur **non-biaisé en échantillon fini** du facteur de rééchelonnement permettant de transformer une recommandation en un portefeuille respectant une VàR cible (§ 2.4.

Nous avons ensuite présenté les deux paradigmes principaux de gestion de portefeuille, soit le modèle de **prévision** (§ 3.3) et le modèle de **décision** (§ 3.4). Pour les deux modèles, nous avons expliqué la dérivation des équations

de rétropropagation du gradient qui permettent l'utilisation transparente d'algorithmes d'apprentissage basés sur la descente du gradient. Nous avons de plus examiné les **difficultés numériques** que pose l'utilisation d'un critère financier basé sur le contrôle de la VàR (§ 3.4.3), et proposé une solution basée sur la **régularisation** de la fonction de coût.

Nous avons examiné les outils modernes de **combinaison de modèles** (§ 4.4), et les avons appliqués au problème du choix systématique des hyperparamètres dans l'entraînement de réseaux de neurones.

Nous avons démontré la viabilité de la **pénalisation sur la norme des entrées** (§ 4.3) comme méthode de contrôle de la capacité dans un réseau de neurones, et sa grande utilité dans le cas où le réseau comporte un grand nombre d'entrées.

6.2 Contributions expérimentales

D'un point de vue **expérimental**, nous tirons les conclusions suivantes :

Comparaison entre les modèles (§ 5.2) Nous avons obtenu, de façon consistante, des modèles qui **battent le marché de façon statistiquement significative**, en performance de généralisation, sur une tâche difficile d'allocation entre les 14 secteurs du TSE 300.

En moyennant à l'échelle de tous les résultats bruts des expériences, nous observons que le modèle de prévision utilisant l'allocation quadratique QOD_{15} est **statistiquement significativement meilleur** que les modèles de décision avec et sans récurrence (*cf.* tableau 5.8).

Nous notons par ailleurs l'importance des hyperparamètres dans la performance du modèle de décision, particulièrement par rapport au nombre d'unités cachées et à la pénalisation sur la norme des entrées. Le modèle de décision semble être **beaucoup plus sensible au choix des hyperparamètres** que le modèle de prévision, comme le révèle une comparaison entre un sous-ensemble des expériences (*cf.* tableau 5.9).

À cause de cette sensibilité, qui influence considérablement la région de viabilité des modèles, nous **ne pouvons tirer de conclusions** définitives sur les qualités des modèles de prévision par rapport aux modèles de décision, tels que comparés dans le présent contexte.

Combinaisons de modèles (§ 5.3) Les méthodes de combinaison de modèles **oblitèrent** les différences significatives entre la performance des modèles sous-jacents. Nous n'observons **pas de différence significative** entre la performance des comités résultant des méthodes de décision ou de prévision (*cf.* tableau 5.14).

Les comités ne sont significativement **jamais pires** que le *meilleur* de leurs sous-jacents (*cf.* tableau 5.15).

Les comités sont souvent significativement **meilleurs** que l'un de leurs membres *choisi au hasard* (*cf.* tableau 5.16).

6.3 Pistes futures

Au-delà des résultats présentement démontrés, nous désirons étendre ce travail pour traiter des aspects suivants :

- Dans le modèle de prévision, nous nous sommes restreints à une allocation moyenne–variance, qui suppose une utilité quadratique chez l'investisseur. Nous désirons permettre à ce dernier de spécifier des utilités plus générales, qui permettraient d'éviter, par exemple, les changements brusques dans le portefeuille (« churn »), ou de spécifier des formes différentes d'aversion au risque, comme l'aversion de type HARA (*hyperbolic absolute risk aversion*) (Viala et Briys 1995). De plus, nous voudrions permettre à l'investisseur de spécifier exactement les portefeuilles admissibles, en introduisant des contraintes de domaines et des contraintes affines.
- Nous désirons explorer de meilleurs prédicteurs de la distribution des rendements qu'un simple perceptron multi-couches. Une avenue pos-

sible est l'emploi de modèles dynamiques de la distribution, comme les IOHMM (Bengio 1999; Bengio, Lauzon et Ducharme 1999), ou de modèles graphiques comme les réseaux bayesiens (Jordan 1998; Cowell, Dawid, Lauritzen et Spiegelhalter 1999) qui incorporent explicitement les connaissances à priori d'experts (humains) sur le comportement des marchés financiers.

– L'ensemble des entrées (variables explicatives) présentement fourni aux réseaux de neurones se limite à quelques variables techniques. Nous voulons considérer l'emploi d'une plus grande variété d'entrées, incluant des variables fondamentales.

– Tous les résultats indiquent l'importance de la pénalisation de la norme des entrées. Nous souhaitons poursuivre davantage ce critère, et caractériser son efficacité pour les problèmes à un très grand nombre d'entrées (plusieurs centaines), et le comparer à des méthodes classiques de sélection de variables. Nous voulons aussi mieux comprendre son interaction avec la pénalisation sur la norme des poids.

– Nous n'avons touché que sommairement aux méthodes de combinaison de modèles, et les premiers résultats obtenus sont extrêmement positifs. Nous désirons analyser le comportement de ces méthodes (particulièrement le gradient exponentiel) pour les séries non-stationnaires, et développer une méthode automatique de sélection des hyperparamètres (η et α dans l'éq. (4.39)).

– Finalement, nous n'avons pas exploré à fond le modèle de prévision implicite (*cf.* figure 3.8). Nous aimerions comparer de façon sérieuse ce modèle avec les autres.

Références

Ahlburg, D. A. (1992), « A Commentary on Error Measures », *International Journal of Forecasting 8*, p. 99–111.

Armstrong, J. S. et F. Collopy (1992), « Error Measures for Generalizing About Forecasting Methods : Empirical Comparisons », *International Journal of Forecasting 8*, p. 69–80.

Bellman, R. (1957), *Dynamic Programming*, NJ : Princeton University Press.

Bengio, Y. (1997), « Training a Neural Network with a Financial Criterion Rather Than a Prediction Criterion », Voir Weigend, Abu-Mostafa et Refenes (1997).

Bengio, Y. (1999), « Markovian Models for Sequential Data », *Neural Computing Surveys*.

Bengio, Y. et C. Dugas (1999), « Learning Simple Non-Stationarities with Hyper-Parameters », Rapport technique 1145, Département d'informatique et recherche opérationnelle, Université de Montréal.

Bengio, Y., V.-P. Lauzon et R. Ducharme (1999), « Experiments on the Application of IOHMMs to Model Financial Returns Series », Rapport technique 1146, Département d'informatique et recherche opérationnelle, Université de Montréal.

Bishop, C. (1995), *Neural Networks for Pattern Recognition*, London, UK : Oxford University Press.

Black, F. et M. Scholes (1973), « The Pricing of Options and Corporate Liabilities », *Journal of Political Economy 81*, p. 637–654.

Bodie, Z., A. Kane et A. J. Marcus (1996), *Investments* (Third ed.)., Boston, MA : Irwin McGraw-Hill.

Bollerslev, T. (1986), « Generalized Autoregressive Conditional Heteroskedasticity », *Journal of Econometrics 31*, p. 307–327.

Box, G. P., W. G. Hunter et J. S. Hunter (1978), *Statistics for Experimenters : An Introduction to Design, Data Analysis, and Model Building*, New York, NY : John Wiley & Sons.

Brown, R. G. (1962), *Smoothing, Forecasting and Prediction of Discrete Time Series*, Englewood Cliffs, N.J. : Prentice-Hall.

Campbell, J. Y., A. W. Lo et A. C. MacKinlay (1997), *The Econometrics of Financial Markets*, Princeton, NJ : Princeton University Press.

Choey, M. et A. S. Weigend (1997), « Nonlinear Trading Models Through Sharpe Ratio Maximization », Voir Weigend, Abu-Mostafa et Refenes (1997).

Cowell, R. G., A. P. Dawid, S. L. Lauritzen et D. J. Spiegelhalter (1999), *Probabilistic Networks and Expert Systems*, Statistics for Engineering and Information Science. Berlin : Springer–Verlag.

Cox, D. (1961), « Prediction by Exponentially Weighted Moving Average and Related Methods », *Journal of the Royal Statistical Society, Series B 23*, p. 414–422.

Engle, R. (1982), « Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of UK Inflation », *Econometrica 50*, p. 987–1008.

Figlewski, S. (1994), « Forecasting Volatility using Historical Data », Working Paper S–94–13, New York University.

Figlewski, S. (1997), « Forecasting Volatility », *Financial Markets, Institutions, and Instruments 6*(1), Blackwell Publishers.

FILDES, R. (1992), « The Evaluation of Extrapolative Forecasting Methods », *International Journal of Forecasting 8*, p. 81–98.

FISHMAN, G. S. (1996), *Monte Carlo : Concepts, Algorithms, and Applications* (Second ed.)., Springer Series in Operations Research. Berlin : Springer-Verlag.

FLETCHER, R. (1987), *Practical Methods of Optimization* (Second ed.)., New York : John Wiley & Sons.

GEMAN, S., E. BIENENSTOCK et R. DOURSAT (1992), « Neural Networks and the Bias/Variance Dilemma », *Neural Computation 4*(1), p. 1–58.

GOURIEROUX, C. et A. MONFORT (1997), *Time Series and Dynamic Models*, Cambridge, UK : Cambridge University Press.

GRAHAM, R., D. KNUTH et O. PATASHNIK (1994), *Concrete Mathematics : A Foundation for Computer Science* (Second ed.)., Reading, MA : Addison-Wesley.

HERBSTER, M. et M. K. WARMUTH (1998), « Tracking the Best Expert », *Machine Learning 32*(2).

HINTON, G. (1987), « Learning translation invariant in massively parallel networks », *Proceedings of PARLE Conference on Parallel Architectures and Languages Europe*, Berlin, Springer-Verlag, p. 1–13.

HORNIK, K., M. STINCHCOMBE et H. WHITE (1989), « Multilayer Feedforward Networks Are Universal Approximators », *Neural Networks 2*, p. 359–366.

HULL, J. C. (1999), *Options, Futures, and Other Derivatives* (Fourth ed.)., Englewood Cliffs, NJ : Prentice-Hall.

JORDAN, M. I. (Édit.) (1998), *Learning in Graphical Models*, Adaptive Computation and Machine Learning. Cambridge, MA : MIT Press.

JORION, P. (1997), *Value at Risk : The New Benchmark for Controlling Market Risk*. Irwin McGraw-Hill.

KIVINEN, J. et M. K. WARMUTH (1997), « Additive Versus Exponentiated Gradient Updates for Linear Prediction », *Information and Computation 132*(1), p. 1–64.

KROGH, A. et J. VEDELSBY (1995), « Neural network ensembles, cross validation and active learning », *Advances in Neural Information Processing Systems 7*, Cambridge MA : MIT Press, p. 231–238.

LARSEN, J., C. SVARER, L. N. ANDERSEN et L. K. HANSEN (1998), « Adaptive Regularization in Neural Network Modeling », Voir ORR et MÜLLER (1998).

L'ÉCUYER, P. (1990, Nov. 1990), « A Unified View of the IPA, SF, and LR Gradient Estimation Techniques », *Management Science 36*(11), p. 1364–1383.

LECUN, Y., L. BOTTOU, G. ORR et K.-R. MÜLLER (1998), « Efficient BackProp », Voir ORR et MÜLLER (1998).

MAGDON-ISMAIL, M. et Y. S. ABU-MOSTAFA (1997), « Systematic Underprediction of Volatility in Maximum Likelihood Methods », Voir WEIGEND, ABU-MOSTAFA et REFENES (1997).

MAGDON-ISMAIL, M. et A. ATIYA (1998), « Neural Networks for Density Estimation », *Advances in Neural Information Processing Systems*,

MARKOWITZ, H. M. (1952), « Portfolio Selection », *The Journal of Finance 7*(1), p. 77–91.

MARKOWITZ, H. M. (1959), *Portfolio Selection : Efficient Diversification of Investments*. John Wiley & Sons.

MARKOWITZ, H. M. (1987), *Mean–Variance Analysis in Portfolio Choice and Capital Markets*, Cambridge, MA : Basil Blackwell.

MCCLELLAND, J. et D. RUMELHART (1986), *Parallel Distributed Processing : Explorations in the Microstructure of Cognition*, Volume 1–2, Cambridge : MIT Press.

MOODY, J. (1998), « Forecasting the Economy with Neural Nets : A Survey of Challenges and Solutions », Voir ORR et MÜLLER (1998).

MOODY, J. et L. WU (1997), « Optimization of Trading Systems and Portfolios », Voir WEIGEND, ABU-MOSTAFA et REFENES (1997).

NARENDRA, P. M. et K. FUKUNAGA (1977), « A Branch and Bound Algorithm for Feature Subset Selection », *IEEE Transactions on Computers 26*(9), p. 917–922.

NEUNEIER, R. et H. G. ZIMMERMANN (1998), « How to Train Neural Networks », Voir ORR et MÜLLER (1998).

ORR, G. B. et K.-R. MÜLLER (Édits.) (1998), *Neural Networks : Tricks of the Trade*, Berlin. Springer–Verlag.

PERRONE, M. P. (1993, may), *Improving Regression Estimation : Averaging Methods for Variance Reduction with Extensions to General Conve Measure Optimization*, Ph. D. thesis, Brown University, Institute for Brain and Neural Systems.

PERRONE, M. P. et L. N. COOPER (1993), *Artificial Neural Networks for Speech and Vision*, Chapter When networks disagree : ensemble methods for hybrid neural networks, p. 126–142, London : Chapman & Hall.

PRECHELT, L. (1998), « Early Stopping—But When ? », Voir ORR et MÜLLER (1998).

PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY et W. T. VETTERLING (1992), *Numerical Recipes in C : The Art of Scientific Computing* (Second ed.)., Cambridge : Cambridge University Press.

RISKMETRICS (1996), « RiskMetrics—Technical Document », Rapport technique, J.P. Morgan, New York, NY, `http ://www.riskmetrics.com`.

RUBINSTEIN, R. (1989), « Sensitivity Analysis and Performance Extrapolation for Computer Simulation Models », *Operations Research 37*(1), p. 72–81.

SCHEFFÉ, H. (1959), *The Analysis of Variance*, New York, NY : John Wiley & Sons.

SHARPE, W. F. (1966, January), « Mutual Fund Performance », *Journal of Business*, p. 119–138.

SHARPE, W. F. (1994), « The Sharpe Ratio », *The Journal of Portfolio Management 21*(1), p. 49–58.

SILVERMAN, B. (1986), *Density Estimation for Statistics and Data Analysis*, London : Chapman & Hall.

SIMONOFF, J. (1996), *Smoothing Methods in Statistics*, New York : Springer.

STONE, M. (1974), « Cross-validatory choice and assessment of statistical predictions », *Journal of the Royal Statistical Society, B 36*(1), p. 111–147.

STUART, A., J. K. ORD et S. ARNOLD (1998), *Kendall's Advanced Theory of Statistics* (Sixth ed.), Volume 2A : Classical Inference and the Linear Model. Edward Arnold.

VAPNIK, V. (1998), *Statistical Learning Theory*. Wiley.

VIALA, P. et . BRIYS (1995), *Éléments de théorie financière*. Nathan.

WEIGEND, A. et N. GERSHENFELD (Édits.) (1993), *Time Series Prediction : Forecasting the future and understanding the past*. Addison-Wesley.

WEIGEND, A. S., Y. ABU-MOSTAFA et A.-P. REFENES (Édits.) (1997), *Decision Technologies for Financial Engineering : Prooceedings of the Fourth International Conference on Neural Networks in the Capital Markets (NNCM '96)*. World Scientific Publishing.

WEIGEND, A. S., D. E. RUMELHART et B. A. HUBERMAN (1991), « Back-Propagation, Weight-Elimination and Time Series Prediction », *Connectionist Models : Proceedings of the 1990 Summer School*, San Mateo, CA, Morgan Kaufmann,