Évaluation et validation de prévisions en loi

Michael Richard

Centre Val de Loire UNIVERSITÉ
Spirit of the Loire Valley

UNIVERSITÉ D'ORLÉANS

UNIVERSITE D'ORLEANS

ÉCOLE DOCTORALE DES SCIENCES DE L'HOMME ET DE LA SOCIÉTÉ LABORATOIRE D'ÉCONOMIE D'ORLÉANS

Thèse présentée par :

Michael RICHARD

soutenue le : **09 Mai 2019**

pour obtenir le grade de : **Docteur de l'Université d'Orléans**

Discipline/ Spécialité : **ÉCONOMIE**

Évaluation et validation de prévisions en loi

Thèse dirigée par :

Jérôme COLLET	Ingénieur-chercheur à EdF R&D
Christophe HURLIN	Professeur de l'université d'Orléans

RAPPORTEURS :

Olivier DARNÉ	Professeur de l'université de Nantes
Peter TANKOV	Professeur à l'Ensae ParisTech

JURY :

Christophe RAULT	Professeur de l'université d'Orléans, Président du jury
Jérôme COLLET	Ingénieur-chercheur à EdF R&D
Olivier DARNÉ	Professeur de l'université de Nantes
Yannig GOUDE	Ingénieur-chercheur à EdF R&D
Christophe HURLIN	Professeur de l'université d'Orléans
Peter TANKOV	Professeur à l'Ensae ParisTech

Remerciements

Après de nombreuses heures de rédaction, il est temps d'apporter la touche finale à ce manuscrit, par le biais de ces remerciements. J'espère n'oublier personne, mais lorsque l'on a la chance d'être aussi entouré que je le suis, ce n'est pas chose facile.

Je commencerais par remercier Jérôme Collet, mon encadrant industriel au sein de EdF et Christophe Hurlin, mon encadrant académique. Ce fut un plaisir de travailler avec vous durant ces trois années, aussi bien sur le plan professionnel que personnel. J'ai appris beaucoup dans le domaine de la statistique et du machine learning, ainsi que sur la façon de mener à bien un projet de recherche, grâce à vos conseils, vos idées et vos nombreux retours, non dénués d'humour. Ce manuscrit n'aurait pas vu le jour sans vous, et je vous en suis très reconnaissant.
Merci également à Olivier, Yannig Clémence et Pierre G. qui ont suivis mes travaux durant ces trois ans à EdF.

Je souhaiterais ensuite remercier Olivier Darné et Peter Tankov qui ont eu la gentillesse d'accepter d'être mes rapporteurs, malgré des emplois du temps que j'imagine très chargés.

Un grand merci également aux membres du LEO, qui m'ont toujours accueilli chaleureusement lors de mes venues à Orléans. En particulier, je pense à Cécile Chamaillard pour sa grande efficacité, à Aziz N'Doye avec qui j'ai partagé un bureau quelques semaines durant, ainsi qu'à Jérémy et Hajare qui m'ont beaucoup aidé.

J'ai ensuite une pensée pour mes amis rencontrés au fil de mes péripéties mathématiques à l'université Paris VII : Anaïs, Olivier, Tania, Patrick, Pei Xia, Ducay, Liantsoa, Christophe M., Mina, Véronique, Jennifer, Fabien, Tristan, Christophe C., Fred et Ric, Elias, Richard, Japhet, Jo, Reda, Daniel, Mohamed, Julie, Jules, Ouarda, Sarah, Truc-Mai, Firdevs, Florian, Sopheary, Guillaume, Andrew, Dilek et Amandine. Certains étaient là depuis le début, d'autres sont arrivés en cours de route, mais chacun de vous à contribué à faire de moi le future Docteur que je m'apprête à devenir. Merci aussi à Jérôme B., Chaneb et Odile, pour m'avoir permis de travailler dans de bonnes conditions à la fin de la thèse ou lorsque j'étais à Campuac.

Je me dois bien entendu de remercier également la R&D de EdF pour m'avoir accueilli durant ces trois années de doctorat, ainsi que les personnes que j'ai eu la chance d'y rencontrer. Tout d'abord, j'aimerais remercier les membres du groupes R39 pour leur accueil et les bon moments passés avec eux. Cyrille, je ne me lasserais jamais de nos discussions sur le cerveau et le sommeil, ni des sessions improvisées dans la salle de musique. De la même façon, je pourrais passer des heures à parler de musique avec Vincent ou à l'écouter nous raconter son passé de jeune cancre. Je voudrais aussi saluer Audrey et son énergie débordante dès 7h du matin, pour son efficacité, son investissement et sa bienveillance. David, Yohann, Gilbert, Dominique et Virgile, les pronostics et débriefs des matchs du PSG, ainsi que les nombreuses blagues que l'on a pu se raconter vont me manquer. Je terminerais en remerciant Thi Thu et Sébastien pour leurs conseils et les nombreux moments de déconne que l'on a partagé, ainsi que Bayram avec qui j'ai passé de très bons moments lorsque nous partagions le même bureau.

Je remercie également les membres du groupe R33. C'était un plaisir de passer du temps avec vous lors de pauses café ou de déjeuners. J'ai beaucoup apprécié, entre autres, les discussions sur des sujets très variés tels que top chef ou le football avec Guillaume, Pascal, Benoît, Isaque et Bruno, sans oublier Michel (et ses piques sur le PSG).

Je n'oublie pas non plus mes camarades du groupe R36 avec qui j'ai partagé beaucoup de très bons moments : Hugo, Olivier, Cécile, Maxime, Paulin, Rebecca et Rodolphe. En particulier, un grand merci au dernier mentionné, avec qui j'ai traversé l'intense épreuve de la rédaction de fin de thèse.

Je continuerais maintenant en adressant mes remerciements à Bérénice pour les nombreux trajets passés en sa compagnie et celle de ses éclats de rire, pour les cours de Lingala et son soutien en cette fin de thèse, ainsi qu'aux amis de passage (de passage à EdF, il va sans dire) qui ont laissé un grand

vide une fois partis : merci à Eliette et Thiziri qui ont été très présentes durant les six derniers mois, et avec qui j'ai passé d'excellents moments, surtout lorsque l'on s'est retrouvés dans le même bureau. Enfin merci à mon marseillais préféré, Steven, ainsi qu'à Patrick et Djiby. Votre soutien, vos encouragements et les nombreux délires en votre compagnie m'ont beaucoup aidé durant les moments difficiles, et je vous en suis reconnaissant.

Pour terminer de remercier ceux qui ont égayé mes journées à Edf, j'aimerais exprimer mon attachement au groupe R32, là où tout à commencé. Un groupe aux personnalités très diverses, tant sur le plan personnel que des activités professionnelles, au sein duquel j'ai tout de suite été intégré. Je ne pouvais pas rêver meilleures conditions pour démarrer la thèse. Je garderais en mémoire beaucoup de bon souvenirs. Entre autres, les taquineries de Slimane, les punchlines bien senties mais toujours teintées d'affection de Guillaume et Pascale, la gentillesse et les bons conseils de Marie, les goûters fromagers avec Christian ou encore les cours de chinois dispensés par Wenkai et Huang Shen. Je finirais en remerciant Thomas, Pierre C. et Pierre G. à nouveau, qui ont été très présents dans les bons comme dans les mauvais moments, tout comme Emma, Anne-Laure et Yan Hui, véritables éclaircies quand le ciel se fait menaçant. Je pourrais écrire énormément à leur sujet pour tout ce qu'ils m'ont apporté. Mais je crois qu'il ne serait pas très approprié d'avoir des remerciements plus long que la thèse en elle même. Je vais donc essayer de faire court, d'autant plus je n'aurais de toute façon jamais assez de mots pour leur exprimer toute ma gratitude. Merci pour votre soutien, votre écoute, ces nombreux conseils, ces moments passés autour d'un verre, d'un café ou d'un repas, avec des discussions parfois absurdes mais tellement drôles, les citations du grand détournement et de Kaamelott (Animaux de la forêt !), les karaokés improvisés et autres pauses musicales, les point C++ ou encore les challenges EdF, pour ne citer qu'une infime partie des très nombreux moments passés à vos cotés.

Pour clôturer ces remerciements, les derniers iront à ma famille, et plus particulièrement à ma mère, mes sœurs et leurs maris, mon oncle et ma tante, avec qui je partageais mes trajets matinaux, à ma deuxième famille Italienne, et à Ruud, Tony et Pascal en particulier et à mes amis de toujours Mickael et Jonathan, ainsi qu'à Géraldine et Émeline. Vous y êtes pour beaucoup dans l'achèvement de ce long projet, durant lequel j'ai toujours pu compter sur votre soutien.

Contents

Liste des figures iv

Liste des tables v

1 Introduction 1
- 1.1 Introduction . . . 1
- 1.2 Notions Préliminaires . . . 2
- 1.3 Chapitre 2 : Machine Learning et prévision quantile . . . 3
- 1.4 Chapitre 3 : Tests de validation de prévisions en loi . . . 5
- 1.5 Chapitre 4 : Recalibration des prévisions en loi . . . 7

2 Machine learning and quantile forecasts 11
- 2.1 Introduction . . . 11
- 2.2 Quantile Forecasting and Machine Learning . . . 11
  - 2.2.1 Support Vector Quantile Regression . . . 12
  - 2.2.2 Quantile Regression Neural Network . . . 14
  - 2.2.3 Additive Quantile Models . . . 15
  - 2.2.4 Quantile Regression Forest . . . 17
  - 2.2.5 Quantile Forecast with Gradient Boosting Machine . . . 17
- 2.3 Data and Empirical Framework . . . 18
- 2.4 Conclusion . . . 22

3 Validation de prévisions en loi 25
- 3.1 Introduction . . . 25
- 3.2 Techniques de prévisions de densité . . . 27
- 3.3 Évaluation des prévisions de densité . . . 28
  - 3.3.1 Les tests statiques de spécification correcte . . . 29
  - 3.3.2 Les tests de spécification dynamique correcte . . . 33
- 3.4 Application . . . 39

4 A generic method for density forecast recalibration 47
- 4.1 Introduction . . . 47
- 4.2 Principle of the method . . . 48
- 4.3 Impact on global score . . . 49
  - 4.3.1 Impact on score: conditions for improvement . . . 49
  - 4.3.2 Impact on score: bounds on degradation . . . 50
- 4.4 Case study . . . 51

5 Conclusion et perspectives 55

A Recalibration on electricity price ensemble forecasts 57

B Appendix chapter 2 61

C Annexe chapitre 3 63

D Appendix chapter 4 **71**
D.1 Impact on score: conditions for improvement 72
D.1.1 Rewriting the difference of L_τ expectation 72
D.1.2 Systematic improvement of the quality 74
D.2 Impact on score: bounds on degradation 75

Bibliography **81**

List of Figures

1.1 *Pinball-Loss* function (from Takeuchi et al., 2006). 3

2.1 The soft margin loss setting for a linear SVM (from Schölkopf and Smola, 2002). 13

3.1 Exemple de prévision de densité : graphique d'évolution possible de l'inflation, Novembre 2017 (Banque d'Angleterre) 27

4.1 Comparison of *EMOS* and our method *CRPS* expectation with that of raw ensemble The plain line corresponds to our method and the dashed one to the *EMOS*. 52

A.1 Comparison of raw OPUS ensembles and recalibrated ones cash-flows (left) and of raw OPUS-PAF ensembles and recalibrated ones cash-flows (right) with the real cash-flows, in absolute value. The plain line corresponds to the raw ensembles and the dashed one to our method. 59

C.1 Rendements logarithmiques des prix à la clôture des trois indices sur la période 01/06/04-01/06/12 65

C.2 Premières prévisions de densité de chaque environnement avec une modélisation GARCH(1,1) pour les résidus 68

C.3 Premières prévisions de densité de chaque environnement avec une modélisation GJR-GARCH(1,1) pour les résidus 69

List of Tables

2.1 Descriptive statistics of the dependent variable . . . 19
2.2 Hyper-parameters selected for mcycle. . . . 20
2.3 Hyper-parameters selected for BostonHousing. . . . 20
2.4 Hyper-parameters selected for GEFCom2014. . . . 21
2.5 Mean of *Pinball-Loss* score for mcycle . . . 21
2.6 Mean of *Pinball-Loss* score BostonHousing . . . 21
2.7 Mean of *Pinball-Loss* score GEFCom2014 . . . 22

3.1 Dates de la première et de la dernière période de chaque environnement . . . 40
3.2 Statistiques descriptives des séries des rendements logarithmiques pour chaque période . 40
3.3 Pourcentages de rejet de l'hypothèse nulle au risque 5% et 1% pour les prévisions à horizon $h = 1$ jour avec modèle GARCH(1,1) pour les résidus. . . . 42
3.4 Pourcentages de rejet de l'hypothèse nulle au risque 5% et 1% pour les prévisions à horizon $h = 10$ jours avec modèle GARCH(1,1) pour les résidus. . . . 43
3.5 Pourcentages de rejet de l'hypothèse nulle au risque 5% et 1% pour les prévisions à horizon $h = 1$ jour avec modèle GJR-GARCH(1,1) pour les résidus. . . . 44
3.6 Pourcentages de rejet de l'hypothèse nulle au risque 5% et 1% pour les prévisions à horizon $h = 10$ jours avec modèle GJR-GARCH(1,1) pour les résidus. . . . 45

4.1 Success rate to 5% K-S test . . . 52

A.1 Forecasting dates for each output . . . 57
A.2 Percentages of raw PIT series passing the Kolmogorov-Smirnov test at level 5% . . . 58
A.3 Percentages of obsrvations out of the ensemble bounds . . . 58
A.4 *Pinball-Loss* score . . . 58

B.1 Seed setting. . . . 61
B.2 Grids search for mcycle. . . . 62
B.3 Grids search for BostonHousing. . . . 62
B.4 Grids search for GEFCom2014. . . . 62

C.1 Paramètres du processus ARMA(3,3) avec modèle GARCH(1,1) pour les résidus . . . 63
C.2 Paramètres du modèle GARCH(1,1) . . . 64
C.3 Ecart-types des estimations des paramètres du processus ARMA(3,3) avec modèle GARCH(1,1) pour les résidus . . . 64
C.4 Ecarts-types des estimations des paramètres du modèle GARCH(1,1) . . . 64
C.5 Paramètres du processus ARMA(3,3) avec modèle GJR-GARCH(1,1) pour les résidus . . 66
C.6 Paramètres du modèle GJR-GARCH(1,1) . . . 66
C.7 Ecart-types des estimations des paramètres du processus ARMA(3,3) avec modèle GJR-GARCH(1,1) pour les résidus . . . 66
C.8 Ecart-types des estimations des paramètres du modèle GJR-GARCH(1,1) . . . 67

Chapter 1

Introduction

1.1 Introduction

Dans ces travaux, nous étudions un type de prévisions de plus en plus utilisées, à savoir les prévisions en loi. Nous nous intéressons à la distribution conditionnelle d'une variable aléatoire Y définie sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$, ou bien muni d'une filtration $(\mathcal{F}_t)_{t\geq 0}$ lorsque l'on manipule des séries temporelles. Il existe principalement deux domaines dans lesquels les prévisions en loi sont très utilisées : le secteur économique et celui de la météorologie, qui serviront à illustrer les différents résultats présentés dans cette thèse.

Les prévisions en loi décrivent donc l'ensemble de la distribution, ce qui permet entre autres, de mieux quantifier l'incertitude autour des prévisions et également de mettre en évidence des changements de régime et des asymétries dans les distributions. Ces prévisions sont donc en opposition avec les prévisions ponctuelles et les prévisions par intervalles. Les prévisions ponctuelles ne fournissent qu'une estimation de la moyenne de la distribution sans se préoccuper de l'incertitude autours des prévisions. Les prévisions par intervalles, quant à elles, se présentent sous la forme d'un intervalle qui définit l'ensemble des valeurs que peut prendre une nouvelle observation à un certain niveau de certitude défini au préalable. Par exemple, un intervalle à $(1-\alpha)\%$ signifie que la réalisation de la variable se situera dans cet intervalle avec une probabilité de $(1-\alpha)\%$. Dans cette thèse, nous nous penchons plus particulièrement sur l'évaluation et la validation des prévisions de densités. En effet, travailler avec des prévisions en loi implique un renouvellement des méthodes d'évaluation et de validation des prévisions dans la mesure où les tests proposés dans un cadre de prévisions ponctuelles ou par intervalles peuvent ne pas s'appliquer. Les tests relatifs aux prévisions ponctuelles sont en général basés sur la distance entre la vraie valeur des observations et la prévision, quand les tests de prévisions par intervalles reposent sur l'étude du processus de *violation* qui prend la valeur 1 si l'observation n'appartient pas à l'intervalle de prévision et 0 sinon. Ainsi, ces tests ne peuvent pas être appliqués directement à des prévisions sous forme de distribution.

Il existe deux moyens de vérifier la validité et la qualité des prévisions en loi, comme le souligne Gneiting et al. (2007, 2014). La première possibilité consiste à les évaluer séparément. Une prévision de densité sera dite valide si elle est bien calibrée, c'est à dire que la variable *PIT* (*Probability Integral Transform*, Rosenblatt, 1952) $F(Y)$, où Y est la variable aléatoire continue que l'on étudie et F sa fonction de répartition, suit une distribution uniforme sur l'intervalle [0,1]. Cette propriété est liée à la fois aux prévisions et aux observations. Pour la qualité, on évalue une propriété liée exclusivement aux prévisions, la *sharpness*. Elle fait référence à la concentration de la prévision et plus elle est concentrée, mieux c'est. On utilise les *sharpness diagrams* pour évaluer cette propriété. L'autre solution consiste à employer des règles de scoring propres. A titre d'exemple, on mentionnera la *MSE* dans le cadre d'une prévision ponctuelle, la *Pinball-Loss* pour les prévisions de quantiles ou encore le *CRPS* pour les fonctions de répartition. Ces fonctions sont dites "propres" dans le sens où leur espérance est minimisée par la vraie valeur de la cible que l'on cherche à prévoir. Elles présentent un intérêt certain car elles permettent de vérifier simultanément la validité de la prévision de densité et sa qualité. En effet, Murphy (1973) propose une décomposition du Brier Score (1950), en trois termes : $BS = Reliability + Resolution + Uncertainty$. Le terme *Reliability* fait référence à la calibration, *Resolution* à la *sharpness* et le dernier terme est un

terme d'incertitude. Cette décomposition a ensuite été généralisée à tous les scores par Bröcker (2009). Tout au long de cette thèse, nous avons utilisé la variable *PIT* pour vérifier la propriété de validité des prévisions et les règles de scoring pour la qualité. Il faut y voir une volonté de lier les deux grands domaines d'application des prévisions en loi.

Cette thèse est motivée par l'utilisation grandissante des prévisions de densité et des questionnements que cela entraîne : comment mettre à profit les avancées récentes dans les autres types de prévisions pour faire des prévisions en loi ? Comment juger de telles prévisions ? Ou encore, comment choisir entre plusieurs prévisions en loi ? Autant de questions auxquelles nous allons tenter de répondre au long des trois chapitres de cette thèse.

Dans la première partie, nous nous pencherons sur l'apport du machine learning vis à vis des prévisions en loi. Le machine learning est un champ de l'informatique, également très proche de la statistique et de l'optimisation qui permet, entre autres, de faire de la prévision sans avoir à spécifier un modèle au préalable. Dans le but de tester l'apport de ce domaine émergent, nous avons testé différents algorithmes de machine learning dans un cadre de prévisions de quantiles sur données réelles. Les données à disposition nous permettent d'évaluer les différentes techniques employées dans des environnements divers, ce qui se traduit par un nombre de régresseurs faible, moyen ou important sur des données individuelles ou des séries chronologiques. Nous tenterons ainsi de mettre en évidence l'intérêt de certaines méthodes selon le type de données auxquelles nous sommes confrontés.

La seconde partie de cette thèse nous permettra d'exposer quelques tests de validation de prévisions en loi présents dans la littérature et de mentionner les avantages et les inconvénients qu'ils comportent. Certains de ces tests seront ensuite appliqués sur données réelles relatives aux log-rendements de trois grands indices boursiers (SP500, Dow Jones et Nasdaq) afin d'évaluer la validité des prévisions dans des environnements différents.

Dans la troisième et dernière partie, nous proposerons une méthode permettant de simplifier le choix d'une prévision de densité en particulier par rapport à d'autres. La méthode que nous avons mise en place est une méthode de recalibration des prévisions qui permet d'obtenir des prévisions valides à partir d'un modèle mal spécifié. Ainsi, puisque le critère de validation est automatiquement vérifié une fois la recalibration appliquée, le choix entre différentes prévisions de densité s'en retrouve simplifié. En effet, elle seront jugées en fonction de leur qualité, leur capacité à approcher la cible, que nous mesurerons en général à l'aide d'une fonction de perte définie en lien avec l'usage. Notre méthode sera enfin illustrée sur des données réelles, telles que des scénarios de prix ou encore des scénarios de température.

1.2 Notions Préliminaires

Nous allons ici rappeler quelques notions importantes pour la suite. La variable *PIT*, issue de la transformation proposée par Rosenblatt est définie comme $F(Y)$, où Y est une variable aléatoire et F sa fonction de répartition. Un résultat important est la suivant :

$$\text{Si } Y \sim F, \text{ alors } F(Y) \sim \mathcal{U}[0,1],$$

sous réserve que F soit continue.

Ainsi, on définit la série des *PIT* $\{z_t\}_{t=1}^T = \{F_t(y_t)\}_{t=1}^T$, avec t la date de prévision, F_t la fonction de répartition de la variable Y_t à la date t et y_t sa réalisation à la même date (sauf cas particulier). La série des *PIT* estimés quant à elle est définie comme $\{\widehat{z}_t\}_{t=1}^T = \{G_t(y_t)\}_{t=1}^T$, avec G_t la densité de prévision à la date t.

On appellera également variable *INT* (*Inverse Normale Transformation*) la variable $\Phi^{-1}(z_t)$, où Φ^{-1} représente la réciproque de la fonction de répartition d'une distribution $\mathcal{N}(0,1)$. Cette variable sera utilisée dans certains tests de validité de la section 1.3 du fait que si $z_t \sim \mathcal{U}[0,1]$, alors $\Phi^{-1}(z_t) \sim \mathcal{N}(0,1)$.

Enfin, il convient d'expliciter plus formellement les deux règles de scoring propres que nous utiliserons dans la suite. La *Pinball-Loss* L_τ est une fonction utilisée pour évaluer les prévisions de quantiles et qui pénalise les erreurs de façon asymétrique :

$$L_\tau(y,q) = \tau(y-q)\mathbf{1}_{\{y \geq q\}} + (q-y)(1-\tau)\mathbf{1}_{\{y<q\}}\,,$$

avec y l'observation, q la prévision, $\mathbf{1}_{\{.\}}$ la fonction indicatrice et $\tau \in [0,1]$ le niveau de quantile choisi.

Cette fonction est illustrée par la figure 1.1, avec $\xi = y - q$.

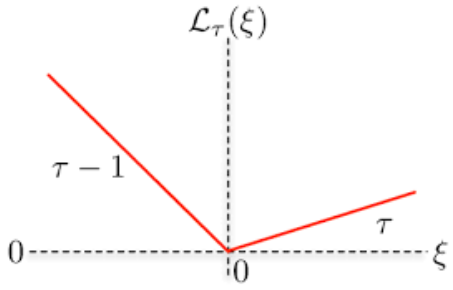

Figure 1.1 – *Pinball-Loss* function (from Takeuchi et al., 2006).

Le *CRPS* (*Continuous Ranked Probability Score*) est défini par :

$$CRPS(G,y) = \int_{-\infty}^{+\infty} (G(x) - \mathbf{1}_{\{y \leq x\}})^2\, dx\,,$$

où G désigne une fonction de répartition et y la réalisation de la variable Y.

Notons que l'on a également la relation suivante :

$$CRPS(G,y) = 2\int_0^1 L_\tau(y, G^{-1}(\tau))\, d\tau\,,$$

ce qui nous permet de généraliser les résultats obtenus d'un score à l'autre.

1.3 Chapitre 2 : Machine Learning et prévision quantile

L'objectif de ce chapitre de la thèse est d'évaluer la contribution des méthodes de type machine learning à la prévision de quantiles. Avant d'évoquer plus en détails les méthodes de machine learning, il convient d'expliquer les différentes manières d'effectuer des prévisions en loi. Une approche évidente de par sa proximité avec les méthodes de prévisions ponctuelles consiste à spécifier un modèle paramétrique avec une hypothèse sur la distribution du terme d'erreur et d'en déduire une distribution pour la variable d'intérêt (méthode du Skeleton). Nous prendrons comme exemple la cas d'une régression linéaire $Y = X\beta^{'} + \epsilon$ avec $\epsilon \sim \mathcal{N}(0,1)$, à partir de laquelle on peut déduire $Y \sim \mathcal{N}(X\beta^t, 1)$. Toutefois, cela ne permet pas de mettre en évidence des asymétries ou des changements de régime, dans la mesure où la densité estimée est liée au terme d'erreur, qui ne reflète pas ces changements.

La modélisation par prévision d'ensemble est également très populaire, notamment pour la modélisation des prix ou des températures, avec de grosses différences. Le schéma européen pour les prévisions météorologiques [1], par exemple, est constitué de 51 scénarios sur la base d'une même distribution, avec des conditions initiales modifiées. En effet, compte tenu du caractère chaotique du système étudié, il suffit d'introduire un bruit sur les conditions initiales pour obtenir des trajectoires divergentes. Ainsi, ces

1. https://www.ecmwf.int/

51 scénarios reflètent l'incertitude autour des prévisions. Dans le cadre de modèles de prix, les scénarios sont en général plus nombreux, et on introduit du bruit tout au long des trajectoires pour obtenir les différents scénarios.

Aussi, on peut voir les prévisions d'ensemble comme des prévisions de quantiles de différents niveaux. Ce qui nous amène à une autre modélisation possible, qui consiste à prévoir toute une succession de quantiles (en général allant de 1% à 99%). La première méthode de prévisions de quantiles, que l'on appelle régression quantile, fut proposée par Koenker et Basset (1978). Avec cette méthode, il n'est plus nécessaire de spécifier des hypothèses sur les deux premiers moments de la distribution. Un autre point important est que l'on peut ainsi voir l'impact différent d'une variable selon le niveau de quantile choisi.

En parallèle des nombreuses avancées en mathématiques, nous avons assisté à l'émergence du machine learning, que l'on classerait dans le domaine de l'informatique, et qui se trouve finalement au confins de la statistique et de l'informatique, tout en étant moteur de nombreuses avancées en optimisation. Les algorithmes de machine learning ont d'abord été proposés afin de résoudre des problèmes de classification, et l'on peut citer, entre autres, les réseaux de neurones de Farley et Clark (1954), les machines à vecteur de support de Vapnik et Chervonenkis (1964) ou encore les arbres de décisions, popularisés par Breiman et al. (1984). Ces différentes méthodes ont ensuite été étendues au cas où les variables d'intérêt ne sont plus binaires mais continues (i.e étendues au cas de la régression) et de nouvelles méthodes telles que les modèles additifs de Friedman et Stuetze (1981), le gradient boosting proposé par Friedman (1999) et les forêts aléatoires développées par Breiman (2001) ont vu le jour en parallèle. Tout naturellement, au même titre que l'on est passé de la régression linéaire à la régression quantile, le machine learning a ensuite été utilisé dans le cadre des prévisions de quantiles.

Le machine learning permet de donner à un ordinateur la capacité d'effectuer certaines tâches sans l'avoir préalablement programmé pour. Dans le cadre de la régression, on peut donc faire de la prévision sans avoir à spécifier d'hypothèses, ce qui fournit des modèles qui sont quasiment exclusivement fondés sur les données à disposition. Cela permet entre autres de mettre en évidence des motifs particuliers ainsi que de prendre en compte des structures complexes contrairement à la régression quantile, qui est utilisée dans un cadre linéaire. Aussi, on assiste depuis ces dernières années à une explosion de la collecte de données, tant sur le nombre que sur le type, ce qui favorise l'utilisation du machine learning. En effet, la variété et le grand nombre de données disponibles posent problèmes lorsque l'on utilise les premières méthodes de prévisions. En revanche, ces questions ne se posent pas lorsque l'on utilise des algorithmes de machine learning. Ces algorithmes nécessitent justement un gros volume de données, peu importe le type (les données satellites ou les images, entre autres, sont très utilisées), et modélisent automatiquement les interactions entre les variables même en très grande dimension, par le biais de paramètres de régularisation, qu'il faut toutefois sélectionner. Ceci nous amène à nous poser les questions suivantes.

Question 1 : La régression par machine learning est-elle plus adaptée que la régression quantile en terme de qualité de prévision, en vue de produire des prévisions d'ensemble ?

Question 2 : Certaines méthodes de machine learning sont-elles nettement plus adaptées que d'autres selon le type de données à disposition ?

Ces questions nous ont amenés à faire un benchmarking de différentes techniques de prévision de quantiles par machine learning avec des données qui nous permettent d'évoluer dans des cadres différents et de les comparer avec des prévisions issues d'une régression quantile. Cette comparaison est effectuée à l'aide de la *Pinball-Loss*.

Pour ce faire, nous avons utilisé des algorithmes de prévision de quantiles appartenant à cinq grandes familles du machine learning, à savoir :

— les machines à vecteurs de support avec une modélisation proposée par Takeuchi et al. (2006)

— les réseaux de neurones tels que présentés par Taylor (2000)

— les forêts aléatoires développées par Meinshausen (2006)

— le gradient boosting utilisé par Landry et al. (2016)

— les modèles additifs par le biais de la méthode Extended Log-F (ELF) de Fasiolo et al. (2017).

Une fois la théorie relative à chaque méthode exposée, nous avons utilisé trois jeux de données pour les comparer dans divers environnements : deux jeux de données individuelles, l'un avec un unique régresseur et l'autre avec de nombreux régresseurs, et un jeu de données de série temporelle, avec un nombre moyen de régresseurs. D'un point de vue algorithmique, nous avons utilisé les fonctions disponibles sous le logiciel R, et avons implémenté des codes qui permettent d'effectuer des prévisions de quantiles successifs lorsque cela n'était pas possible avec les fonctions de base. De plus, nous avons pris soin de mettre en place une procédure de réarrangement des quantiles afin d'éviter le croisement de quantiles, ainsi qu'une méthode de sélection des paramètres par k-fold cross-validation.

Les résultats obtenus nous ont permis de mettre en évidence certaines propriétés et limites des approches par méthodes de machine learning et par régression quantile. Nous avons également constaté que certains algorithmes de machine learning se démarquent des autres selon le type de données à disposition. En effet, les résultats obtenus sur un jeu de données avec un seul régresseur et une variable à prédire avec beaucoup de variance sont peu convaincants. Pour le jeu de données avec de nombreux régresseurs, la méthode des forêts aléatoires tire son épingle du jeu. Les autres algorithmes présentent quant à eux des résultats mitigés selon le quantile étudié. Nous avons également remarqué que les prévisions issues de la régression quantile semblent de bonne qualité pour les quantiles bas. Les résultats obtenus pour les autres quantiles sont en revanche moins convaincants, mais ne sont pas pour autant les moins bon résultats obtenus sur ce jeu de données. Les résultats concernant le jeu de données de série temporelle ont permis de montrer que les méthodes ELF et gradient boosting machine sont les plus adaptées et que la régression quantile dans ce cadre là ne l'est pas du tout. Cela nous laisse penser que l'utilisation des méthodes de machine learning est plus adaptée que la régression quantile pour la prévision de densité.

On notera toutefois qu'il faut prendre ces résultats avec beaucoup précautions car nous n'avons testé ces méthodes que sur trois jeux de données. De plus, le jeu de données de série temporelle contient des variables explicatives qui sont celles sélectionnées par Gaillard et al. (2016) parmi un plus grand nombre de variables, ce qui avantage la méthode ELF. Quand bien même, il est important de remarquer aussi que l'on s'est limité ici à la prévision de six quantiles et non de l'ensemble des quantiles de la distribution, ce qui ne nous permet que de tirer des conclusions partielles quant à l'utilisation du machine learning pour les prévisions de densité, la validité des prévisions ne pouvant être vérifiée ici.

1.4 Chapitre 3 : Tests de validation de prévisions en loi

Dans cette partie, nous proposons une synthèse des différents tests de validation des prévisions en loi présents dans la littérature. Les premiers tests n'étant plus applicables dans ce contexte de par leur construction, il a fallu repenser les tests d'évaluation. En se basant sur la distribution théorique des variables *PIT*, de nombreux tests d'uniformité d'une distribution ont été avancés.
Ainsi, les tests que nous allons présenter dans la suite s'intéressent à la variable suivante (ou une transformation de celle-ci) :

$$z_t = F_t(y_t|\Omega_{t-1}, \theta_0) = \int_{-\infty}^{y_t} f_t(u|\Omega_{t-1}, \theta_0)\, du \ , t = 1 \cdots T,$$

où F_t est la fonction de répartition de la variable Y_t à la date t, f_t sa densité, y_t est la réalisation à la date t, Ω_{t-1} l'ensemble d'information disponible à la date t, θ_0 un vecteur de paramètres et T le nombre de prévisions.

Avant d'évoquer plus en détails les différents tests proposés, il convient d'apporter quelques précisions sur ces derniers. On distingue deux grandes classes de tests. Les tests de spécification correcte, qui reposent sur une évaluation absolue des prévisions, ont pour but de vérifier sous l'hypothèse nulle la validité des prévisions. La deuxième classe correspond aux tests basés sur une évaluation relative

des prévisions, c'est-à-dire que l'on teste les performances d'un nombre fini de modèles potentiellement tous mal spécifiés par rapport à un modèle de référence qui peut lui aussi être mal spécifié. Il faut également distinguer deux types de tests au sein de ces deux classes. Les tests de spécification statiques ont pour vocation de vérifier que la forme paramétrique de la densité spécifiée appartient à la bonne famille de distributions. Les tests de spécification dynamique s'attachent en plus de cela à vérifier que les variables transformées z_t sont indépendantes, et ne concernent donc que des prévisions à horizon $t = 1$.

La plupart des tests proposés sont des tests de type Kolmogorov-Smirnov. En effet, le test de Kolmogorov-Smirnov est un des plus populaires dès lors qu'il s'agit de tester l'uniformité d'une distribution, et a donc servi de point de départ. Le premier test de spécification correcte d'une densité est sans conteste celui de Diebold et al. (1998) qui suggèrent de comparer graphiquement directement la fonction de répartition de la série de *PIT* estimée à celle d'une distribution uniforme, l'idée étant que si la densité de prévision n'est pas correctement spécifiée, on observera un écart important entre les deux fonctions de répartition. Clements et Smith (2001) proposent d'utiliser les valeurs critiques exactes de la Statistique de Kolmogorov-Smirnov afin de construire un intervalle de confiance à $(1 - \alpha)\%$ autour de la fonction de répartition théorique de la distribution uniforme. Une autre façon de procéder est d'utiliser directement le test de Kolmogorov-Smirnov. Mais plusieurs problèmes se posent alors. En effet, l'ensemble d'informations n'est en général pas observé dans son intégralité. De plus, le vecteur de paramètres θ_0 n'est pas observé mais estimé, ce qui entraîne la présence de paramètres de nuisance qui rendent inutilisable la distribution asymptotique usuelle du test d'adéquation.

Les travaux de Bai (2003) proposant un test de spécification statique couplé à une transformation martingale de Khmaladze (1982) permettent de régler les problèmes engendrés par l'estimation de θ_0. Ce test n'a une puissance asymptotique unitaire que dans le cas d'alternatives pour lesquelles il y a violation de l'uniformité. Cependant, ce test est peu puissant contre les alternatives pour lesquelles il y a violation de l'indépendance. Les tests de Hong et Li (2005) et de Corradi et Swanson (2006) présentent quant à eux l'avantage d'être puissants pour des alternatives pour lesquelles il y a à la fois violation de l'uniformité et de l'indépendance. Le test de Hong et Li est un test de spécification dynamique correcte qui repose sur la comparaison de la densité jointe de (z_t, z_{t-j}) à celle de deux variables suivant une distribution uniforme par le biais d'un noyau Kernel modifié. Tout comme le test de Bai, ce test est libre de paramètres de nuisance et utilise une distribution asymptotique standard. en revanche une difficulté subsiste dans la mesure où l'on doit estimer un paramètre de lissage, chose souvent peu aisée. Le test de Corradi et Swanson n'est pas concerné par l'estimation d'un paramètre de lissage et converge à un taux paramétrique mais souffre de paramètres de nuisance qui impliquent de simuler les valeurs critiques par des procédures de Boostrap.

Le test de Hong et Li a par la suite inspiré ceux de Park et Zhang (2010) et Lin et Wu (2017). L'idée principale reste la même, mais plutôt que d'utiliser un test de Kolmogorov-Smirnov, ils utilisent les tests lissés de Neyman (1937) qui englobent la distribution uniforme sur l'intervalle [0,1] dans un ensemble plus large de densités dites alternatives. Le choix de ne pas utiliser les tests de type Kolmogorov-Smirnov est motivé par le fait que ces derniers ne fournissent que peu d'indications quant à l'éloignement d'une distribution uniforme lorsque l'hypothèse nulle est rejetée.

Forts de la propriété relative à la distribution des variables *INT*, et en avançant que les tests de normalité sont plus nombreux que ceux d'uniformité et qu'il est plus facile de tester l'auto-corrélation pour une distribution normale que pour une distribution uniforme (Mitchell et Wallis, 2001), des tests basés sur cette transformation ont été proposés. Nous verrons plus en détails dans le chapitre 3 le test de Knüppel (2011) qui s'intéresse aux "moments bruts" de la série des $\Phi^{-1}(z_t)$ et les tests de mauvaise spécification dynamique de Kalliovirta (2012).

Enfin, notons que nombre de tests de spécification correcte, qu'ils soient statiques ou dynamiques, supposent que le processus étudié est stationnaire, ce qui n'est en pratique pas toujours le cas. Ainsi, on a assisté à l'émergence de tests qui n'utilisent pas cette hypothèse sous l'hypothèse nulle de spécification correcte. Pour ce qui est des tests statiques, nous étudierons le test de Rossi et Sekhposyan (2013) dont le test de Corradi et Swanson est un cas particulier et les tests de spécification dynamique proposés par González-Rivera et Sun (2017).

La dernière partie de ce chapitre sera consacrée à l'application des tests de Corradi et Swanson et de Rossi et Sekhposyan sur des prévisions de log-rendements de trois indices boursiers (Nasdaq, Dow Jones et SP500) issues de modèles paramétriques "simples", évaluées dans des environnements différents (pré-crise 2008, durant la crise et post-crise). Les modélisations choisies sont des modélisations classiques en économétrie : processus ARMA(3,3) et modèles GARCH(1,1) ou GJR-GARCH(1,1) pour les résidus et les prévisions sont effectuées à horizons $h = 1$ jour, et $h = 10$ jours.

Pour les prévisions à horizon $h = 1$ jour, on constate que les deux tests et leur déclinaisons donnent des résultats assez similaires, que l'on utilise le modèle GARCH(1,1) ou GJR-GARCH(1,1). Nous sommes très souvent amenés à ne pas rejeter l'hypothèse nulle de spécification correcte de la densité en période post-crise, alors qu'en période de crise, l'inverse se produit. En période pré-crise, on rejette assez souvent l'hypothèse nulle, mais on constate tout de même que la modélisation GJR-GARCH(1,1) semble plus adaptée, car le taux de rejet est moins important qu'avec la modélisation GARCH(1,1), en particulier pour l'actif Dow Jones.

Pour les prévisions à horizon $h = 10$ jours avec modélisation GARCH(1,1), les résultats obtenus sont comparables à ceux issus des prévisions à horizon $h = 1$ jour pour tous les tests considérés. On remarque toutefois que les taux de rejet de l'hypothèse nulle en période de crise pour les tests de Rossi et Sekhposyan sont parfois très élevés. Les résultats obtenus avec la modélisation GJR-GARCH(1,1) sont une nouvelle fois proches des résultats précédents, hormis en période pré-crise, où seul l'actif Dow Jones semble difficile à modéliser. On note également que le taux de rejet de l'hypothèse nulle en période de crise est en général moins élevé pour les tests de Rossi et Sekhposyan avec cette modélisation, quand on observe le phénomène inverse pour les tests de Corradi et Swanson.

1.5 Chapitre 4 : Recalibration des prévisions en loi

Nous proposons dans cette troisième et dernière partie une méthode de recalibration de prévisions issues d'un modèle mal spécifié afin de les rendre valides. Cela nous permet par la même occasion de simplifier le choix entre plusieurs prévisions de densité.

On se place ici dans le cas où l'on dispose d'une prévision en loi estimée (ou plusieurs) issue d'un modèle non valide, i.e la distribution de la série de *PIT* zt ne correspond pas à une distribution uniforme sur l'intervalle [0,1]. On notera par la suite G la distribution estimée et F la vraie distribution. Un problème se pose alors, car la validité est une notion clé lorsque l'on parle de prévision de densité. En effet, l'avantage des prévisions de densité (par rapport aux prévisions ponctuelles) est qu'elles permettent de mieux gérer le risque. Or, la gestion de risque implique souvent plusieurs parties prenantes (internes ou externes à l'entreprise), aux objectifs différents, et qui doivent donc avoir confiance dans leur vision commune du risque. C'est ce qui se passe pour les prévisions de VaR des banques et assurances, qui sont vérifiées par le régulateur financier. On peut aussi mentionner la gestion du risque de déséquilibre grave entre l'offre et la demande. EdF avait pour contrainte réglementaire de prouver à la Commission de Régulation de l'Énergie que la probabilité d'un tel déséquilibre était inférieure à 1%. Cette contrainte, qui n'est plus réglementaire mais liée à l'image d'EdF (et s'applique toujours à RTE), impose d'utiliser des prévisions fiables.

Question 1 : comment recalibrer des prévisions de quantiles issues d'un modèle mal spécifié afin de les rendre valides ?

L'idée que nous proposons est d'utiliser le biais constaté lorsque l'on compare la distribution des z_t à celle d'une distribution uniforme pour recalibrer les futures prévisions.

Afin d'expliquer comment recalibrer les prévisions, plaçons-nous dans le cas suivant : à chaque instant, on connaît l'état du monde $e \in E$ dans lequel on se trouve (dans le cadre d'une régression par exemple, E sera l'ensemble des valeurs possibles des régresseurs) et l'on dispose de prévisions de densité conditionnelle à e, que l'on note G_e.

Résultat 1. la fonction de répartition de la variable $G(Y)$ est donnée par :

$$C(y) \equiv \Pr(\, G(\, \boldsymbol{Y}) \leq y) \equiv \sum_{e} p_e \, F_e \circ G_e^{-1}(y),$$

avec p_e la fréquence d'apparition de l'état e, sous réserve que G_e soit inversible pour tout $e \in E$.

On notera que l'on a pris ici une distribution discrète pour caractériser E, mais que le résultat reste vrai avec une distribution continue. Nous allons par la suite utiliser la fonction C pour obtenir des prévisions de densité valides. En effet, comme on peut le voir ici

Résultat 2. la fonction de répartition de la variable $C \circ G(y)$ est donnée par :

$$\Pr(\, C \, \circ \, G(\, \boldsymbol{Y}) \leq y) = y,$$

à la condition que F soit inversible.

Autrement dit, la variable corrigée suit une distribution uniforme sur l'intervalle [0,1], propriété que l'on attend d'une prévision valide. D'un point de vue pratique, la correction d'une prévision de quantile permet d'illustrer clairement les choses. Considérons que l'on dispose de prévisions de densités conditionnelles $G_{e,t}$, $t = 1 \cdots T$. Plutôt que de fournir le quantile estimé brut $G^{-1}_{e,T+1}(\tau)$ à la date $T+1$, avec $\tau \in [0,1]$ le niveau de quantile, on préférera $G^{-1}_{e,T+1}(\tau_c)$ où $\tau_c \in [0,1]$ est le quantile de niveau τ de la variable $G(Y)$ et sera estimé à l'aide de la fonction de répartition empirique de la série $\{z_t\}_{t=1}^{T}$.

Ainsi, nous proposons une méthode qui permet de rendre valides les prévisions et permet donc de simplifier le choix entre plusieurs prévisions de densité, qui ne sera plus basé que sur un critère de qualité. Cela nous amène à nous poser une seconde question.

Question 2 : Quelle est la qualité des prévisions recalibrées ?

En effet, bien que la propriété de validité soit fondamentale, il ne faut toutefois pas négliger la qualité des prévisions, qui, si elle est trop dégradée, peut être aussi dommageable que le fait de ne pas passer les tests de validité. Afin d'évaluer la qualité des prévisions recalibrées, nous avons comparé les valeurs de l'espérance de la *Pinball-Loss* obtenue selon que l'on utilise les quantiles estimés bruts ou corrigés. Soit L_τ la *Pinball-Loss* obtenue lorsque l'on utilise les quantiles estimés bruts et L_{τ_c} celle avec les quantiles recalibrés. Sous certaines hypothèses de régularité sur les fonctions G_e, F_e et leur dérivées, et sous réserve que certaines propriétés concernant les densités f_e et l'écart entre la densité estimée et la vraie densité soit respectées, on obtient la majoration suivante :

Résultat 3.

$$0 \leq \mathrm{E}_Y[\, L_\tau - L_{\tau_c}\,].$$

En d'autres termes, sous ces conditions, notre méthode de correction améliore systématiquement la qualité des prévisions. On montre également que cela reste vrai lorsque l'on utilise le *CRPS* plutôt que la *Pinball-Loss*.

Cependant, il s'agit là d'un résultat théorique, car avec un échantillon fini, il est évident qu'on ne peut pas obtenir la vraie valeur de τ_c mais seulement une valeur approchée $\widehat{\tau}_c$. On s'intéresse donc ensuite à la valeur de l'espérance de la *Pinball-Loss* obtenue selon que l'on utilise $G^{-1}(\tau)$ ou $G^{-1}(Q_\tau)$, où Q_τ est une variable aléatoire dont $\widehat{\tau}_c$ serait une réalisation et vérifie la propriété suivante :

$$Q_\tau \longrightarrow N(\tau_c, \lambda n^{-1})$$

en distribution, avec λ qui dépend de τ_c et de f_Y la vraie densité.

Sous certaines hypothèses de régularité (entre autres) sur les fonctions G_e, F_e et leur dérivées, nous obtenons ainsi une borne concernant la dégradation de la qualité des prévisions.

Résultat 4. Nous avons la majoration

$$|\,\mathrm{E}_Y[L_{\widehat{\tau}_c} - L_\tau\,]\,| \leq 2\,\varepsilon^2\,\xi + \frac{C\,\lambda}{n}.$$

Ici, la constante ε correspond à une borne supérieure de la distance entre la fonction de répartition estimée et la vraie distribution en valeur absolue pour tout $e \in E$ et prend ces valeurs entre 0 et 1. ξ est une constante strictement positive dont l'inverse permet d'obtenir une borne inférieure de la prévision de densité G_e pour tout $e \in E$ sur un intervalle précis. Enfin, C est une constante qui dépend de α, M et β, des constantes strictement positives qui permettent d'obtenir des bornes pour les dérivées des densités de prévision conditionnelles et des vraies densités conditionnelles ainsi que pour la vraie densité évaluée en $G_e^{-1}(\tau_c)$ respectivement, et cela pour tout $e \in E$. Nous montrons de cette façon que dans le pire des cas, la qualité des prévisions n'est que très légèrement dégradée. A nouveau, une majoration similaire est obtenue lorsque l'on utilise le *CRPS* comme règle de scoring.

Nous avons appliqué notre méthode de recalibration sur des prévisions d'ensemble dans différents domaines : le domaine météorologique avec des scénarios de température et le domaine économique avec des scénarios de prix. On notera toutefois que la qualité dans le domaine météorologique à été évaluée via le *CRPS*, quand les scénarios de prix sont évalués selon les valeurs de *Pinball-Loss*. De plus, nous avons choisi de n'utiliser qu'un simple test de Kolmogorov-Smirnov pour tester la validité des prévisions (pour l'application sur les températures tout du moins, le nombre de données n'étant pas assez important dans le contexte des scénarios de prix). Notre volonté était d'employer la recalibration sur des applications qui sont relatives aux secteurs les plus concernés par les prévisions en loi.

Pour l'application sur données météorologiques, nous avons implémenté plusieurs fonctions avec le logiciel R, qui permettent, entre autres, d'obtenir les valeurs des niveaux de quantile corrigés τ_c, les prévisions de quantiles recalibrées ainsi que les valeurs de *PIT* recalibrées. Nous avons confronté notre recalibration à une autre méthode de pre-processing assez populaire et développée par Gneiting et al. (2005), la méthode *EMOS* (*Ensemble Model Output Statistics*). Alors que le test de validité des prévisions avec l'ensemble brut et la méthode *EMOS* est rarement positif, notre méthode montre de très bon résultats. Pour ce qui est de la qualité des prévisions, nous avons choisi le *CRPS*, et nous avons comparé la différence entre le *CRPS* avec ensemble brut avec celui obtenu en utilisant soit notre méthode de recalibration, soit la méthode *EMOS*. Nous avons constaté que les prévisions issues de notre recalibration et de la méthode *EMOS* sont meilleures que celles issues de l'ensemble brut pour les petits horizons. Pour les horizons suivants, on constate que les scénarios recalibrés sont parfois moins performants que les ensembles bruts. Toutefois, ils restent meilleurs que les scénarios recalibrés par *EMOS*, et on remarque également que lorsque la qualité est dégradée, la différence avec celle des prévisions brutes n'est que légère.

Concernant notre application sur les scénarios de prix, nous avons utilisé les mêmes algorithmes que pour les scénarios de température. Précisons que l'on avait à disposition deux modèles bruts différents issus d'un outils de gestion de l'entreprise EdF, *OPUS*. La différence entre ces deux modèles vient du fait que l'un d'entre eux considère les pays étrangers dans leur ensemble pour produire des scénarios (modèle OPUS) quand le second les modélise individuellement (modèle OPUS-PAF). Nous avons ensuite évalué les prévisions brutes et les prévisions recalibrées sur la base de la *Pinball-Loss* pour cinq niveaux de quantile (1%, 4%, 50 %, 96% et 99%). Nous avons également étudié le pourcentage d'observations se situant en dehors des bornes de l'ensemble avant et après recalibration. Enfin, nous avons considéré un indicateur métier et ainsi calculé les valeurs de différents cash-flows pour trois commodités : thermique à flamme, hydraulique et nucléaire. Les résultats obtenus montrent que le fait d'utiliser notre méthode permet de mieux approcher la valeur réelle des cash-flows (en valeur absolue) dans 68% des cas sur un peu plus d'un an de données.

Chapter 2

Machine learning and quantile forecasts

2.1 Introduction

Nowadays, quantile forecasting is an important challenge in many fields. One can cite the banking sector with the VaR previsions due to the introduction of Basel II accord or energy field, in which forecasts of quantile for electricity demand are very important. The first approach in quantile forecast has been proposed by Koenker and Basset (1978). They propose a framework very close to that used for linear parametric regression. The principal differences lie in the asumptions about the error term, and the loss function used to estimate the parameter. Indeed, they do not specify restrictions on the two first moments, namely $E(\epsilon)$ and $Var(\epsilon)$, but instead that $F_\epsilon(0) = \tau$, with F_ϵ the c.d.f of the random variable ϵ. Moreover, the loss function used is the *Pinball-Loss* rather than the *MSE* used in linear regression. The *Pinball-Loss* is defined as follow:

$$\rho_\tau(u) = (\tau - 1)u\mathbf{1}_{\{0>u\}} + \tau u\mathbf{1}_{\{u\geqslant 0\}},$$

with $\mathbf{1}_{\{.\}}$ the indicator function and $\tau \in [0, 1]$ the quantile level.

However, face of the incredible expansion of data collected, linear modeling (whatever the loss function used) is a very limited framework. Indeed, regression in high dimensional space is not easy, for example when one needs to find relevant interactions between variables, and cannot handle some type of data like pictures. Moreover, models can have more complex structure than the linear formulation.

A credible alternative to regression is the use of machine learning technics, as shown in many papers devoted to the forecast comparison between parametric approaches (linear or non linear) and machine learning technics. For example, one can cite the works of Wang et al. (2009) and Ahmed et al. (2010). Those papers are interested in conditional expectation, though, and to the best of our knowledge, there is no paper concerned by comparison of other forecasts, particularly quantile forecasts. Thus, the goal of this paper is to benchmark the different machine learning algorithms used to make quantile forecasts and compare them with quantile regression.

The remaining of this chapter is organized as follows. The first section introduces machine learning, its pros and its cons. The second section presents some machine learning technics to make quantile forecasts and the third is dedicated to case study in different contexts. Finally, the fourth concludes the paper.

2.2 Quantile Forecasting and Machine Learning

Machine learning is a field of Computer Science (linked to Statistics and Optimization), which gives computers the ability to make good forecasts without explicitly programming them, and shows good performances in forecast. Today, machine learning is used for both classification and regression, and covers a large set of methods, including among many others artificial neural networks, support vector

machine, boosting method and random forest or additive models, that we will describe in the next sections.

As in regression framework, the goal is to make forecasts given some inputs data, but the approach of the problem is quite different. When economists are interested in the distribution of the data, machine learning is concerned by predictions only. Here is one of the principal differences between regression technics and machine learning: econometric models are in general parametric (or semi-parametric), and seek for the best model by a consistent estimation of the parameter β, based on some mathematical theories and properties and machine learning models are funded almost exclusively on the data and don't care about properties about estimators, as long as it produces good forecasts. In consequence, economists penalize their models by their complexity ex-post, contrarily to machine learning which penalizes the objective function ex-ante. It implies also another validation procedure of the parameters. Usually, economists use in-sample validation. To do so, they separate the available dataset in three parts: training, validation and test sets. The model is obtain with the training set and then validated on the validation sample. Finally, the model is tested on the last sample. Things are different with machine learning since the parameters are generally validated via k-fold cross validation. The idea is to separate the dataset into two parts: one for the estimation/validation, the other to test the model. Within the first sample, we create k "folds" of equal size and at each step, one uses a fold as validation set and the $k-1$ other folds as training sample, so that each fold is used only one time as validation set. As result, one obtain forecast for each observation of the sample and we can evaluate those forecasts. It is a way to make "out-of-sample" validation. This method is robust and approximates the optimal complexity, which produces good forecasts.

As previously mentioned, machine learning is an alternative to regression, and presents some advantages. First, since the models are almost exclusively based on the data and use forecast errors during the learning phasis to improve the next forecasts, they can capture particular patterns and complex structures and the fewer the number of data, the better. Second, machine learning find automatically the relevant interactions between variables using hyper-parameters, which is very interesting when there are a lot of explanatory variables. It controls the complexity of a model and plays a role of regularizer. For example, one can cite the minimum number of observations in a final node of a tree or its depth. Moreover, machine learning allows us to use variables with strong colinearity and we don't need apply pre-processing procedure to the data to work with.

However, machine learning is not a miracle remedy and has its own limits. It can be seen as a "black box", and in consequence, it is not easy to interpret the models estimated. Moreover, the absence of strong hypotheses conduct to non-consistent estimators, which is an important condition to quantify an inferior bound for the variance, as economists do. Thus, it is important to take into account that machine learning is used to predict, and don't conclude systematically about mathematical properties for estimators issued from those algorithms. A good example will be the coefficient estimated by penalized regression, which often don't show the properties required when using linear regression. Besides, we don't really know how the variables are selected by the algorithms, and that could be a problem in term of model selection since two models can have the same forecast performances. For more details, the interested reader can consult the articles of Mullainathan et al. (2017), Charpentier et al. (2017) and Athey (2018).

In the same fashion that we can slide from the linear regression to quantile regression, it is possible, principally modifying the loss function used in the algorithms, to slide from machine learning regression to quantile regression. The next section presents some technics to produce quantile forecasts by machine learning.

2.2.1 Support Vector Quantile Regression

Support Vector Machine was first developed by Vapnik and Chervonenkis (1964). It is a supervised learning algorithm, and was initially used for classification problems. The idea is to find an hyperplane which separates optimally the data in two groups. Key concepts are the margin and the Kernel trick. The margin is the distance between the hyperplane and the closest observations (called support vector) and the goal is to find the hyperplane maximizing the margin. Thus, we need to find $h(x) = \langle \omega, x \rangle + b$, with

$x \in \mathbb{R}^d$, a vector of explanatory variables, $b \in \mathbb{R}$ and $\omega \in \mathbb{R}^d$ the normal (not necessarily normalized) vector to the hyperplane. The bigger the margin, the better, so we need to minimize $\|\omega\|^2$. To do so, we need to solve a quadratic optimization programming, expressed in primal or dual form. However, the problem is sometimes unfeasible, that's why we use slack variables, link to the concept of soft margin. Moreover, the dataset is sometimes not linearly separable. A way to fix this is to use that we call the Kernel trick. The dot product is replaced by a nonlinear Kernel function which mappes the dot product in higher feature space where the data are separable.

Vapnik et al. (1997) extend the SVM to regression problem (SVR). To use SVM for regression, one needs to use the ϵ-insensitive function, where ϵ represents the proportion of forecasting error tolerated. It creates a "tube" within the errors are not penalized, as shown in figure 2.1.

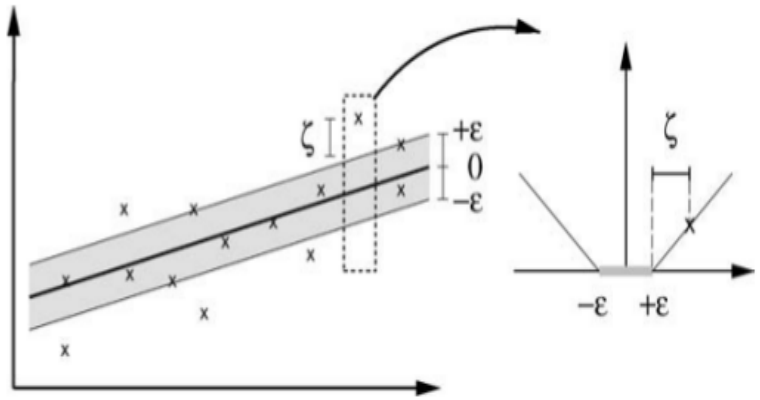

Figure 2.1 – The soft margin loss setting for a linear SVM (from Schölkopf and Smola, 2002).

In other words, we do not care about error forecasts as long as they are less than ϵ in absolute value. Thus, we need to find a function $f(x) = \langle \omega, x \rangle + b$ and to minimize $\frac{1}{2}||\omega||^2$ subject to $y_i - (\omega \cdot x_i + b) \leq \epsilon$, $(\omega \cdot x_i + b) - y_i \leq \epsilon$. If the problem is not feasible, one can use slack variables and it can be rewritten as the following primal programming:

$$\underset{\omega, \xi_i, \xi_i^*}{min} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*),$$

subject to $y_i - (\omega \cdot x_i + b) \leq \epsilon + \xi_i$, $(\omega \cdot x_i + b) - y_i \leq \epsilon + \xi_i^*$, $\xi_i, \xi_i^* \geq 0$, with $C > 0$ a penalization term which is a trade-off between flatness of $f(x)$ and the amount up to which deviations larger than ϵ are tolerated. In other words, it controls for the number of slack variables ξ_i, ξ_i^*.

From the formulation of SVR, one can obtain quantiles replacing the ϵ-insensitive function by the *Pinball-Loss* function (SVQR), as proposed by Hwang and Shim (2005) and Takeuchi et al. (2006). This involves finding a function $Q_y(\tau|x) = \langle \omega_\tau, x \rangle + b_\tau$ and to solve the following primal programming:

$$\underset{\omega_\tau, b_\tau}{min} \frac{1}{2}\|\omega_\tau\|^2 + C_\tau \sum_{i=1}^{n} \rho_\tau(y_i - (\omega_\tau \cdot x_i + b_\tau)), \tag{2.1}$$

where $C_\tau > 0$ plays the same role as in SVR framework. By introducing slack variables ξ_i , ξ_i^*, we can rewrite equation 2.1 as a quadratic primal programming:

$$\underset{\omega_\tau, b_\tau, \xi_i, \xi_i^*}{min} \frac{1}{2}\|\omega_\tau\|^2 + C_\tau \sum_{i=1}^{n} \tau\xi_i + (1-\tau)\xi_i^*, \tag{2.2}$$

subject to $y_i - (\omega_\tau \cdot x_i + b_\tau) \leq \xi_i$, $(\omega_\tau \cdot x_i + b_\tau) - y_i \leq \xi_i^*$ and $\xi_i, \xi_i^* \geq 0$.

Now, one can construct the dual programming using Lagrange function and according to Karush-Kuhn-Tucker conditions, we can express 2.2 as the following minimization programming:

$$\underset{\alpha_{\tau,i},\alpha^*_{\tau,i}}{min} \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_{\tau,i}-\alpha^*_{\tau,i})(\alpha_{\tau,j}-\alpha^*_{\tau,j})K(x_i,x_j)-\sum_{i=1}^{n}(\alpha_{\tau,i}-\alpha^*_{\tau,i})y_i,$$

subject to $\sum_{i=1}^{n}(\alpha_{\tau,i}-\alpha^*_{\tau,i})=0$, $\alpha_{\tau,i}\in[0,\tau C]$ and $\alpha^*_{\tau,i}\in[0,(1-\tau)C]$, with $\alpha_{\tau,i},\alpha^*_{\tau,i}$ the Lagrange multipliers and $K(\cdot,\cdot)$ a Kernel function.

Denote $I_{SV}=\{i=1,2\cdots,n|0<\alpha_{\tau,i}<\tau C_\tau\,,\,0<\alpha^*_{\tau,i}<(1-\tau)C_\tau\}$, the index ensemble of support vectors. Finally, we can estimate $Q_y(\tau|x)$ by:

$$\widehat{Q}_y(\tau|x)=\widehat{\omega}_\tau\cdot x+\widehat{b}_\tau,$$

with $\widehat{\omega}_\tau=\sum_{i=1}^{n}(\alpha_{\tau,i}-\alpha^*_{\tau,i})x_i$, $\widehat{b}_\tau=\frac{1}{|I_{SV}|}\sum_{i'\in I_{SV}}b_{\tau,i'}$ and $b_{\tau,i'}=y_{i'}-\sum_{i=1}^{n}(\alpha_{\tau,i}-\alpha^*_{\tau,i})K(x_i,x_{i'})$.
It is important to note that the final result is expressed in function of the support vectors only.

Takeuchi et al.(2006) extend their work to propose methods which avoid quantile-crossing or to ensure monotonicity with respect to some variables, as for the standard support vector regression. For example if we want to produce growth curve, we need monotonicity in function of the age.

Many extensions of the SVQR exist in the literature. Hwang and Shim (2010), propose a method based on a quadratic loss function, which is a transformation of the *Pinball-Loss*. This involves solving an optimization problem using iterative reweighted least square. Seok et al. (2010) use an asymmetric ϵ-insensitive loss function, which provide sparsity contrarily to SVQR, since the *Pinball-Loss* is not differentiable in 0. Provide sparsity refers to the ability to obtain the regression function using the smallest number of variables. In SVM context, this is characterized by a small number of support vectors. One can also mention the work of Xu et al. (2015). They propose to use SVQR with a weighted version of the *Pinball-Loss* (a real weighted version, not a transformation of the *Pinball-Loss*, as Hwang and Shim (2010) do).

2.2.2 Quantile Regression Neural Network

The idea of neural network was first to mimic human brain to take decisions for classification problems. As a human brain, the algorithm is trained, for example by pictures labeled by the user, in order to "learn" and achieve the capacity to classify future pictures with the appropriated label.

The neural network is composed by layers: an inputs layer, one or more hidden layer and an output layer. Each layer consists in few neurons. The input layer has a number of neurons equal to the number of explanatory variables. The neurons in a layer are not connected between them, but are all connected to each neurons of the next layer. When inputs are send to a neuron, it creates a weighted sum with bias, which is then passed in an activation function attributed to the neuron. This produces an output which will be send as input to the next hidden layer's neurons and so on, until achieve the output layer. From a mathematical point of view, we take as illustration a neural network with one hidden layer. It results a function:

$$f(x_t)=g_2\left(\sum_{j=0}^{m}\nu_i g_1\left(\sum_{i=0}^{n}\omega_{ji}x_{it}+b_j\right)+b\right), \tag{2.3}$$

where g_1 and g_2 are activation functions (commonly a sigmoid function and a linear one respectively [1]), ν_i and ω_{ji} are output-layer and input-layer weights respectively, with m the number of nodes in the layer, and n the number of explanatory variables (determining the complexity of a model), and b_j and b the bias. Then, we need to solve an optimization problem to estimate the weights and the bias.

1. The function g_2 is often the identity. Other possible choices for g_1 and g_2 are threshold functions for example

In order to estimate quantiles, Taylor (2000) suggests to use the *Pinball-Loss* score as objective function [2]. Indeed, if we take the function defined in 2.3 as input, minimizing the *Pinball-Loss* conduct to estimate the weights and bias such that the function 2.3 produces quantiles. However, since a too complex model can lead to overfitting, we need regularizer parameters. Thus the optimization problem of Taylor becomes:

$$\underset{\omega,\nu,b}{min} \sum_{t=1}^{T} \rho_\tau(y_t - f(x_t)) + \lambda_1 \sum_{i,j} \omega_{j,i}^2 + \lambda_2 \sum_i \nu_i^2,$$

with $\rho_\tau(\cdot)$ the *Pinball-loss* and λ_1 and λ_2 the regularizer parameters.

Kulczycki and Schioler (1998) propose a method to obtain quantile estimation based on neural network and Kernel estimation. In Kernel estimation theory, an estimator of the conditional density of a random variable Y is defined as:

$$h(Y|X) = \frac{1}{mV(r)} \sum_{i=1}^{m} H\left(\frac{Y - y_i}{r}\right) H\left(\frac{X - x_i}{r}\right),$$

where m represents the number of available observations, y_i the i^{th} observation, x_i the i^{th} realization of the explanatory variable, $H(\cdot)$ the chosen Kernel function and r the bandwidth parameter.

The idea is to consider this estimator as output of a neural network, where the function $H(\cdot,\cdot)$ is viewed as activation function and r, y_i and x_i as weights and bias. They propose a transformation of this estimator to obtain estimation of cumulative distribution, directly used to forecast quantiles. Moreover, they suggest another transformation to provide compression (i.e a neural network where only a subset of the available data is used rather than the whole observations, which will lead to an important number of neurons).

We find also in the recent literature the methods proposed by Hatalis and Kishore (2017) and Cannon (2018). The method of Hatalis and Kishore seems work very well for time series. They use particular activation functions in order to produce Fourier transform at the end [3]. Cannon propose a method to avoid quantile-crossing for multiple estimates, says monotone composite quantile regression neural network. It is a combination of the monotone quantile regression neural network (which is itself a combination of the monotone multi-layer perceptron proposed by Zhang and Zhang (1999) and the quantile regression neural network proposed by Taylor) and the composite quantile regression neural network of Xu et.al (2017). Xu et al. propose a method structurally very close to that of Taylor. The difference lie in the quantile regression error function used, which is summed over K (generally equally spaced) value of τ, with $\tau_k = \frac{k}{K+1}, k = 1 \cdots K$ for example. It is important to note that even if we sum over K values of τ, the function is τ-independant and used to forecast only one quantile.

2.2.3 Additive Quantile Models

Additive models have been first proposed by Friedman and Stuetzle (1981). The model is written as follows:

$$\mu_t = \beta_0 + \sum_{i=1}^{m} f_i(x_i),$$

with μ_t the mean of the variable of interest, β_0 a constant, m the number of regressors, and f_i are unknown smooth functions fit from the data.

In the same fashion that we can extend the linear regression to the quantile regression, we can use additive models to forecast quantiles.

2. In practice, since the algorithm uses gradient descent to solve the problem and the *Pinball-Loss* is not differentiable at 0, a smooth approximation, says the Huber Loss, is used.

3. Notice that the smooth approximation of the *Pinball-Loss* used here is different to that used by Taylor

Generally speaking, an additive quantile model has the following formulation:

$$y_t = c_\tau + \sum_{i=1}^{m} q_i(x_i) + \epsilon_\tau, \tag{2.4}$$

with c_τ a constant term, ϵ_τ an error term such that $P(\epsilon_\tau \leq 0) = \tau$ and q_i are smooth functions written as $q_i(x_i) = \sum_{r=1}^{R} \beta_{ri} b_{ri}(x_i)$, with β the regression parameters, b_{ri} the basis function (spline, polynomial or Kernel for example) and R the degree of smoothing. It is important to note that the basis functions can be different from a variable to another. Then, the common way to find the different parameters is to minimize the *Pinball-Loss* or a function which is often a combination of the *Pinball-Loss* and a penalization term, using the interior point algorithm.

To forecast quantiles, Doksum and Koo (2000) propose spline estimate for the additive component in 2.4. De Gooijer and Zerom (2003) and Horowitz and Lee (2005) suggest marginal integration and two step estimate respectively. The work of De Gooijer and Zerom was extended by Dette and Scheder (2011), using marginal integration and non-increasing rearrangements to avoid quantile crossing. Cheng et al. (2011) propose two Kernel-based estimators. The first one is an internally normalized Kernel smoother, which can be seen as an alternative to that of De Gooijer and Zerom. The second one involves sequential fitting by univariate local polynomial quantile regression for each additive component with the other additive components replaced by the corresponding estimates from the first proposed estimator. One can also mentioned the works of Koenker (2011), where the use of the total variation roughness penalties is suggested to control the smoothness of the additive component or Sherwood and Wang (2016), where a partially linear additive quantile regression is presented.

Another approach is proposed by Fasiolo et al. (2017), which is an extension of the work of Gaillard et al. (2016). In their article, Fasiolo et al. work in a Bayesian framework, supposing prior Gaussian distribution on the regression parameter β and use cubic splines as smooth functions. However, since quantile regression is based on a loss function and not a density, it is not possible to apply Bayes rule to update the prior. The general prior belief updating proposed by Bissiri et al. (2016) is used to circumvent this problem. This conduct to the scaled "Gibbs posterior":

$$p(\beta|y) \propto \prod_{i=1}^{n} \widetilde{p}_F\{y_i|\mu, \sigma, \tau, \lambda\} p(\beta),$$

with $\widetilde{p}_F$ the Extended Log-F (ELF) density which is linked to Kernel quantile estimators and defined as:

$$\widetilde{p}_F\{y_i|\mu, \sigma, \tau, \lambda\} = \frac{e^{(1-\tau)\frac{y-\mu}{\sigma}}(1 + e^{\frac{y-\mu}{\lambda\sigma}})^{-\lambda}}{\lambda\sigma \mathrm{Beta}[\lambda(1-\tau), \lambda\tau]}$$

where $\mu(x)$ is the quantile, $\sigma(x) = \sigma_0 \exp\left\{\sum_{j=1}^{m} f_j(x)\right\}$ (μ and σ implicitly depend on β), σ_0 is the reciprocal of the learning rate ν, which determines the relative weight of the loss and of the prior, $\mathrm{Beta}(\cdot,\cdot)$ the Beta function, τ the quantile level and λ is such that $\lambda\sigma$ is the bandwidth parameter. The additive terms f_j are fixed, random or smooth effects and their purpose is to modulate the learning rate. The ELF density embed large set of densities and particularly, the asymmetric Laplace density pAL. This density is expressed in function of the *Pinball-Loss* and was first proposed by the authors for their work. However, since pAL is non-differentiable at its mode and $\log pAL$ is piecewise linear, standard optimizer cannot be used, and the framework of Wood et al.(2016) for prior hyper-parameter (i.e smoothing parameter) is not respected.

Maximizing the Gibbs posterior permits to obtain the regression parameters ($\beta = (\beta^\mu, \beta^\sigma)$). Moreover, Fasiolo et al. use a penalized loss proportional to the Gibbs posterior to estimate the regression coefficient, the Laplace Approximate Marginal Loss criterion to estimate the smoothing parameters and the Integrated Kullback-Leibler divergence to calibrate σ_0. Finally, the parameter λ is selected using the relation $\lambda = \frac{\epsilon}{2\log(2)\sigma \sup_y f(y)}$, with f the p.d.f of y and ϵ is an acceptable value selected by the user such that $|F(\mu^*) - F(\mu_0)| \leq \epsilon$, where F is the c.d.f of y, μ_0 the true quantile and μ^* minimize the

negative expectation of $\widetilde{p}_F$. To obtain $\sup_y f(y)$, Fasiolo et al. propose to model y by a Gaussian additive model, where both the mean and the variance are allowed to vary with x in an heteroscedastic setting.

2.2.4 Quantile Regression Forest

The use of random forest in a linear regression framework was proposed by Breiman (2001) and works as follow. First, we need to growth K single trees, each constructed from a subset of the available covariates and based on a bagged version of the estimation sample. For each leaf of a single tree, a weight is associated, averaging over all the observations in this leaf. Thus, a weight is associated to a new data point $X_i = x$ in the following way:

$$\omega_i(x, T_k) = \frac{\mathbf{1}_{\{X_i \in R_{l(x,T_k)}\}}}{\#\{j : X_j \in R_{l(x,T_k)}\}},$$

with $\mathbf{1}_{\{\cdot\}}$ the indicator function, T_k the k-th tree, and $R_{l(x_i,T_k)}$ the leaf X_i belongs to.

Then, the final weights are obtained averaging the weights from each tree:

$$\omega_i(x) = \frac{1}{K} \sum_{k=1}^{K} \omega_i(x, T_k),$$

and finally, we have:

$$\widehat{\mu}(x) = \sum_{i=1}^{n} \omega_i(x) Y_i,$$

with n the number of observations and $\widehat{\mu}(x)$ the estimation of the mean of the random variable Y.

To the best of our knowledge, the only work concerned by quantile forecast by random forest is due to Meinshausen (2006). To do so, we only need to consider $\mathbf{1}_{\{Y_i \leq y\}}$ rather than Y_i to obtain:

$$\widehat{F}(y|X = x) = \sum_{i=1}^{n} \omega_i(x) \mathbf{1}_{\{Y_i \leq y\}},$$

with $\widehat{F}$ the estimated conditional c.d.f of y. Finally, we can deduce from this estimator the desired quantile.

2.2.5 Quantile Forecast with Gradient Boosting Machine

The idea of Gradient Boosting is to obtain a forecast iteratively by gradient descent, combining "weak learners". Thus, we estimate y by:

$$\widehat{f}(x) = \sum_{m=1}^{M} \eta_m g_m(x) + \text{cste},$$

where M is the number of iteration, η_m are multipliers and g_m are functions we need to estimate, that we will describe in the next.

In a general framework, the estimation of the parameters η_m, the constant and the functions g_m consists in few steps. First, the model is initialized by a constant value (which will be the constant in the final model) $f_0(x)$, wich minimize $\sum_{i=1}^{n} L(y_i, \cdot)$, where n is size of the available dataset, y_i the observations and L the chosen loss function.

Then, we need to update the forecast. To do so, for each iteration $m = 1$ to M, we need to compute:

$$U_i = -\frac{\partial L(y_i, f)}{\partial f}|_{f=f_{m-1}(x_i)}, i = 1 \cdots n,$$

the negative gradient (pseudo residuals) of the loss function L w.r.t f evaluated at $f_{m-1}(x_i)$. Here, n represents the number of available data.

This produces a vector of gradient U which will be fitted to the set of explanatory variables by the chosen weak learners (linear regression or trees). The function estimated to fit U to the set of explanatory variables is the function g_m.

The next step is to find:

$$\eta_m = \underset{\eta}{min} \sum_{i=1}^{n} L(y_i, f_{m-1}(x_i) + \eta g_m(x_i))$$

and finally, to update the forecast:

$$f_m(x) = f_{m-1}(x) + \eta_m g_m(x).$$

Besides, to avoid overfitting, a shrinkage parameter ν (also known as learning rate) is often introduced in the forecast updating step and we have $f_m(x) = f_{m-1}(x) + \nu\eta_m g_m(x)$.

When Gradient Boosting is used to produce quantile forecasts, the chosen loss function is obviously the *Pinball-Loss*, and the difference between some implementations resides in the method employed to fit the vector U on the set of explanatory variables. Zheng (2012) and Ben Taieb at al. (2015) use a least square regression. Note that the method proposed by Ben Taieb et al.(2015) combines additive model and gradient boosting. Indeed, they specify an additive model and a gradient boosting is used such that each weak learner correspond to one input variable. At each iteration m, the best weak learner is selected and in consequence, only a subset of the available explanatory variables is selected to the model. They use dummy variable for categorical variable, and linear and non linear effect (P-spline) for continuous variables. Another weak learner largely used is tree, as Landry et al.(2016) do for the Global Energy Forecasting Competition 2014 (GEFCom2014).

2.3 Data and Empirical Framework

In order to evaluate the predictive performances of machine learning algorithms and compare them with forecasts issued from linear approach, as the quantile regression (Qreg) proposed by Koenker and Basset (1978), we use five different methods: the Support Vector Quantile Regression (SVQR) of Takeuchi et al. (2006), the Quantile Regression Neural Network (QRNN) method of Taylor (2000), the Quantile Regression Forest (QRF) proposed by Meinshausen (2006), the Gradient Boosting Machine (GBM) method employed by Landry et al. (2016) and finally the Extended Log-F (ELF) method used by Fasiolo et al. (2017).

We have selected 3 datasets for this case study, in order to highlight the performances of machine learning algorithms in different contexts. Thus, we dispose of two datasets composed by individual i.i.d data, and one dataset concerned about time series. Among the i.i.d datasets, we have chosen a dataset with a lot of regressors and another with a small number of regressors.

The first dataset is the "mcycle" dataset (available in the *R* package *VarReg*) and is commonly used for illustration of quantile regressions. It consists in 133 observations from a simulated motorcycle accident, used to test crash helmets, of 2 variables. We try to forecast the acceleration of the head measured in g during an accident in function of the time of impact in second.

The second one is another widely used dataset. The "BostonHousing" dataset (available in the *mlbench* *R* package), collecting Housing data for 506 census tracts of Boston from the 1970 census. It contains 14 variables: the dependent variable which is the median value of owner-occupied homes in USD 1000's, and 13 regressors.

The last one is a time series dataset used for the GEFCom2014 and is concerned by electricity loads. It contains 6 variables, including the electricity loads in GW, the dependent variable, and 5 regressors.

We have almost 7 years of data, from 01/01/05 to 01/12/11, with one observation per day which is the electricity loads at 12:00. Thus, there are 2525 observations available. The table 2.1 present some descriptive statistics of the dependent variable of each dataset.

Dataset	Mean	Variance	Skewness	Kurtosis	Number of observations
Mcycle	-25.55	2335.00	-0.50	-0.35	133
Bostonhousing	22.53	84.59	1.10	1.45	506
GEFCom2014	150.97	1779.88	0.68	-0.52	2525

Table 2.1 – Descriptive statistics of the dependent variable

All the algorithms are implemented with R software:

— The *qrsvm* package is based on the article of Takeuchi et al. (2006) and used to implement the support vector quantile regression method. By default, the Kernel function used is the Radial Basis Function $K(x, x^{'}) = \exp(-\sigma||x - x^{'}||^2)$. Thus, the hyper-parameters to select are σ and obviously the regularizer C_τ.

— The *qrnn* package is based on Taylor's work (2000) and is used to implement quantile regression neural network. It contains only one hidden layer by default. However, it doesn't permit to penalize the input-hidden weights as Taylor does. That's why we have only two hyper-parameters to chose, says the number of hidden nodes m, and the regularizer λ_1.

— The *quantregForest* package is used to implement the Meinshausen's random forest method, and we need to use four hyper-parameters: the number of trees growth, the number of variables randomly selected to split at each nodes, the minimum observations of terminal nodes and the maximum number of terminal nodes trees in the forest can have.

— The gradient boosting machine method is implemented in the *gbm* package, and need also four hyper-parameters. Similarly to the random forest method, since it uses trees, we need to select the minimum observations of terminal nodes, the maximum number of terminal nodes trees and the number of trees. The last parameter is the shrinkage parameter, also known as learning rate or step-size reduction.

— The ELF method is based on the work of Fasiolo et al. (2017) and implemented in the *qgam* package. There is no hyper-parameter to select since they are all automatically calibrated. We can, though, chose an ϵ value different that the value proposed by default to calculate the bandwidth parameter. In this study, we let the default value, which is $\epsilon = 0.05$.

We have separated each dataset in two parts, says a train sample to select the best hyper-parameters and a test sample. For the "mcycle" dataset, we use 75% of the data (one hundred observations) for the train sample and 25% (33 observations) for the test set. The partition for "BostonHousing" is approximatively 80% for the train (400 observations) and 20% for the test (106 observations). For the time series dataset, it was important to keep temporal structure in the sample. Thus, we have used the five first years of the dataset as train sample (1825 observations, representing approximatively 72% of the whole dataset) and the remaining for the test (approximatively 28% of the dataset or 700 observations). Since a lot of those algorithms have a stochastic part, we set seed to obtain reproducible results.

To choose the hyper-parameters, we use 10-fold cross validation for "mcycle" and "BostonHousing" datasets, and 5-fold cross validation for the time series data. For the 10-fold cross validation, each fold is created randomly, without replacement, so that each observation belongs to only one test set. Things are different for time series data, since the dependence between observations is important. Thus, we use the fixed-origin forward chaining method proposed by Tashman (2000) which consists of separating the data in equal subsets respecting their chronological order, using a fold as validation set and all the precedents folds as training set. In the same paper, Tashman (2000) proposes another, more accurate evaluation method; we could not use it due to excessive computational burden. For the same reason, the forecasts

on the test sets after tunning hyper-parameters are made using directly the entire sets rather than one by one for the support vector quantile regression and the quantile regression neural network. For the other algorithms, though, the forecasts on the test sets are made one by one. We have compared the forecasts with that obtained using directly the test sets and observed results with the same order of magnitude. Thus, we are confident in the order of magnitude of the forecasts for all algorithms.
Moreover, for the support vector quantile regression and the quantile regression neural network, we use standardized data, which is often advised with those algorithms. The grids search used for each hyper-parameter are either selected by ourselves, when the number of reasonable choices is acceptable (as the number of trees growth for example), or chosen randomly when there is a lot of possibilities (as for the number of observations in terminal nodes).

The quantiles predicted are the quantiles at level 1%, 5%, 25%, 75% and 95%, and to evaluate the quality of the forecasts, we use the *Pinball-loss* score. To avoid the (potential) quantile-crossing problem, we have implemented functions to predict successively the different quantiles desired [4] and we have reordered the forecasts when necessary.

Tables 2.2, 2.3 and 2.4 show the hyper-parameters selected for the "mcycle" dataset, the "BostonHousing" dataset and the GEFCom2014 dataset, respectively [5].

Method	parameters	$\tau = 1\%$	$\tau = 5\%$	$\tau = 25\%$	$\tau = 75\%$	$\tau = 95\%$	$\tau = 99\%$
SVQR	C_τ	1000	1000	100	1000	1000	100
	σ	0.25	1	1	0.25	1	1
QRNN	nb neurons	2	2	2	2	2	2
	λ_1	0.1	0.0001	0.0001	0.001	0.0001	0.0001
QRF	depth	5	5	5	5	5	5
	nb obs leaf	16	17	18	18	15	17
	nb split var	1	1	1	1	1	1
	nb trees	100	100	100	1000	100	1000
GBM	depth	3	5	3	2	3	4
	nb obs leaf	11	10	10	10	10	11
	shrinkage	0.01	0.1	0.1	0.1	0.1	0.1
	nb trees	100	100	500	100	100	100

Table 2.2 – Hyper-parameters selected for mcycle.

Method	Parameters	$\tau = 1\%$	$\tau = 5\%$	$\tau = 25\%$	$\tau = 75\%$	$\tau = 95\%$	$\tau = 99\%$
SVQR	C_τ	500	20	500	100	1000	100
	σ	0.001	0.01	0.01	0.1	0.01	0.01
QRNN	nb neurons	1	1	1	3	2	2
	λ_1	0.1	0.01	0.0001	0.001	0.001	0.001
QRF	depth	13	13	14	14	14	13
	nb obs leaf	24	20	17	24	15	21
	nb split var	12	10	8	7	13	12
	nb trees	100	100	100	500	100	100
GBM	depth	5	4	4	4	6	5
	nb obs leaf	23	25	17	12	12	19
	shrinkage	0.1	0.1	0.1	0.1	0.1	0.1
	nb trees	100	500	500	1000	100	100

Table 2.3 – Hyper-parameters selected for BostonHousing.

4. When a simultaneous predictions is not allowed by the algorithms.
5. The seed used and the grids search can be found in Appendix B.

Method	Parameters	$\tau = 1\%$	$\tau = 5\%$	$\tau = 25\%$	$\tau = 75\%$	$\tau = 95\%$	$\tau = 99\%$
SVQR	C_τ	10	100	500	1000	100	500
	σ	0.5	0.1	0.1	0.1	0.1	0.1
QRNN	nb neurons	2	3	3	2	3	3
	λ_1	0.01	0.001	0.01	0.01	0.01	0.01
QRF	depth	18	18	18	18	12	14
	nb obs leaf	31	37	45	26	34	33
	nb split var	5	5	5	3	4	3
	nb trees	100	100	100	100	500	100
GBM	depth	11	16	20	7	7	7
	nb obs leaf	31	28	27	28	27	28
	shrinkage	0.1	0.1	0.1	0.1	0.01	0.1
	nb trees	1000	1000	1000	100	500	100

Table 2.4 – Hyper-parameters selected for GEFCom2014.

The tables 2.5, 2.6 and 2.7 show the mean of the *Pinball-Loss* obtained using the different algorithms for different quantile levels.

Quantile	SVQR	QRNN	QRF	GBM	ELF	Qreg
1%	10.80	145.69	1.90	0.98	2.85	2.05
5%	4.55	8.53	3.17	2.69	5.51	14.50
25%	9.23	15.70	11.65	10.80	14.44	45.76
75%	7.15	19.46	11.88	11.64	32.40	8.40
95%	4.28	4.37	3.59	3.46	3.38	3.96
99%	0.64	0.91	0.72	0.69	0.65	1.05

Table 2.5 – Mean of *Pinball-Loss* score for mcycle

The table 2.5 shows the *Pinball-Loss* for the "mcycle" dataset. Overall, We remark that all the algorithms present poor results. More precisely, one can see that quantile regression neural network model obtains very high value for the quantile at level 1%. Those bad results are probably due to the number of regressors and the high variability of the dependent variable.

Quantile	SVQR	QRNN	QRF	GBM	ELF	Qreg
1%	0.17	0.17	0.15	0.43	0.15	0.15
5%	0.67	0.41	0.43	0.74	0.42	0.43
25%	1.26	1.10	1.12	1.25	1.69	1.26
75%	2.23	2.86	1.47	1.20	2.90	2.95
95%	0.66	0.88	0.51	0.57	1.36	1.64
99%	0.27	0.20	0.16	0.34	1.16	0.67

Table 2.6 – Mean of *Pinball-Loss* score BostonHousing

The table 2.6 shows the *Pinball-Loss* obtained on the "BostonHousing" dataset. One can see that quantile regression forest seems adapted to dataset with a lot of explanatory variables. It shows the best results for the quantile at level 1%, 95% and 99%. Moreover, the results obtained for quantile at levels 5% and 25% are close to that of the neural network, which are the best results presented for those quantile levels. Furthermore, this algorithm shows also results close to that of the quantile regression

forest for extreme quantiles. For the quantile at level 75%, the quantile regression forest ranks second, when the best result is obtained using gradient boosting machine. The gradient boosting machine shows also good result for the quantile at level 95%, but works quite bad for the other quantiles. The support vector quantile regression and the ELF method seem not adapted here, even if they obtain good results for the quantile at level 1% and the quantile at level 5% for the ELF method. For the support vector quantile regression, the optimization algorithm used in the algorithm is not adapted when we dispose of a lot of regressors. For the ELF method, a possible reason is that it is very sensitive to the variables included in the model and needs a variable selection procedure. In this case study, we do not use such a procedure since the idea is to use directly the available dataset. The quantile regression shows mixed results. Indeed, the results obtained for the quantile at level 1% et 5% are close to that of the random forest, but this algorithm seems work bad for the other quantiles. This method needs probably a variable selection procedure and a modeling of the interactions to improve the results.

Quantile	SVQR	QRNN	QRF	GBM	ELF	Qreg
1%	0.65	0.95	0.44	0.59	0.38	0.65
5%	1.54	2.66	1.55	1.14	1.00	3.01
25%	3.40	6.00	4.39	2.64	2.58	12.71
75%	4.05	4.48	3.68	2.87	2.58	15.33
95%	1.98	1.32	1.41	1.26	0.89	4.64
99%	1.08	0.36	0.46	0.32	0.27	1.18

Table 2.7 – Mean of *Pinball-Loss* score GEFCom2014

We can read in the table 2.7 the *Pinball-Loss* for the GEFCom2014 dataset. One can see that the ELF method shows systematically the best results. We remark, also that the gradient boosting machine shows results close to that of the ELF method, excepted for quantile at levels 1% and 95%. Furthermore, the support vector quantile regression, the quantile regression forest and the quantile regression neural network seem not very adapted for this dataset, even if the two latest algorithms mentioned present good results for quantile at levels 1% and 99% respectively. The time series framework and the number of explanatory variable can explain those results. Besides, it is important to note that the dataset used here is that used by Gaillard et al. (2016) for the GEFCom2014 after a variable selection procedure among a large set of variables. Thus, the ELF method, which is an extension of the work of Gaillard et al. (2016), has an advantage on the others. Finally, one can see that the quantile regression method shows the worst results for all quantiles considered.

2.4 Conclusion

We have presented in this chapter some methods to forecast quantiles using machine learning algorithms. Through our empirical framework, we highlight some properties and limitations of those algorithms and show that some of them are more adapted in function of the type of dataset at hand. Indeed, the first dataset used has only one regressor and the dependent variable has a lot of variability. In this case, all the machine learning technics show poor results. The second dataset contains a lot of regressors and we remark that the quantile regression forest is very efficient, when the other algorithms present mixed results, varying with the quantile considered. In time series framework, the ELF method and the gradient boosting machine seem more adapted. Remember though that the dataset contains variables selected by Gaillard et al. (2016) for the GEFCom2014 among a large set of variables.

Moreover, we have compared the machine learning algorithms performances with that of the quantile regression proposed by Koenker and Basset (1978). Here again, the results obtains for the "mcycle" dataset are poor. In time series framework, we can see that this method is not adapted, since all the machine learning algorithms are more efficients. However, results are mitigate for the dataset with a lot of regressors. Indeed, the results are quite good for the low quantiles, and for the other quantiles, the

Pinball-Loss score obtained are not necessary the greater. Those results could probably be improved with a variable selection procedure and a modeling of the interactions between variables. However, we want to compare the results using directly the available dataset since one of the principal advantages of machine learning is that it can produce forecast without programming the computer. Moreover, another advantage of machine learning is that there is no need to model the interactions between variables.

Chapitre 3

Validation de prévisions en loi

3.1 Introduction

La prise en compte de la non linéarité (changement de régimes, ruptures, etc.) a profondément changé les approches de l'économétrie appliquée à la macroéconomie et à la finance. Cette évolution est sans nul doute comparable à celle qu'a pu connaître la micro-économie lorsque l'on a progressivement abandonné l'univers de référence walrassien, que nous pouvons assimiler à la modélisation linéaire en économétrie, pour s'orienter vers les multiples formes de la concurrence imparfaite, auxquelles nous pouvons assimiler les innombrables modélisations non linéaires. Il existe aujourd'hui un très grand nombre d'études tendant à démontrer la présence de changements structurels et l'existence d'asymétries dans la dynamique des principaux agrégats macro-économiques ou des séries financières. Étant donné qu'il est impossible de rendre compte de ce type de phénomènes à partir des modèles linéaires autorégressifs usuels qu'ils soient univariés (ARMA) ou multivariés (VAR), de très nombreuses modélisations non linéaires ont été proposées afin de reproduire ces caractéristiques. On peut citer ici, parmi d'autres, les modèles à changement de régimes markoviens, les modèles avec ruptures, etc.

Ces modèles non-linéaires conduisent à un profond renouvellement de la problématique de la prévision à court et moyen terme des séries macro-économiques et financières. En particulier, ces modélisations impliquent de repenser la définition même de ce que doit être une prévision : prévision ponctuelle, prévision par intervalle ou prévision de densité. Dans le cas des modèles linéaires, les prévisions se limitent généralement à de simples prévisions ponctuelles parce que la densité de la prévision ne fait que refléter les propriétés de la loi conditionnelle de la variable dépendante, qui elle-même dépend fondamentalement de la loi supposée du terme d'erreur. Ainsi par exemple, un modèle linéaire ne génère pas par lui-même d'asymétrie ou de multimodalité dans la densité de la prévision : ces propriétés, si elles existent, ne proviennent pas de la loi supposée des erreurs.

Au contraire, l'estimation ou le calcul par simulations d'une prévision de densité prend tout son sens dans le cas des modèles non linéaires, par exemple dans le cas d'un modèle à changement de régime : cet exercice permet de révéler d'éventuelles asymétries ou multimodalités de la distribution des prévisions que ne pouvaient générer de façon endogène les modèles linéaires. Les prévisions par intervalle de confiance ou *High Density Regions* (HDR) permettent alors de révéler la présence d'asymétrie ou de multimodalités des prévisions. Ces intervalles non nécessairement symétriques et non nécessairement continus, peuvent être très utile aux décideurs publics pour par exemple établir des scénarios. De la même façon, les prévisions de densité asymétriques et multimodales fournissent de précieux renseignements sur l'incertitude autour de la prévision ponctuelle. L'exemple typique est celui des prévisions d'inflation de la Banque Centrale d'Angleterre par densité, reportées dans les fameux *Inflation Fan Chart Reports* (voir figure 3.1) qui permettent aux opérateurs de marché de parfaitement visualiser l'incertitude des prévisions de la Banque Centrale et donc de mieux anticiper ses actions sur les taux directeurs. Au-delà des prévisions d'inflation, cette évolution a touché de nombreux domaines des prévisions financières et macroéconomiques.

Ces différentes formes de prévisions qui sont de plus en plus utilisées en pratique, nécessitent de nouvelles méthodes de validation des modèles de prévision (Corradi et Swanson (2006)). Globalement, l'évolution de ces méthodes de validation est allée dans le sens d'une plus grande prise en compte de l'incertitude autour de la prévision ponctuelle. En effet, tester la validité d'un modèle de prévision ne peut se résumer à tester la validité de sa prévision (ponctuelle) moyenne. Il est évident qu'il est important pour une Banque Centrale de tester la validité de la prévision ponctuelle de l'inflation fournie par son modèle interne (2% par exemple). Mais si ce modèle conduit à une prévision symétrique autour de cette valeur alors que dans les faits il y a deux fois plus de chances d'observer une inflation supérieure à 2% qu'inférieure à ce seuil, alors la validité du modèle doit être rejetée.

Dans ce chapitre, nous proposons une synthèse des principales méthodes de validation des prévisions de densité. En ce qui concerne les techniques usuelles de validation des prévisions ponctuelles et de prévision par intervalle de confiance, le lecteur intéressé trouvera plus de précisions dans l'article « Modèles non linéaires et prévisions » de Colletaz et Hurlin (2007). Plusieurs remarques doivent être faites à ce stade.

Premièrement, les techniques d'évaluation des prévisions de densité diffèrent sensiblement des techniques de validation des prévisions ponctuelles et par intervalle de confiance. Pour ces dernières, les tests de validation sont généralement fondés sur des fonctions de pertes associées aux erreurs de prévisions définies par rapport aux prévisions ponctuelles ou aux séquences de violation dans le cas de prévision par intervalle de confiance. On définit une violation comme une situation dans laquelle la valeur ex-post de la variable prévue n'est pas comprise au sein de l'intervalle construit ex-ante. Dans le cas de l'évaluation des prévisions de densité, les tests de validation ne se fondent plus nécessairement sur une erreur de prévision. Ils visent généralement à tester l'adéquation entre la densité conditionnelle de prévision et la densité du processus générateur des données (*Data Generating Process*, DGP) telle que révélée par les réalisations de la variable à prévoir. Un modèle de prévision est alors jugé valide si la densité conditionnelle de prévision est « proche » du DGP.

Deuxièmement, quelle que soit la forme de la prévision à évaluer, on oppose généralement l'évaluation des prévisions sur la période d'estimation du modèle (*in-sample*) à l'évaluation hors période d'estimation (*out-of-sample*). Ce problème, classique en économétrie, n'est pas spécifique à la validation des densités et il ne fera donc pas l'objet de discussion dans ce chapitre. Toutefois il convient de garder à l'esprit, que l'évaluation *out-of-sample* requiert une estimation du modèle de type récursive, fixe ou glissante (*rolling estimate*), et donc un ajustement des formules des statistiques de tests et de leur distribution.

La troisième remarque porte sur l'opposition entre les méthodes dites d'évaluation « absolue » et des méthodes d'évaluation « relative » des modèles de prévisions. En effet, quelle que soit la forme de la prévision (ponctuelle ou densité), il existe deux grandes classes de tests de validation. La première classe regroupe tous les tests qui s'attachent à tester la validité des prévisions issues d'un modèle en particulier et qui admettent pour hypothèse nulle la validité du modèle de prévision. A l'inverse, d'autres tests reposent au contraire sur une évaluation relative des prévisions issues de deux modèles ou plus généralement d'un nombre fini de modèles. L'exemple typique est celui du test de Diebold et Mariano (1995) qui permet d'évaluer si les prévisions ponctuelles issues de deux modèles alternatifs sont équivalentes au regard d'une fonction de perte calculée à partir des erreurs de prévisions. Dans ce cas, le rejet de l'hypothèse nulle implique que les prévisions d'un modèle A sont meilleures que celles issues d'un modèle B au regard d'une certaine norme, mais ce rejet n'implique en rien que le meilleur modèle soit correctement spécifié. Dans ce contexte, certains tests d'évaluation des prévisions de densité permettent de comparer un nombre fini de modèles alternatifs potentiellement tous mal spécifiés par rapport à un modèle de référence qui lui-même peut être mal spécifié. Ces tests permettent ainsi de choisir parmi un ensemble de modèle le moins mauvais (ou le meilleur) par rapport à un modèle de référence, mais en aucun cas ils ne garantissent la validité (au sens de la distance avec la densité du DGP) de la prévision de densité obtenue à partir du meilleur modèle testé.

Ce chapitre sera organisé de la façon suivante. Dans les deux premières sections, nous étudierons en détail les prévisions de densité et présenterons les approches d'évaluations proposées dans la littérature académique. Nous verrons tout d'abord les tests statiques de spécification correcte d'une densité, avec ou

sans hypothèse de stationnarité. Puis, nous nous concentrerons sur les tests de spécification dynamique correcte. Enfin, dans la troisième et dernière section, nous mettrons en pratique certains tests évoqués en les appliquant à des prévisions de densité de rendements de différents indices boursiers.

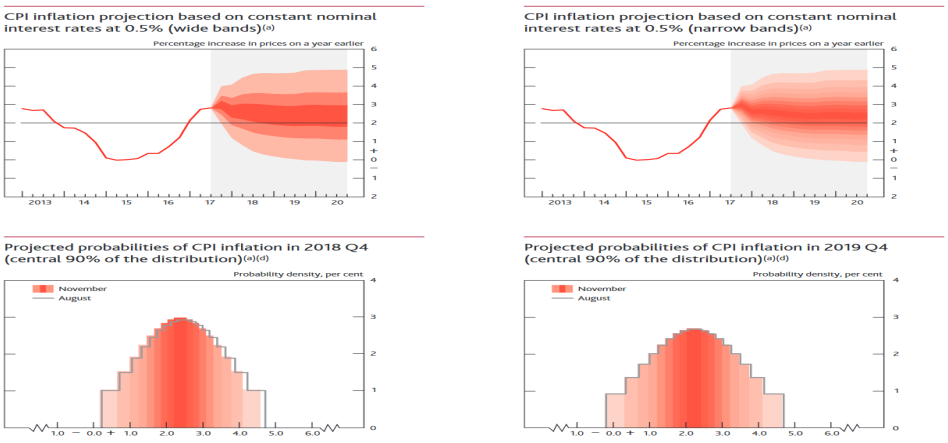

FIGURE 3.1 – Exemple de prévision de densité : graphique d'évolution possible de l'inflation, Novembre 2017 (Banque d'Angleterre)

3.2 Techniques de prévisions de densité

La prévision de densité fournit bien plus d'informations au décideur économique que la seule espérance de cette distribution qui est reportée dans le cas de la prévision ponctuelle. En particulier, elle permet d'appréhender (*i*) les éventuelles asymétries de la distribution conditionnelle de la prévision et (*ii*) la possibilité d'existence de plusieurs modes. En effet, contrairement au cas d'un modèle linéaire, un modèle non linéaire peut engendrer une distribution de prévisions asymétrique, et cela même dans le cas d'une distribution symétrique des résidus. Or, comme le note Teräsvirta (2006), cette information est essentielle : si par exemple, dans le cas d'une prévision d'inflation, la densité de la prévision présente une skewness positive, cela implique que les erreurs associées à la prévision ponctuelle ont une plus forte probabilité d'être positives que négatives. Ce qui peut nécessiter des réponses de politique économique différentes. L'examen des prévisions de densité permet en outre de mettre en évidence l'existence d'éventuelles caractéristiques bi-modales ou multi-modales, même si ce phénomène est a priori relativement rare sur séries macro-économiques. On distingue principalement trois techniques de prévisions de densités : l'approche paramétrique, l'approche non paramétrique et les prévisions d'ensemble.

L'approche paramétrique se déroule de la façon suivante : on postule une distribution paramétrique sur le terme d'erreur, on estime les paramètres et on en déduit la prévision de densité sur l'endogène. On peut citer par exemple la régression linéaire. Même si l'on utilise cette régression afin d'effectuer des prévisions ponctuelles, le modèle associé nous permet également d'obtenir une prévision de densité. Cependant, il est clair que cette modélisation a ses limites. En effet, cela ne permet pas de modéliser des variables discrètes, ni de tenir compte d'une relation non linéaire entre la variable expliquée et les variables explicatives. Afin de pallier ces problèmes, Nelder et Wedderburn (1972) proposent de ne plus se limiter à une distribution Normale des termes d'erreur, mais d'étendre les distributions possibles de la variable à expliquer aux distributions exponentielles, ce qui inclut bien entendu la loi Normale. Ce type de modèle se nomme « modèle linéaire généralisé » (*Generalized Linear Model* - à ne pas confondre avec le modèle linéaire général). Un GLM est constitué de 3 composantes : une composante aléatoire, une composante déterministe et une fonction de lien. La composante aléatoire fait référence à la distribution de la variable à expliquer, et s'écrit de la façon suivante : $f_{\alpha_i}(y_i) = \exp\left(\frac{\alpha_i y_i - b(\alpha_i)}{a(\phi)} + c(y_i, \phi)\right)$, avec a, b, c des fonctions spécifiées selon le type de la fonction exponentielle et ϕ un paramètre de dispersion, constant. On notera que seul les α_i sont estimés ici. Le choix de la distribution se fait en général natu-

rellement selon le type de données à disposition : Bernoulli pour des données binaires, Poisson pour des données de comptage etc... La composante déterministe est un prédicteur linéaire qui n'est autre qu'une combinaison linéaire des variables explicatives : $\eta(x_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. La fonction de lien est une fonction g inversible qui relie l'espérance de la variable à expliquer à la composante déterministe : $g(\mathbf{E}[Y_i]) = \eta(X_i)$. La fonction de lien g la plus souvent retenue est la fonction dite canonique qui est telle que $g(u) = (b^{'})^{-1}(u)$. Ainsi, l'espérance de la variable Y s'écrit $b^{'}(\eta(X_i))$. On peut également montrer que sa variance s'écrit $\phi b^{''}(\eta(X_i))$. Pour ce qui est d'estimer les paramètres β_i du modèle, la méthode des moindres carrés re-pondérés itérativement pour le calcul du maximum de vraisemblance est très populaire.

L'approche non paramétrique est basée sur un estimateur de noyau (ou autre) de la densité obtenue à partir des réalisations des tirages de Bootstrap ou de Monte Carlo. En effet, les simulations de Monte Carlo ou de Bootstrap $\widehat{y}^{(i)}_{t+h|t}$ peuvent être considérées comme des réalisations de la distribution conditionnelle théorique $g\left(y_{t+h} | \, \Omega_t\right)$ dont l'espérance correspond à la prévision optimale ponctuelle $\widehat{y}_{t+h|t} = E\left(y_{t+h} | \, \Omega_t\right)$. A partir de ces réalisations, on peut dès lors construire un estimateur à noyau de la densité conditionnelle.

Enfin, un autre moyen d'effectuer des prévisions de densité est de faire des prévisions d'ensemble. Ce type de prévisions est très utilisé dans le domaine de la météorologie. La caractéristique importante de la dynamique atmosphérique est qu'elle est très sensible, dans certaines conditions, à la moindre fluctuation (ce que l'on nomme « effet papillon », à la suite d'Edward Lorenz). Par ailleurs, on connaît assez bien, mais pas parfaitement, l'état actuel de l'atmosphère, et ses lois d'évolution. C'est pourquoi plusieurs prévisions numériques sont réalisées à l'aide de conditions initiales légèrement différentes mais plausibles étant données les limites de résolution des observations et des équations. On peut alors utiliser ces prévisions afin d'obtenir une estimation de la densité du processus étudié. Le contexte est très différent dans les domaines financier et économétrique, et ce genre de prévisions n'y est actuellement pas très courant. On note toutefois, que la Banque d'Angleterre utilise des méthodes qui s'en rapprochent (agrégation d'experts), pour produire ses graphiques d'évolution possible de l'inflation [1] comme on peut le voir dans les articles de Wallis (2003), de Mitchell et Hall (2005) ou encore de Galbraith et van Norden (2012) par exemple.

3.3 Évaluation des prévisions de densité

Les premiers tests d'évaluation des prévisions de densité se sont attachés à déterminer si les réalisations des prévisions étaient issues d'une distribution identique à la vraie distribution des données, c'est-à-dire à la densité issue du processus générateur des données (DGP). Ce sont donc avant tout des tests de spécification correcte de la prévision de densité, qui vérifient que la forme fonctionnelle de la prévision de densité est correcte.

Le test pionnier dans cette perspective est sans conteste celui de Diebold, Gunther et Tay (1998) qui repose sur la transformation probabiliste (*Probability Integral Transform* - PIT) de Rosenblatt (1952). Rappelons le principe de cette transformation. Soit $f_t\left(y | \, \Omega_{t-1}, \theta_0\right)$ la densité conditionnelle des prévisions (à l'horizon $h = 1$) de la variable y_t obtenue dans le modèle de référence (AR, SETAR, ou autre) conditionnellement à l'ensemble d'information Ω_{t-1} disponible à la date $t-1$ et où θ_0 désigne un ensemble de paramètres connus. Soit $\{y_t\}_{t=1}^{T}$ la séquence des réalisations de la variable y sur la période d'évaluation des prévisions du modèle. Sous l'hypothèse que la distribution conditionnelle des prévisions associée à ce modèle corresponde effectivement au DGP, alors les variables transformées :

$$z_t = F_t\left(y_t | \, \Omega_{t-1}, \theta_0\right) = \int_{-\infty}^{y_t} f_t\left(u \, | \Omega_{t-1}, \theta_0\right) du \quad t = 1, ..., T$$

sont identiquement et indépendamment distribuées selon une loi Uniforme sur $[0, 1]$. Autrement dit, lorsque à chaque date t, la densité conditionnelle des prévisions associée au modèle testé correspond à la vraie distribution conditionnelle associée au DGP, la séquence des variables transformées $\{z_t\}_{t=1}^{T}$ est *i.i.d.* $U_{[0,1]}$. Par conséquent, dans cette perspective, une manière évidente d'évaluer si la spécification de

1. Inflation fan charts. https ://www.bankofengland.co.uk/

la densité conditionnelle des prévisions est correcte revient à tester l'adéquation de la distribution des variables transformées z_t à une loi Uniforme.

Diebold, Gunther et Tay proposent ainsi de comparer sur un même graphique la fonction de répartition empirique des variables transformées et la fonction de répartition théorique de la loi Uniforme, *i.e.* la droite à 45°. Si le modèle utilisé pour effectuer la prévision ne permet pas d'obtenir une spécification correcte de la densité conditionnelle des prévisions, on doit alors observer un écart important entre les deux fonctions de répartition. Une première façon de tester si la prévision de densité est correcte consiste à construire un intervalle de confiance à $1-\alpha\%$ autour de la fonction de répartition théorique de la loi Uniforme. Clements et Smith (2001) utilisent pour cela les valeurs critiques exactes de la statistique de Kolmogorov-Smirnov pour des petits échantillons de taille T. Pour un intervalle de confiance à 95% ($\alpha = 0.025$), cet intervalle est donné [2] par la droite à $45° \mp \sqrt{\ln(1/\alpha)/(2T)}$.

Une seconde façon consiste à utiliser directement un test non paramétrique d'adéquation de loi, comme par exemple le test de Kolmogorov-Smirnov. Le principe est alors le même : il s'agit de tester l'adéquation entre la fonction de répartition $F_t(y \mid \Omega_{t-1}, \theta_0)$ et celle d'une loi Uniforme sur $[0,1]$. Mais deux problèmes se posent à ce niveau-là. Tout d'abord, les paramètres θ_0 du modèle utilisé pour effectuer la prévision ne sont généralement pas connus mais estimés. L'erreur d'estimation de ces paramètres vient alors « polluer » la distribution conditionnelle des prévisions, ce qui peut notamment induire un rejet à tort de l'hypothèse nulle de spécification correcte. Sur le plan technique, la distribution asymptotique des tests usuels d'adéquation de loi (Kolmogorov-Smirnov) dépend alors de paramètres de nuisance. De plus, l'ensemble d'information Ω_{t-1} n'est généralement pas observé dans son intégralité, seul un sous ensemble tronqué est observé.

Par la suite, on a assisté à l'émergence des tests de spécification dynamique correcte, qui vérifient non seulement que les variables transformées z_t suivent bien une distribution Uniforme, mais également qu'elles sont indépendantes. De ce fait, ils ne sont utilisés que pour évaluer des prévisions à horizon $h = 1$, car les prévisions à horizons supérieurs entraînent fatalement une autocorélation des erreurs de prévisions et donc une dépendance, même si l'on utilise la vraie distribution du processus étudié. L'hypothèse alternative de mauvaise spécification peut dès lors se traduire par une violation de l'une et/ou de l'autre hypothèse.

3.3.1 Les tests statiques de spécification correcte

Lorsque l'on effectue des tests sur les prévisions de densité, il va de soi que de nombreuses hypothèses sont prises en compte. L'une d'entre elle, très souvent mentionnée, se réfère à une stationnarité supposée du processus. Afin d'alléger cette partie, nous évoquerons des résultats obtenus sous hypothèse nulle (qui peut varier selon les tests). On notera que d'autres hypothèses que celle concernant la stationnarité du processus sont en réalité nécessaires. Pour ces hypothèses, nous renvoyons aux articles originaux cités dans notre chapitre. Lorsque l'estimation du paramètre θ n'est pas explicitée, on considérera que les auteurs utilisent pour leurs applications l'estimateur du maximum de vraisemblance ou un estimateur cohérent qui converge à vitesse $\sqrt{T}$, où T est la taille de l'échantillon à disposition. Ces remarques sont également valables pour les tests de spécification dynamique correcte.

Tests statiques avec hypothèse de stationnarité

Bai (2003) propose un test général de spécification correcte d'une distribution conditionnelle qui règle les problèmes issus de l'estimation des paramètres. De façon générale, l'hypothèse nulle du test est la suivante : la distribution cumulative conditionnelle de la variable y_t est dans la famille paramétrique $F_t(r \mid \Omega_{t-1}, \theta_0)$. En d'autres termes :

$$H_0 : \mathbb{P}(y_t \leq y | \Omega_{t-1}, \theta_0) = F_t(y | \Omega_{t-1}, \theta_0), \;\; \theta_0 \in \Theta,$$

avec F_t la vraie distribution et Θ l'ensemble des paramètres. Le principe du test de Bai reste alors similaire au test de Kolmogorov-Smirnov : il s'agit de vérifier l'adéquation de la fonction de répartition de la séquence des processus $z_t = F_t(y_t \mid \Omega_{t-1}, \theta_0)$ avec celle d'une loi Uniforme sur $[0,1]$ pour $t = 1, ..T$.

2. Pour plus de détails, voir Miller (1956).

Mais contrairement au cas standard, les paramètres θ_0 ne sont pas connus mais seulement estimés. De plus, Bai suppose que seul un sous ensemble de l'ensemble d'information est observable, et donc utilisable pour la prévision. Soit $\widetilde{\Omega}_{t-1} = \{y_{t-1}, .., y_1, z_{t-1}, .., z_1\}$ un sous ensemble observable et tronqué de Ω_{t-1}. On note $\widehat{\theta}$ l'estimateur de θ_0. Les variables transformées établies à partir des paramètres estimés $\widehat{\theta}$ sont notées $\widehat{z}_t$ et vérifient :

$$\widehat{z}_t = F_t\left(y_t|\,\widetilde{\Omega}_{t-1}, \widehat{\theta}\right).$$

Bai propose un test de type Kolmogorov-Smirnov couplé à une transformation martingale de Khmaladze (1982), ce qui permet d'obtenir un test libre de paramètres de nuisance et dont les valeurs critiques sont faciles à calculer.

Soit $\widehat{V}_T(r)$ un processus empirique défini à partir de la séquence des variables transformées $\{\widehat{z}_t\}_{t=1}^T$ tel que :

$$\widehat{V}_T(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[\mathbb{I}\left(\widehat{z}_t \leq r\right) - r\right].$$

Soit $V_T(r)$ le processus équivalent défini quant à lui, à partir des vraies transformées probabilistes z_t établies sur la base de la vraie distribution conditionnelle, *i.e.* $z_t = F_t\left(y_t|\,\Omega_{t-1}, \theta_0\right)$.

Bai montre que le processus $\widehat{V}_T(r)$ admet la représentation asymptotique suivante :

$$\widehat{V}_T(r) = V_T(r) - \overline{g}(r)' \sqrt{T}\left(\widehat{\theta} - \theta_0\right) + o_p(1), \tag{3.1}$$

avec $\overline{g}(r)$ la probabilité limite de la somme (divisée par $1/T$) pour $t = 1 \cdots T$, des dérivée partielles de la vraie fonction de répartition du processus par rapport à θ, évaluée en $F_t^{-1}(r \mid \Omega_{t-1}, \theta_0)$.

Le deuxième terme de droite de la relation (3.1) met en évidence le fait que le processus $\widehat{V}_T(r)$ dépend asymptotiquement de la vraie distribution $F_t\left(r|\,\Omega_{t-1}, \theta_0\right)$ et de ce fait, des vrais paramètres θ_0.

Toute l'astuce du test de Bai consiste alors à appliquer une transformation qui permet de retirer le terme $\overline{g}(r)' \sqrt{T}\left(\widehat{\theta} - \theta_0\right)$ et d'obtenir un processus transformé admettant une distribution asymptotique correspondant à un mouvement Brownien. Cette transformation est la suivante :

$$\widehat{W}_T(r) = \widehat{V}_T(r) - \int_0^r \left[\dot{g}(s)' C^{-1}(s) \int_s^1 \dot{g}(\tau)\, d\widehat{V}_T(\tau)\right] ds,$$

où $g(r) = \left(r, \overline{g}(r)'\right)'$, $\dot{g}(r) = (1, \dot{\overline{g}}(r)')'$ et $C(r) = \int_r^1 \dot{g}(\tau)\, \dot{g}(\tau)'\, d\tau$. Bai montre que cette statistique converge vers un mouvement Brownien standard, c'est-à-dire vers une distribution totalement indépendante du modèle considéré et de la vraie valeur des paramètres de ce modèle. Il propose de considérer la statistique suivante :

$$T_T = \sup_{0 \leq r \leq 1} \left|\widehat{W}_T(r)\right| \xrightarrow{d} \max_{0 \leq r \leq 1} |W(r)|,$$

où $W(.)$ est un mouvement Brownien standard. L'obtention des valeurs critiques de ce test se fait par simulation, et les valeurs suivantes à un niveau de significativité 10%, 5% et 1% sont respectivement 1.94, 2.22 et 2.80.

Le test de Bai n'a une puissance asymptotique unitaire que dans le cas d'alternatives dans lesquelles il y a une violation de l'hypothèse d'uniformité. En revanche, ce test est peu puissant contre des alternatives pour lesquelles il y a violation de l'indépendance. En effet, en cas de mauvaise spécification dynamique, bien que ce ne soit pas la propriété testée, le test échouera à rejeter l'hypothèse nulle.

Un autre test très important dans la littérature est celui de Corradi et Swanson (2006). Les auteurs proposent un test de spécification correcte de la densité robuste à une mauvaise spécification dynamique. En d'autres termes, ils considèrent sous l'hypothèse nulle, à l'instar de Bai, que seul un sous ensemble de l'ensemble d'information Ω_{t-1} est observé et utilisé pour la prévision [3] mais cherchent à vérifier si

3. Les ensembles d'information $\widetilde{\Omega}_{t-1}$ sont liés au modèle de prévision, et non au choix du test de Bai ou Corradi et Swanson.

la distribution conditionnelle de y_t est correctement spécifiée compte tenu de l'ensemble d'information retenu. On a donc :

$$H_0 : \mathbb{P}(y_t \leq y|\widetilde{\Omega}_{t-1}, \theta^\dagger) = F_t(y|\widetilde{\Omega}_{t-1}, \theta^\dagger), \;\; \theta^\dagger \in \Theta\,.$$

Ce test présente l'avantage de converger à un taux paramétrique. En contrepartie, puisqu'il prend en compte sous l'hypothèse nulle un ensemble d'information qui ne contient pas nécessairement toutes les informations historiques utiles, la distribution asymptotique de leur test n'est pas libre de paramètres de nuisance. Ainsi, les valeurs critiques associées doivent être simulées selon des procédures assez sophistiquées de Bootstrap. Le test proposé considère la statistique suivante :

$$\widehat{V}_{1T} = \sup_{r\in[0,1]} |\widehat{V}_T(r)|\,,$$

où $\widehat{V}_T(r)$ est identique au processus évoqué dans le test de Bai, à la différence possible près, que les auteurs estiment le paramètre θ par maximum de vraisemblance. On notera cependant que contrairement au test de Bai, il n'y a pas besoin d'une manipulation qui vise à annuler la « pollution » due aux erreurs d'estimations.

On remarque que si $\widetilde{\Omega}_{t-1} = \Omega_{t-1}$, alors le test devient un test de spécification dynamique correcte. Pour ce qui est du test en lui-même, un théorème nous dit que sous H_0 :

$$\widehat{V}_{1T} \Rightarrow \sup_{r\in[0,1]} |V_1(r)|\,,$$

avec V_1 un processus Gaussien de moyenne nulle et de fonction de covariance $K(r, r^{'})$. [4] Corradi et Swanson proposent également un deuxième test qui n'est plus basé sur l'étude des variables transformées z_t mais qui peut être perçu comme une extension du test de Kolmogorov conditionnel de Andrews (1997) [5].

Cependant, les tests de type Kolmogorov-Smirnov présentent les désavantages d'être en général peu puissant, et de ne pas donner plus d'indications quant à l'éloignement d'une distribution Uniforme lors du rejet de l'hypothèse nulle. C'est pourquoi on note un certain intérêt depuis quelques années pour les tests de type *Neyman's smooth test* (que nous évoquerons plus tard dans la section 3.3.2) ainsi que pour l'étude d'une transformation des $\widehat{z}_t$. La plus répandue d'entre elle consiste à étudié la série $\{\Phi^{-1}(\widehat{z}_t)\}_{t=1}^T$, où Φ^{-1} représente la fonction réciproque de la fonction de répartition d'une variable qui suit une loi Normale standard. Cette transformation est appelée *Inverse Normal Transformation* (INT). Si $z_t \sim \mathcal{U}[0,1]$, alors $\Phi^{-1}(z_t) \sim \mathcal{N}(0,1)$. Les tests de normalité sont plus nombreux que les tests d'uniformité et il est plus simple de tester l'hypothèse d'autocorélation pour une distribution Normale que pour une distribution Uniforme (Mitchell et Wallis, 2001) ce qui explique que l'on utilise généralement cette transformation.

Knüppel (2011) propose un test d'adéquation qui ne repose plus sur la comparaison de la fonction de répartition des z_t avec celle d'une distribution Uniforme, mais sur l'étude de N moments. Plus précisément, il s'intéresse aux moments ordinaires, appelés ainsi en opposition aux moments centrés ou standardisés (centrés réduits) tels que la variance et la skewness. Par exemple, les moments ordinaires d'ordres 2 et 3 d'une variable Z sont $\mathbb{E}[Z^2]$ et $\mathbb{E}[Z^3]$, quand le moment centré d'ordre 2 est la variance $\mathbb{E}[(Z-\mu)^2]$ et le moment standardisé d'ordre 3 est la skewness $\mathbb{E}\left[\left(\frac{Z-\mu}{\sigma}\right)^3\right]$. Ce test est basé sur une transformation de la série des z_t (pas nécessairement INT, bien que ce soit l'exemple choisi dans l'article), et n'est pas libre de paramètres de nuisance. L'hypothèse nulle est la suivante :

$$H_0 : \{z_{\phi,t}\}_{t=1}^T \sim \mathcal{N}(0,1)\,,$$

où $z_{\phi,t} = \Phi^{-1}(\widehat{z}_t)$, avec $\widehat{z}_t$ défini comme précédemment. La statistique du test proposée est :

$$\widehat{\alpha}_{r_1\, r_2 \cdots r_N} = T\, \widehat{D}^{'}_{r_1\, r_2 \cdots r_N}\, \widehat{\Omega}^{-1}_{r_1\, r_2 \cdots r_N}\, \widehat{D}_{r_1\, r_2 \cdots r_N}\,,$$

4. Pour plus de détails sur la variance, le lecteur peut consulter l'article « *predictive density evaluation* » de Corradi et Swanson (2006).

5. Pour plus de détails concernant les tests de Bai et Corradi et Swanson dans un cadre d'évaluation *out-of-sample*, le lecteur peut consulter également l'article « *predictive density evaluation* » de Corradi et Swanson (2006).

avec $\widehat{D}_{r_1\, r_2 \cdots r_N} = [\widehat{m}_{r_1} - m_{r_1}, \widehat{m}_{r_2} - m_{r_2}, \cdots, \widehat{m}_{r_N} - m_{r_N}]'$, $\widehat{\Omega}_{r_1\, r_2 \cdots r_N}$ la matrice de covariance du vecteur $d_t = [z_{\phi,t}^{r_1} - m_{r_1}, z_{\phi,t}^{r_2} - m_{r_2} \cdots, z_{\phi,t}^{r_N} - m_{r_N}]'$, $m_{r_i} = \mathbb{E}[z_{\phi,t}^{r_i}]$, le moment ordinaire d'ordre r_i associé et $\widehat{m}_{r_i}$ sa contrepartie empirique. On suppose par ailleurs $r_1 < r_2 < \cdots < r_N$. Sous l'hypothèse nulle de spécification correcte de la densité, la statistique converge vers un χ^2 à N degrés de liberté.

Ce test est facile (i) à mettre en place, (ii) peut s'appliquer à tous les moments jugés importants, (iii) utilise des valeurs critiques standard et (iv) a une taille correcte asymptotiquement. En revanche, des problèmes de puissance sont mis en évidence sur de petits échantillons. Cependant, Knüppel propose un autre test qui a les mêmes propriétés asymptotiques mais plus puissant, et basé sur le fait que si la densité de la transformation choisie est symétrique en 0, alors $cov(z_{\phi,t}^{r_i} - m_{r_i}, z_{\phi,t}^{r_j} - m_{r_j}) = 0$ si $r_i + r_j$ est impair. Cette statistique est la suivante :

$$\widehat{\alpha}^0_{r_1\, r_2 \cdots r_N} = \widehat{\alpha}^{pair}_{r_1\, r_2 \cdots r_N} + \widehat{\alpha}^{impair}_{r_1\, r_2 \cdots r_N},$$

où les $\widehat{\alpha}^{\bullet}_{r_1\, r_2 \cdots r_N}$ sont calculés de la même façon que précédemment, mais en utilisant uniquement les moments ordinaires pairs et impairs respectivement.

Test statique sans hypothèse de stationnarité

Comme le montre la section précédente, bon nombre de tests de spécification correcte de la prévision de densité supposent que le processus étudié est stationnaire. En pratique, cela n'est pas toujours le cas ... C'est pourquoi des tests de spécification correcte robustes à la présence d'instabilités (dues à la non-stationnarité du processus étudié, hypothèse qui est donc prise en compte dans ce type de tests) ont été mis en place. Rossi et Sekhposyan (2013) proposent une méthode d'évaluation de la spécification correcte de la prévision de densité basée sur des tests de type Kolmogorov-Smirnov ou Cramér-von Mises. Leurs tests sont robustes à la mauvaise spécification dynamique, même lorsque cette spécification ne concerne qu'une petite portion de l'échantillon à disposition (à cause des instabilités).

Leurs tests reposent sur une évaluation de la prévision de densité *out-of-sample*. On considère que l'on dispose d'un échantillon de taille $T+h$ (h étant l'horizon pour lequel la prévision est effectuée), que l'on a divisé en deux parties : une partie *in-sample* de taille R qui sert à estimer les paramètres et une partie *out-of-sample* de taille P à partir de laquelle les prévisions sont effectuées, puis testées. On a par ailleurs l'égalité $R + P - 1 + h = T + h$. Ici, on se place dans le cas suivant : P, $R \to \infty$ lorsque P, $T \to \infty$ et $\lim_{T\to\infty} P/R = \pi$, $0 < \pi < \infty$.

Ces tests sont une extension de celui de Corradi et Swanson, lorsque l'on se replace dans un cadre d'évaluation *out-of-sample* et pour un horizon de prévision $h > 1$. La différence se situe principalement au niveau de l'hypothèse nulle du test :

$$H_0 : F_t\left(y_{t+h} | \widetilde{\Omega}_t\right) = F_0\left(y_{t+h} | \widetilde{\Omega}_t, \theta^\dagger\right) \ \forall\ t = R \cdots T,$$

avec $F_0\left(y_{t+h} | \widetilde{\Omega}_t, \theta^\dagger\right) \equiv \mathbb{P}\left(y_{t+h} \leq y | \widetilde{\Omega}_t, \theta^\dagger\right)$ et $\theta^\dagger$ la limite en probabilité de $\widehat{\theta}_{t,R}$, qui représente ici l'estimateur du maximum de vraisemblance pour la période d'estimation considérée. Le but n'est plus de tester si la série des $\widehat{z}_t$ est Uniforme en moyenne sur la période d'évaluation mais plutôt de vérifier qu'elle suit une distribution Uniforme en tout point dans le temps durant cette période. Aussi, le fait d'évaluer les densités en prenant en considération la limite en probabilité $\theta^\dagger$ sous l'hypothèse nulle permet de corriger la distribution asymptotique (qui n'est donc pas libre de paramètres de nuisance) sans avoir recours à la méthode Bootstrap. D'un point de vue mathématique, la différence d'hypothèse nulle se traduit par la prise en compte dans le test d'un nouvel élément construit de façon similaire à $\widehat{V}_T$:

$$\widehat{V}_{P_\tau}(\tau, r) = \frac{1}{\sqrt{P}} \sum_{t=R}^{R+[\tau P]} \left[\mathbb{I}\left(F_t\left(y_{t+h} | \widetilde{\Omega}_t, \widehat{\theta}_{t,R}\right) \leq r\right) - r\right],$$

avec $[\tau P]$ la partie entière de τP, $\tau \in \Upsilon \subset (0,1)$.

Rossi et Sekhposyan proposent d'étudier deux statistiques, respectivement de type Kolmogorov-Smirnov et Cramér-von Mises :

$$\begin{aligned} \kappa_p &\equiv \sup_{\tau\in\Upsilon}\sup_{r\in[0,1]} Q_P(\tau,r) \text{ et} \\ C_p &\equiv \int_\tau\int_r Q_P(\tau,r)\,d\tau\,dr \ , \end{aligned}$$

où $Q_P(\tau,r)$ est le carré d'une combinaison de $\widehat{V}_{P_\tau}(\tau,r)$ et $\widehat{V}_T$. Bien que ces tests soient mis en œuvre sur $Q_P(\tau,r)$ plutôt que sur la valeur absolue de $\widehat{V}_T$ comme le font Corradi et Swanson, les tests restent équivalents car cela ne change que les valeurs critiques. Rossi et Sekhposyan montrent que leurs tests sont plus puissants et qu'ils possèdent de bonnes performances sur des échantillons de petites tailles.

Synthèse

En résumé, le test de Bai a une distribution libre de paramètres de nuisance, mais qu'il est peu puissant contre les alternatives pour lesquelles il y a violation de l'indépendance. De ce fait, il est préférable de l'utiliser pour des données non chronologiques.
Pour les séries temporelles, on pourra utiliser le test de Corradi et Swanson, qui n'a pas ce problème relatif à l'indépendance. De plus, sa distribution asymptotique converge à un taux paramétrique. Cependant, le principal inconvénient de ce test réside dans la nécessité de simuler les valeurs critiques par Bootstrap car il n'est pas libre de paramètres de nuisance. Par ailleurs, il est moins puissant que les tests de Rossi et Sekhposyan, qui présentent en plus l'avantage de ne pas considérer l'hypothèse de stationnarité du processus étudié. Aussi, la correction des distributions asymptotiques de leurs tests est moins coûteuse que la méthode Bootstrap.
Enfin, le test de Knüppel se distingue par l'utilisation d'une double transformation. Cela conduit à vérifier l'adéquation de la série des variables transformées à une distribution Normale, ce qui n'est pas sans intérêt. En effet, les tests de normalité sont plus nombreux que les tests d'uniformité et tester l'hypothèse d'autocorélation est plus simple pour une distribution Normale que pour une distribution Uniforme. Cependant, ce test n'est pas libre de paramètres de nuisance.

3.3.2 Les tests de spécification dynamique correcte

Afin de corriger les problèmes de puissance des tests statiques face à des alternatives dans lesquelles il y a violation de l'indépendance, de nombreux tests de spécification dynamique correcte ont été mis en place. Une fois encore, cette section sera séparée en fonction de la stationnarité ou non du processus étudié.

Tests de spécification dynamique correcte avec hypothèse de stationnarité

Hong et Li (2004) proposent un test de spécification dynamique correcte pour des prévisions à horizon $h = 1$ sous l'hypothèse nulle, c'est-à-dire :

$$H_0 : \{z_t\}_{t=1}^T \ \ i.i.d \ \mathcal{U}[0,1].$$

Ce test est fondé sur la comparaison de la densité non paramétrique jointe de (z_t, z_{t-j}), avec $j > 0$, à la densité jointe de deux variables Uniformes sur $[0,1]$. Pour cela, ils introduisent un estimateur à noyau modifié qui permet d'obtenir un bon estimateur non paramétrique y compris sur les bords du domaine de définition, c'est-à-dire aux alentours de 0 et de 1. Définissons :

$$\widehat{\phi}(u_1,u_2) = (n-j)^{-1}\sum_{t=j+1}^{n} K_\lambda(u_1,\widehat{z}_t)K_\lambda(u_2,\widehat{z}_{t-j})\,,$$

avec

$$K_\lambda(x,y) = \begin{cases} \lambda^{-1}\left(\frac{x-y}{\lambda}\right) / \int_{-(x/\lambda)}^{1} k(u)\,du & \text{si } x \in [0,\lambda[\\ \lambda^{-1}\left(\frac{x-y}{\lambda}\right) & \text{si } x \in [\lambda, 1-\lambda[\\ \lambda^{-1}\left(\frac{x-y}{\lambda}\right) / \int_{-1}^{(1-x)/\lambda} k(u)\,du & \text{si } x \in [1-\lambda, 1]. \end{cases}$$

Ici, λ représente le paramètre de lissage (*bandwidth parameter*) et $k(\cdot)$ est une fonction de noyau.

Définissons également :

$$\widehat{M}(j) = \int_0^1 \int_0^1 \left(\widehat{\phi}(u_1,u_2) - 1\right)^2 du_1\, du_2\,,$$

et

$$\widehat{Q}(j) = \left((n-j)\widehat{M}(j) - A_\lambda^0 \right) / V_0^{1/2}\,,$$

avec

$$A_\lambda^0 = \left((\lambda^{-1} - 2) \int_{-1}^1 k^2(u)\,du + 2\int_0^1 \int_{-1}^b k_b(u)\,du\,db \right)^2 - 1\,,$$

$$k_b(.) = k(.) / \int_{-1}^b k(v)\,dv\,,$$

et

$$V_0 = 2\left(\int_{-1}^1 \left(\int_{-1}^1 k(u+v)k(v)\,dv \right)^2 du \right)^2 .$$

Sous l'hypothèse nulle de spécification dynamique correcte du modèle, $\widehat{Q}(j)$ converge vers une distribution Normale standard.

Effectuer le test pour différentes valeurs de j permet d'obtenir des informations quant au nombre de retards à partir duquel on s'éloigne significativement d'une distribution Uniforme sur [0,1]. Cependant, cela peut poser problème lorsque l'on compare deux modèles, car il est possible que pour un j donné, l'un soit meilleur que l'autre, et que ce ne soit plus le cas pour une autre valeur de j. Pour contourner ce problème, les auteurs proposent un test statistique de type porte-manteau. Ce test peut être vu comme une généralisation des tests d'autocorélation de Box-Pierce ou Ljung-Box, initialement utilisés dans un contexte de séries temporelles linéaires étudiées dans un cadre *in-sample* et que l'on étendrait aux séries non linéaires dans un cadre *out-of-sample*. Considérons :

$$\widehat{W}(p) = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \widehat{Q}(j)\,,$$

où p est un nombre arbitraire de retards. Sous H_0, $\widehat{W}(p)$ converge vers une distribution Normale standard.

En cas de rejet de l'hypothèse nulle, il est intéressant de connaître quelle est la source de ce rejet. Est-ce dû à une violation de la condition d'uniformité, de celle d'indépendance ou les deux ? Afin de savoir pourquoi l'on rejette l'hypothèse nulle du test, Hong et Li préconisent l'utilisation d'un test d'uniformité robuste à une mauvaise spécification dynamique ainsi qu'un test d'indépendance robuste à la violation de la condition d'uniformité proposés par Hong (2001).

Tout comme la statistique de Bai, celle proposée par Hong et Li présente en outre l'avantage d'avoir une distribution asymptotique standard et libre de paramètres de nuisance. L'inconvénient principal de cette approche réside dans le fait qu'elle implique le choix d'un paramètre de lissage dans la phase d'estimation non paramétrique.

Cependant, comme nous l'avons mentionné précédemment, les tests de types Kolmogorov-Smirnov ou Cramér-von Mises présentent certains désavantages qui ont conduit à s'intéresser à d'autres type de tests d'uniformité, les tests de type *Neyman's smooth test* (test lisse de Neyman). L'idée de Neyman (1937) est d'englober la distribution Uniforme sur l'intervalle [0,1] dans un ensemble plus large de densités dites alternatives, de la forme suivante :

$$g(y;\gamma) \equiv \begin{cases} \frac{\exp\left(\sum_{i=1}^{K} \gamma_i h_i(y)\right)}{z(\gamma)} \text{ , si } y \in [0,1] \\ \\ 0 \text{ sinon} \end{cases}$$

où les h_i sont des fonctions bien choisies, qui respectent certaines conditions que nous n'évoquerons pas ici [6], et $z(\gamma)$ est une constante de normalisation de telle sorte que l'intégrale de g soit égale à 1 :

$$z(\gamma) \equiv \int_0^1 \exp\left(\sum_{i=1}^{K} \gamma_i h_i(y)\right) dy\,.$$

Ainsi, tester si g correspond à la densité d'une distribution Uniforme sur [0,1] revient à tester γ=0. En effet, si c'est le cas, on obtient $g = 1$ pour $y \in$ [0,1], ce qui correspond bien à la densité d'une distribution Uniforme sur l'intervalle convoité. Park et Zhang (2010) et Lin et Wu (2017) ont eu l'idée de procéder à un test qui repose sur l'idée de Hong et Li de s'intéresser à la densité jointe de (z_t, z_{t-j}) , en utilisant un test de Neyman plutôt que Kolmogorov-Smirnov.

Park et Zhang proposent un test joint d'uniformité et d'indépendance des z_t. Ainsi, leur hypothèse nulle est :

$$H_0 : g^j(z_t, z_{t-j}) = 1\,,$$

contre une large classe d'alternatives H_A. Les fonctions h_i^j choisies dans l'article pour réaliser le test de Neyman sont de la forme : $b_i(z_t)\, b_i(z_{t-j})$, avec $b_i(z_t)$ le polynôme de Legendre orthonormal d'ordre i en z_t. Puisque vérifier l'hypothèse nulle revient à vérifier $\gamma = 0$, on peut réécrire :

$$H_0^{'} : \gamma = 0.$$

Le paramètre θ_0 étant inconnu, les auteurs divisent leur échantillon d'observations en un échantillon de taille R pour l'estimation (par maximum de vraisemblance) et un échantillon de taille P utilisé pour les prévisions. De plus, on se place dans le cas où $\lim_{T\to\infty} P_j/R = \pi$, $0 < \pi < \infty$ avec $P_j = P - j$. Le test n'est clairement pas libre de paramètres de nuisance du fait de l'estimation de θ_0. Cependant, il est possible de corriger cela grâce à un résultat obtenu par West et McCracken (1998, Lemme 4.2).

Aussi, il est évident que le choix du nombre K d'exponentielles influe sur le bon déroulement des opérations (en termes de puissance du test). Afin de choisir le K optimal, les auteurs préconisent l'utilisation d'une méthode à mettre au crédit de Inglot et Ledwina (2006), qui utilisent les critères AIC ou BIC, en fonction des données à disposition (*data-driven test*). Une fois leur statistique obtenue pour un retard j donné, Park et Zhang mettent en place un test de type porte-manteau, de façon similaire à Hong et Li.

Une comparaison est effectuée avec le test de Hong et Li. Ils reprennent pour cela les mêmes expérimentations de simulation que dans un article de Hong, Li et Zhao (2007). En termes de taille, le test de Park et Zhang semble meilleur (pour les modèles considérés). Pour ce qui est de la puissance, les résultats sont plus discutables, bien qu'en faveur de Park et Zhang. Pour le premier des trois modèles considérés, leur performance est nettement meilleure, pour le second elle est identique et légèrement moins bonne pour le troisième.

Comme nous avons pu le remarquer, ce test offre de nombreux avantages. Cependant, Lin et Wu pointent le fait que lorsque l'on constate un rejet de H_0, on ne sait pas quelle condition est violée. Ainsi,

6. Le lecteur peut se référer à la page https ://onlinelibrary.wiley.com/doi/full/10.1002/0471667196.ess0693.pub2.

ils proposent une autre solution, qui teste séquentiellement la spécification correcte d'une densité, basée sur un test de type Neyman.

L'idée évoquée dans le test de Lin et Wu est d'effectuer dans un premier temps un test d'indépendance des variables transformées $\widehat{z}_t$ robuste à la violation de la propriété d'uniformité, et de procéder ensuite, si l'hypothèse d'indépendance est validée, à un test d'uniformité. Le choix d'effectuer les tests dans ce sens est motivé par le fait que le test d'uniformité est soumis à une hypothèse d'indépendance des z_t. Ce test se déroule donc en deux étapes, qui sont toutes deux très proche du test évoqué précédemment. En effet, le test d'uniformité est une version univariée du test de Park et Zhang où les fonctions h_i^j ne sont plus restreintes aux polynômes de Legendre orthonormaux, mais sont des fonctions $h_i^j : [0,1] \to \mathbb{R}$, orthonormales avec la distribution Uniforme. Cependant, puisqu'il s'agit d'un test univarié, la mise en œuvre du test de type porte-manteau n'est pas nécessaire. On notera également que le paramètre est à nouveau estimé par maximum de vraisemblance, mais que l'on se place cette fois-ci dans le cas $\lim_{T\to\infty} P/R = \pi$, $0 \leq \pi < \infty$.

Une copule est une fonction qui permet de caractériser la dépendance entre des variables aléatoires. Ainsi, tester l'indépendance entre z_t et z_{t-j} est équivalent à tester l'hypothèse selon laquelle la densité de leur copule est constante et égale à 1. A partir de ce constat, Lin et Wu proposent un test de type Neyman basé sur la copule entre z_t et z_{t-j} afin de vérifier l'hypothèse d'indépendance des z_t. Une nouvelle fois, une différence avec le test de Park et Zhang va se situer dans le choix des fonctions h_i^j. En effet, sont concernées ici des fonctions $h_{i_1 i_2} = h_{i_1} h_{i_2}$. De plus, les auteurs utilisent une estimation empirique (et donc non paramétrique) des z_t. Ainsi, le test est libre de paramètres de nuisance. Il est en revanche cette fois-ci nécessaire de mettre en place un test de type porte-manteau à partir de la statistique obtenue.

Lin et Wu comparent leur test avec ceux de Hong et Li et Park et Zhang. Pour cela, ils reprennent eux aussi les expérimentations de l'article de Hong, Li et Zhao (2007). Les deux tests de types Neyman sont relativement proches en termes de taille, et donc meilleurs que celui basé sur un test de type Kolmogorov-Smirnov. En revanche, le test de Lin et Wu se démarque de par sa puissance, généralement supérieure aux autres tests.

Kalliovirta (2012) propose quant à elle trois tests qui requièrent la transformation INT évoquée dans la section 3.3.1, mais il s'agit ici de tests de mauvaise spécification. Les tests en question peuvent être interprétés comme des tests du Multiplicateur de Lagrange (*Lagrange Multiplier*, *LM*). Ainsi, ils sont asymptotiquement optimaux contre des alternatives locales. Par ailleurs, l'erreur d'estimation des paramètres est maintenue sous l'hypothèse nulle. De ce fait, l'obtention de la distribution asymptotique des tests ne nécessite pas de correction.

Le premier des trois tests est un test d'autocorélation. L'hypothèse nulle est donc :

$$H_0 : Corr(R_{t,\theta_0}, R_{t-k,\theta_0}) = 0, \ \forall \ t \text{ et } k > 0\,,$$

avec $R_{t,\theta} = \Phi^{-1}(F_t(y_t|\widetilde{\Omega}_{t-1}, \theta))$.

L'auteur suppose que les K_1 premières autocovariances suffisent à mettre en évidence une inadéquation entre le modèle et les données. On définit ensuite la fonction $g : \mathbb{R}^{K_1+1} \to \mathbb{R}^{K_1}$:

$$g(r_{t,\theta}) = \left[r_{t,\theta} r_{t-1,\theta}, \cdots, r_{t,\theta} r_{t-K_1,\theta}\right]',$$

avec $r_{t,\theta}$ la réalisation associée à $R_{t,\theta}$.

La statistique du test et sa distribution asymptotique sous H_0 sont données par :

$$A_{K_1} = (T - K_1)\, \frac{1}{T - K_1} \sum_{t=1+K_1}^{T} g(r_{t,\widehat{\theta}_T})' \, \widehat{\Omega}_T^{-1} \sum_{t=1+K_1}^{T} g(r_{t,\widehat{\theta}_T}) \sim \chi^2(K_1)\,,$$

avec $\widehat{\theta}_T$ l'estimateur du maximum de vraisemblance et $\widehat{\Omega}_T^{-1}$ la matrice de covariance asymptotique qui dépend de $g(r_{t,\theta})$.

L'interprétation du test LM s'inspire des travaux de Hosking (1981). Définissons un modèle auxiliaire pour les variables transformées :

$$R_{t,\theta} = \rho^{'}\mathbf{R}_{t-1,\theta} + \varepsilon_t\,,$$

avec $\varepsilon_t \sim \mathcal{N}(0,1)$, $R_{t,\theta}$ comme précédemment et $\mathbf{R}_{t-1,\theta} = [R_{t-1,\theta}, \cdots, R_{t-p,\theta}]'$, $t = 1, \cdots, T$, $R_{t,\theta} = 0$ pour $t \leq 0$. On constate que ces variables transformées sont indépendantes si $\rho = 0$. Ainsi, effectuer un test LM basé sur $\frac{\partial \widetilde{l}(\widehat{\theta}_T, 0, y)}{\partial \rho}$, avec $\widetilde{l}(\theta, \rho, y)$ la fonction de log-vraisemblance associée au modèle auxiliaire, permet de retrouver la statistique proposée par Kalliovirta.

Le second test concerne l'hétéroscédasticité conditionnelle. Il est motivé par le fait que si la série des z_t suit une distribution $\mathcal{U}[0,1]$, alors la série transformée doit vérifier la propriété d'homoscédasticité conditionnelle. L'hypothèse nulle est donc la suivante :

$$H_0 : Corr(R^2_{t,\theta_0}, R^2_{t-k,\theta_0}) = 0,\ \forall\ t \text{ et } k > 0\,.$$

Le test est basé sur une version modifiée des covariances empiriques. Une fois encore, un faible nombre K_2 d'autocovariances suffit à mettre en lumière une inadéquation entre les données et le modèle. Ainsi, on définit la fonction $g : \mathbb{R}^{K_2+1} \to \mathbb{R}^{K_2}$:

$$g(r_{t,\theta}) = [(r^2_{t,\theta} - 1)r_{t-1,\theta}, \cdots, (r^2_{t,\theta} - 1)r_{t-K_2,\theta}]^{'}\,.$$

On obtient ensuite la statistique de test et sa distribution asymptotique sous l'hypothèse nulle :

$$H_{K_2} = (T - K_1)\,\frac{1}{T - K_2} \sum_{t=1+K_2}^{T} g(r_{t,\widehat{\theta}_T})^{'}\,\widehat{\Omega}_T^{-1} \sum_{t=1+K_2}^{T} g(r_{t,\widehat{\theta}_T}) \sim \chi^2(K_2).$$

L'interprétation du test LM s'inspire cette fois des travaux de Engle (1982). L'idée est de définir le modèle auxiliaire suivant pour les variables transformées :

$$R_{t,\theta} = h_t^{-1/2}\varepsilon_t\,,$$

avec $\varepsilon_t \sim \mathcal{N}(0,1)$, $h_t = 1 + \alpha^{'}\mathbf{R}^2_{t-1,\theta}$, $\mathbf{R}^2_{t-1,\theta} = [R^2_{t-1,\theta}, \cdots, R^2_{t-p,\theta}]'$, $t = 1, \cdots, T$, $R_{t,\theta} = 0$ pour $t \leq 0$ et $R_{t,\theta}$ définit comme précédemment. La propriété d'homoscédasticité est vérifiée pour $\alpha = 0$. On retrouve la statistique proposée de la même façon que pour le premier test évoqué, mais en utilisant cette fois-ci $\frac{\partial \widetilde{l}(\widehat{\theta}_T, 0, y)}{\partial \alpha}$.

Le troisième et dernier test proposé est un test de normalité. L'hypothèse nulle est basée sur 3 moments [7] de la série des variables transformées :

$$H_0 : \mathbb{E}[R^2_{t,\theta_0} - 1\ \ R^3_{t,\theta_0}\ \ R^4_{t,\theta_0} - 3]^{'} = 0\ \forall\ t\,.$$

La fonction $g : \mathbb{R}^3 \to \mathbb{R}$ associée est :

$$g(r_{t,\theta}) = [r^2_{t,\theta} - 1\ \ r^3_{t,\theta}\ \ r^4_{t,\theta} - 3]^{'}\,.$$

L'approche de type LM est valable lorsque la série des variables transformées appartient à la famille des résidus de Pearson. De plus, contrairement aux autres tests basés sur des résidus de Pearson, on utilise $r^2_{t,\theta_0} - 1$ car il améliore les résultats obtenus sur de petits échantillons dans le cadre de modèles mixtes non linéaires. Cependant, ce terme doit être enlevé si la variance asymptotique de la série des résidus transformés est égale à 1, afin de ne pas altérer les résultats asymptotique que nous allons mettre en évidence. Sous H_0, la statistique de test et sa distribution sont données par :

$$H_N : N = T\,\frac{1}{T}\sum_{t=1}^{T} g(r_{t,\widehat{\theta}_T})^{'}\,\widehat{\Omega}_T^{-1}\,\sum_{t=1}^{T} g(r_{t,\widehat{\theta}_T}) \sim \chi^2(3).$$

7. Le moment centré d'ordre 2 et les moments standardisés d'ordre 3 et 4, qui coïncident ici avec les moments ordinaires.

L'interprétation du test LM est obtenue à partir des travaux de Jarque et Bera (1987). On considère les distributions de la famille Pearson, caractérisées par l'équation différentielle suivante :

$$\frac{d\,log(f_\beta(u))}{du} = -\frac{u}{b_0 + b_1 u + b_2 u^2}\ ,\ -\infty < u < \infty\,,$$

où $f_\beta(u)$ est la densité d'une variable U et $\beta = [b_0, b_1, b_2]^{'}$ est un vecteur de paramètres. Si l'on choisit $\beta = [1, 0, 0]^{'} \equiv \beta_0$, alors $f_\beta(u)$ correspond à la densité d'une distribution $\mathcal{N}(0, 1)$. Il s'agira donc de tester $\beta = \beta_0$.

Test de spécification dynamique correcte sans hypothèse de stationnarité

Les tests de spécification dynamique correcte évoqués jusqu'ici supposent que le processus étudié est stationnaire. Afin de pouvoir effectuer un test même lorsque cette hypothèse n'est pas vérifiée, González-Rivera et Sun (2017) proposent des tests de spécification correcte (tests joint de spécification dynamique correcte des moments auxquels on s'intéresse et de spécification correcte de la forme fonctionnelle de la prévision de densité) robustes à la présence d'instabilités. Les distributions asymptotiques de leurs différentes statistiques de tests présentent l'avantage d'être libre de paramètres de nuisance et montrent de bonnes performances en termes de taille et de puissance. Ils utilisent pour leurs tests l'approche par AutoContouR (ACR) mise en place par González-Rivera, Senyuz et Yoldas (2011) et généralisée (G-ACR) par González-Rivera et Sun (2015).

Pour une série $\{z_t\}_{t=1}^T$ donnée, on définit les G-ACRs comme étant des carrés d'aires différentes, inclus dans le carré de côté 1 (dans le cas univarié). Plus précisément, sous l'hypothèse nulle de spécification dynamique correcte :

$$H_0 : \{z_t\}_{t=1}^T \ \ i.i.d\ \ \mathcal{U}[0, 1],$$

G-ACR$_{\alpha_i,k}$ est défini comme l'ensemble $B(\cdot)$ de points dans le plan (z_t, z_{t-k}) de telle sorte que le carré de côté $\sqrt{\alpha}_i$ contienne au moins $\alpha_i\%$ des observations, c'est-à-dire :

$$\text{G-ACR}_{\alpha_i,k} = \{B(z_t, z_{t-k}) \subset \mathbb{R}^2\,||\,0 \le z_t \le \sqrt{\alpha}_i \text{ et } 0 \le z_{t-k} \le \sqrt{\alpha}_i,\ \text{s.t} : z_t \times z_{t-k} \le \sqrt{\alpha}_i\}\,.$$

Des exemples graphiques d'ACR pour différentes distributions sont disponibles dans les articles de González-Rivera et alii. (2011) et González-Rivera et Sun (2015). L'utilisation du G-ACR est motivée par le fait, d'une part, qu'il est très sensible à un éloignement de l'uniformité, dans n'importe quelle direction que ce soit, et d'autre part car il est possible de visualiser la forme des G-ACRs et ainsi de savoir d'où vient le rejet de l'hypothèse nulle.

Définissons :

$$I_t^{k,\alpha_i} = \mathbb{I}((z_t, z_{t-k}) \subset\ \text{G-ACR}_{\alpha_i,k}) = \mathbb{I}(0 \le z_t \le \sqrt{\alpha}_i\,,\ 0 \le z_{t-k} \le \sqrt{\alpha}_i)\ \ \text{et}$$

$$\widehat{\alpha}_i = \frac{\sum_{t=k+1}^T I_t^{k,\alpha_i}}{T-k}$$

A partir de cet indicateur, les auteurs proposent un t-test et deux statistiques du χ^2 afin de tester l'hypothèse nulle. La statistique du t-test et sa distribution asymptotique, pour un ACR α_i donné et un retard k donné sont :

$$s \equiv \frac{\sqrt{T-k}(\widehat{\alpha}_i - \alpha_i)}{\sigma_{k,i}} \xrightarrow{d} \mathcal{N}(0, 1)\,,$$

avec $\sigma_{k,i}$, la variance asymptotique de $\widehat{\alpha}_i$.

Les χ^2-statistiques, que nous nommerons C et L, sont construites respectivement à retard k et ACR α_i fixés. La première statistique du χ^2 et sa distribution asymptotique sont :

$$\mathbf{C}_k^{'}\Omega_k^{-1}\mathbf{C}_k \xrightarrow{d} \chi^2_{\mathrm{C}}\,,$$

avec $\mathbf{C}_k = (c_{k,1}, \cdots, c_{k,\mathrm{C}})^{'}$ un vecteur de taille C × 1, $c_{k,i} = \sqrt{T-k}(\widehat{\alpha}_i - \alpha_i)$ et Ω_k la matrice de variance-covariance asymptotique du vecteur $\mathbf{C}_k$.

La seconde statistique du χ^2 et sa distribution asymptotique sont données par :

$$\mathbf{L}_{\alpha_i}^{'} \Lambda_{\alpha_i}^{-1} \mathbf{L}_{\alpha_i} \xrightarrow{d} \chi^2_{\mathrm{K}},$$

avec $\mathbf{L}_{\alpha_i} = (l_{1,\alpha_i}, \cdots, l_{\mathrm{K},\alpha_i})^{'}$ un vecteur de taille K × 1, $l_{k,\alpha_i} = \sqrt{T-k}(\widehat{\alpha}_i - \alpha_i)$ et Λ_k la matrice de variance-covariance asymptotique du vecteur $\mathbf{L}_{\alpha_i}$.

González-Rivera et Sun se concentrent sur les prévisions de densité *out-of-sample*. On se place dans le cas suivant : $\lim P/R = 0$ lorsque $T \to \infty$, $R \to \infty$ et $P \to \infty$, avec T la taille de l'échantillon à disposition, R la partie *in-sample* et P la partie *out-of-sample* ($T = R + P$). L'idée est de former des sous-échantillons de taille r au sein de l'échantillon *out-of-sample* et de calculer les statistiques s, C et L pour chaque sous-échantillon. Ainsi, on obtient trois ensembles de taille $n \equiv T - r - R + 1$ de tests, c'est-à-dire $\{s_j\}_{j=1}^n$, $\{C_j\}_{j=1}^n$ et $\{L_j\}_{j=1}^n$ à partir desquels les auteurs proposent de construire des tests de type Sup (basé sur le supremum) et Avg (basé sur la moyenne) afin de détecter les instabilités.

Synthèse

En résumé, le test de Hong et Li a une distribution asymptotique libre de paramètres de nuisance, mais présente l'inconvénient de devoir définir un paramètre de lissage. Le test de Park et Zhang, quant à lui, se distingue par l'utilisation d'un test de type *Neyman's smooth test*, qui donne plus d'indications sur l'éloignement d'une distribution Uniforme. Il est lui aussi libre de paramètres de nuisance, et présente de meilleurs résultats en termes de taille que le test de Hong et Li. Tout comme Park et Zhang, Lin et Wu utilisent des tests de type *Neyman's smooth test*, mais se différencient dans le choix des fonctions utilisées et par leur approche séquentielle du problème. En effet, ils testent d'abord l'hypothèse d'indépendance, et selon les cas, l'uniformité. En comparaison avec les deux tests précédemment cités, ils obtiennent des résultats similaires à Park et Zhang en termes de tailles, mais sont bien meilleurs en termes de puissance, pour les modèles testés.
Les tests de Kalliovirta présentent les avantages de maintenir l'erreur d'estimation des paramètres sous l'hypothèse nulle (ainsi, la distribution asymptotique des tests ne nécessite pas de correction) et de bénéficier d'une interprétation de type Multiplicateur de Lagrange.
Enfin, les tests de González-Rivera et Sun se démarquent de par leur robustesse quant aux instabilités. Aussi, ils sont libres de paramètres de nuisance et s'accompagnent d'une interprétation graphique par visualisation de la forme des ACR empiriques en cas de rejet de l'hypothèse nulle, ce qui donne des indications sur l'éloignement à une distribution Uniforme.

3.4 Application

La prévision des rendements financiers est un des domaines où l'utilisation des prévisions de densité est particulièrement utile. Ce type de prévision permet de rendre compte de l'incertitude autour de la prévision ponctuelle (espérance de la distribution) qui est souvent peu informative, notamment dans le cas de marchés efficients. Elle permet en outre de rendre compte du risque associé à la détention de l'actif, puisque le calcul de la densité est une alternative au calcul des mesures usuelles (volatilité, VaR, ES, etc.). Dans le cadre de cette application, nous considérerons les rendements logarithmiques quotidiens de 3 indices actions, à savoir le Nasdaq, le Dow Jones et le SP500 sur la période du 01/06/04 au 01/06/12. Les rendements quotidiens des 3 actifs sont représentés sur la figure C.1 en annexe C.

Afin d'obtenir des prévisions de densité, nous avons considéré un processus ARMA(3,3) pour modéliser la dynamique de l'espérance conditionnelle des rendements, et des modèles GARCH(1,1) ou GJR-GARCH(1,1) [8] pour modéliser leur variance conditionnelle. On suppose en outre que la distribution conditionnelle des rendements est Normale. Le modèle GARCH(1,1) est dit symétrique, quand le

8. Dans notre application, ces modèles sont définis par $\epsilon_t = \sigma_t z_t$, avec $z_t \sim \mathcal{N}(0,1)$ et la dynamique de σ_t^2 est donnée par $\sigma_t^2 = a_0 + a_1\epsilon_{t-1}^2 + b_1\sigma_{t-1}^2$ pour le modèle GARCH(1,1) et $\sigma_t^2 = \omega + a_1\epsilon_{t-1}^2 + b_1\sigma_{t-1}^2 + \gamma_1\epsilon_{t-1}^2\mathbf{1}_{\{\epsilon_{t-1}<0\}}$ pour le modèle GJR-GARCH(1,1).

second est asymétrique. Ce choix est motivé par le fait qu'il s'agit de modèles usuels capables de capter l'heteroscédasticité conditionnelle des rendements. Les paramètres ont été estimés sur 90 échantillons [9] de notre jeu de données, sur des périodes qui ont été judicieusement choisies afin de tester la validité des modèles dans des environnements différents. En effet, les 30 premières périodes se situent avant la crise de 2008, les 30 suivantes durant la crise et les dernières nous permettront d'évaluer la validité des modèles dans un contexte post-crise. Chacun de ces échantillons est de taille 250, ce qui correspond approximativement à une période d'un an en terme de jours ouvrés. Une fois le premier échantillon lié à chaque environnement sélectionné, on se décale ensuite d'une journée pour chaque nouvel échantillon. Les dates de début et de fin du premier et du dernier échantillon de chaque environnement sont listés dans le tableau 3.1.

Environnement	Première période	Dernière période
Période pré-crise	28/05/04 - 24/05/05	12/07/04 - 06/07/05
Période de crise	27/08/07 - 21/08/08	08/10/07 - 02/10/08
Période post-crise	11/01/10 - 05/01/11	23/02/10 - 16/02/11

TABLE 3.1 – Dates de la première et de la dernière période de chaque environnement

Pour la suite de cette application, les résultats présentés concerneront uniquement la première période associée à chaque environnement, hormis pour les résultats des différents tests employés. Le tableau 3.2 présente les statistiques descriptives des rendements logarithmiques de chaque période.

Période	Indice	Moyenne	Variance	Skewness	Kurtosis
Période pré-crise :	Nasdaq	1.5 e^{-04}	9.1 e^{-05}	-2.2 e^{-01}	-1.9 e^{-01}
du 28/05/04	Dow Jones	1.2 e^{-04}	4.5 e^{-05}	9.2 e^{-03}	-7.6 e^{-02}
au 24/05/05	SP500	2.5 e^{-04}	4.6 e^{-05}	-5.7 e^{-02}	-2.0 e^{-01}
Période de crise :	Nasdaq	-3.2 e^{-04}	2.1 e^{-04}	9.5 e^{-02}	-8.5 e^{-02}
du 27/08/07	Dow Jones	-6.3 e^{-04}	1.5 e^{-04}	4.5 e^{-02}	2.3 e^{-01}
au 21/08/08	SP500	-5.9 e^{-04}	1.7 e^{-04}	8.6 e^{-02}	3.4 e^{-01}
Période post-crise :	Nasdaq	6.1 e^{-04}	1.6 e^{-04}	-2.8 e^{-01}	1.7
du 11/01/10	Dow Jones	4.0 e^{-04}	1.0 e^{-04}	-1.8 e^{-01}	2.1
au 05/01/11	SP500	4.4 e^{-04}	1.3 e^{-04}	-2.1 e^{-01}	1.9

TABLE 3.2 – Statistiques descriptives des séries des rendements logarithmiques pour chaque période

Les paramètres du processus ARMA(3,3), du modèle GARCH(1,1) et du modèle GJR-GARCH(1,1), ainsi que les écart-types des estimations de ces paramètres sont présentés en annexe C.

Une fois les paramètres estimés, nous avons fait des prévisions à horizons $h = 1$ et $h = 10$ jours de la façon suivante, pour une période donnée : on estime les paramètres du modèle sur ladite période, puis on fait des prévisions à horizon $h = 1$ jour pour les N jours qui suivent la période d'estimation (pour $h = 10$, la période de test est donc décalée de 10 jours par rapport à celle utilisée à horizon $h = 1$ jour), en conservant toujours le même jeu de paramètres. Pour notre étude, nous avons choisi $N = 250$. On a donc au final pour chaque environnement 30 échantillons d'estimation de taille 250, et 30 échantillons de validation de taille 250 eux aussi. A titre d'illustration, les premières prévisions de densité de chaque environnement sont présentées en figure C.2 et C.3 en annexe C, qui ont respectivement été obtenues en modélisant les résidus avec un modèle GARCH(1,1) et un modèle GJR-GARCH(1,1).

9. Ce type d'estimation étant très sensible à la première date choisie (Hansen et Timmermann (2012)) il est préférable d'effectuer les tests sur plusieurs échantillons.

On cherche ensuite à tester la validité des prévisions de densité par deux types de tests : ceux de Rossi et Sekhposyan ainsi que celui de Corradi et Swanson [10] (pour lequel une statistique de test de type Cramér-von Mises a été implémentée). Pour rappel, il s'agit de tests statiques de spécification correcte et l'idée est de vérifier que la série des $\widehat{z}_t$ suit une distribution Uniforme sur l'intervalle [0,1]. Cependant, les tests de Rossi et Sekhposyan se démarquent par le fait qu'ils ne supposent pas que le processus étudié soit stationnaire et qu'ils visent à vérifier que la série des $\widehat{z}_t$ suit une distribution Uniforme en tout point dans le temps durant la période d'évaluation, quand le test de Corradi et Swanson teste l'uniformité de la série en moyenne sur cette période.

Dans le tableau 3.3 sont reportés les pourcentages de rejet de l'hypothèse nulle des différents tests au risque 5% et 1 % pour la prévision à horizon $h = 1$ jour avec modélisation GARCH(1,1) pour les résidus, pour les trois environnements étudiés [11]. K et CvM font référence respectivement aux tests de type Kolmogorov-Smirnov et Cramér-von Mises, et CS et RS respectivement à ceux de Corradi et Swanson et Rossi et Sekhposyan. Les indices 95 et 99 font quant à eux référence au niveau de risque. Les tableaux 3.4, 3.5 et 3.6 reportent respectivement les pourcentages de rejet de l'hypothèse nulle au risque 5% et 1% pour les prévisions à horizon $h = 10$ jours avec modélisation GARCH(1,1) pour les résidus, les prévisions à horizon $h = 1$ jour avec modélisation GJR-GARCH(1,1) pour les résidus et enfin les prévisions à 10 jours avec modélisation GJR-GARCH(1,1). Ils sont construits selon la même logique que le tableau 3.3. Rappelons que l'on rejette l'hypothèse nulle au risque α lorsque $K_\bullet > K_\bullet^{1-\alpha}$, avec $K_\bullet$ la statistique de test obtenue, et $K_\bullet^{1-\alpha}$ la valeur critique associée au niveau de risque considéré. Le même raisonnement s'applique aux statistiques CvM.

Pour les prévisions à horizon $h = 1$ jour avec modélisation GARCH(1,1), on observe que les deux versions du test de Rossi et Sekhposyan conduisent très souvent à ne pas rejeter l'hypothèse nulle de spécification correcte de la densité pour les périodes post-crise, que ce soit au risque 1% ou 5%. En revanche, pour les périodes pré-crise, on rejette assez souvent l'hypothèse nulle, hormis pour l'indice Nasdaq. Enfin, on remarque qu'en période de crise, on est très souvent amené à rejeter l'hypothèse nulle pour les 3 actifs avec parfois un taux de rejet de 90%, même si les résultats des tests au niveau de risque 1% restent mitigés pour l'indice SP500. Cela s'explique bien évidemment par le fait que la crise rend l'estimation de la densité difficile. De plus, les périodes d'estimation associées débutent aux alentours du 18/09/07, soit quelques jours après l'effondrement de Northern Rock, souvent cité comme le début de la crise et les périodes de test démarrent à des dates proches du 15/09/08, soit le jour de la faillite de JP Morgan. Les résultats sont similaires pour le test de Corradi et Swanson, à la différence près que le taux de rejet de l'hypothèse nulle en période de crise est moins élevé. Lorsque l'on modélise les résidus avec un modèle GJR-GARCH(1,1), on retrouve des résultats proches de ceux obtenus avec le modèle GARCH(1,1) pour les deux tests considérés et leurs déclinaisons. Cette modélisation semble cependant plus adaptée en période pré-crise, où seule l'indice Dow Jones reste difficile à prévoir.

Pour les prévisions à horizon $h = 10$ jours avec modélisation GARCH(1,1), les résultats obtenus sont comparables à ceux issus des prévisions à horizon $h = 1$ jour pour tous les tests considérés. On remarquera toutefois que les taux de rejet de l'hypothèse nulle en période de crise pour les tests de Rossi et Sekhposyan sont parfois très élevés (97% et 90% pour les actifs Dow Jones et SP500 respectivement d'après le test de type Cramér-von Mises à 5%) et même de 100% pour l'actif Nasdaq si l'on considère les tests à 5%. Les résultats obtenus avec la modélisation GJR-GARCH(1,1) sont une nouvelle fois proche des résultats précédents. Là encore, la différence se fait principalement en période pré-crise, où seul l'actif Dow Jones semble difficile à modéliser, bien que le taux de rejet de l'hypothèse nulle soit moins élevé avec le modèle GARCH asymétrique. On notera également que le taux de rejet de l'hypothèse nulle en période de crise est en général moins élevé pour les tests de Rossi et Sekhposyan avec cette modélisation, quand on observe le phénomène inverse pour les tests de Corradi et Swanson.

10. Compte tenu de la difficulté de mise en œuvre des tests de spécification dynamique, de la disponibilité des codes et de la taille limitée du chapitre, nous ne présenterons pas d'application de ces tests.

11. En pratique, nous n'avons pas toujours utilisé toutes les prévisions disponibles, certains échantillons fournissant des prévisions pour lesquelles les algorithmes de test ne fonctionnaient pas. Cela ne concerne toutefois que peu d'échantillons.

Période	Indice	K_{CS}^{95}	K_{CS}^{99}	CvM_{CS}^{95}	CvM_{CS}^{99}
Période pré-crise :	Nasdaq	34	34	34	34
	Dow Jones	83	79	83	83
	SP500	67	53	57	53
Période de crise :	Nasdaq	70	70	70	70
	Dow Jones	70	67	67	67
	SP500	60	50	57	50
Période post-crise :	Nasdaq	20	20	20	20
	Dow Jones	37	37	37	37
	SP500	32	32	32	32

Période	Indice	K_{RS}^{95}	K_{RS}^{99}	CvM_{RS}^{95}	CvM_{RS}^{99}
Période pré-crise :	Nasdaq	34	34	34	34
	Dow Jones	83	79	83	83
	SP500	67	53	57	53
Période de crise :	Nasdaq	90	70	90	70
	Dow Jones	83	67	73	67
	SP500	77	53	77	53
Période post-crise :	Nasdaq	20	20	20	20
	Dow Jones	37	37	37	37
	SP500	32	32	32	32

TABLE 3.3 – Pourcentages de rejet de l'hypothèse nulle au risque 5% et 1% pour les prévisions à horizon $h = 1$ jour avec modèle GARCH(1,1) pour les résidus.

Conclusion

Les tests de spécification de prévision de densité se distinguent tout d'abord selon qu'ils testent l'appartenance de la prévision de densité à une famille de distribution donnée (test statiques) ou qu'ils s'attachent également à vérifier l'indépendance de la série des $\widehat{z}_t$ (tests dynamiques).

Parmi les tests statiques, les 2 piliers sont ceux de Bai et Corradi et Swanson. Il s'agit de tests de type Kolmogorov-Smirnov libres de paramètres de nuisance. L'astuce de Bai pour rendre la distribution asymptotique du test libre de paramètres de nuisance est assez simple à mettre place, mais le test souffre de problème de puissance contre des alternatives pour lesquelles il y a violation de l'indépendance. Le test de Corradi et Swanson, en revanche, n'a pas ce problème, mais la méthode de Bootstrap utilisée pour corriger la distribution asymptotique est assez sophistiquée. En marge de ces deux tests, on retrouve ceux de Rossi et Sekhposyan qui eux ne font pas d'hypothèse sur la stationnarité du processus étudié, contrairement aux deux autres précédemment cités. De plus, ils sont plus puissants que le test de Corradi et Swanson (qui en est un cas particulier) et la correction de la distribution asymptotique afin de le rendre libre de paramètres de nuisance est moins coûteuse que la méthode de Bootstrap. Une alternative à ces tests d'uniformité de type Kolmogorov-Smirnov de la série des $\widehat{z}_t$ est l'utilisation d'une double transformation, comme le préconise Knüppel avec l'INT. Toutefois, ce test présente l'inconvénient d'être affecté par des paramètres de nuisance.

Au sein des tests de spécification dynamique correcte, un des premiers tests mis en œuvre est celui de Hong et Li. Ce test est libre de paramètre de nuisance mais nécessite de définir un paramètre de

Période	Indice	K_{CS}^{95}	K_{CS}^{99}	CvM_{CS}^{95}	CvM_{CS}^{99}
Période pré-crise :	Nasdaq	34	34	34	34
	Dow Jones	76	69	76	69
	SP500	57	53	57	53
Période de crise :	Nasdaq	77	73	77	70
	Dow Jones	70	70	73	70
	SP500	60	57	60	60
Période post-crise :	Nasdaq	23	20	20	20
	Dow Jones	41	30	37	33
	SP500	36	32	36	32

Période	Indice	K_{RS}^{95}	K_{RS}^{99}	CvM_{RS}^{95}	CvM_{RS}^{99}
Période pré-crise :	Nasdaq	34	34	34	34
	Dow Jones	72	69	76	69
	SP500	53	53	57	53
Période de crise :	Nasdaq	100	83	100	70
	Dow Jones	83	70	97	70
	SP500	77	57	90	60
Période post-crise :	Nasdaq	23	20	20	20
	Dow Jones	41	33	37	33
	SP500	36	32	36	32

TABLE 3.4 – Pourcentages de rejet de l'hypothèse nulle au risque 5% et 1% pour les prévisions à horizon $h = 10$ jours avec modèle GARCH(1,1) pour les résidus.

lissage. L'idée de ce test est reprise par Park et Zhang et Lin et Wu qui se différencient par l'utilisation de tests de type *Neyman's smooth test* plutôt que Kolmogorov-Smirnov, et par une approche en deux temps du problème pour Lin et Wu quand Park et Zhang testent une hypothèse jointe d'uniformité et d'indépendance des $\widehat{z}_t$. Ces tests sont également libre de paramètres de nuisance. Il en résulte de meilleurs résultats en terme de taille par rapport à Hong et Li et également de puissance pour le test séquentiel. Il existe également dans le cadre de tests de spécification dynamique correcte des alternatives à l'utilisation des tests d'uniformité. Les tests de Kalliovirta utilisent la transformation INT, sont libre de paramètres de nuisance et présentent l'avantage de pouvoir être interprétés comme des tests de type Multiplicateur de Lagrange. Les tests de González-Rivera et Sun quant à eux, sont des tests libre de paramètres de nuisance qui se démarquent par l'utilisation des ACR et par le fait qu'ils ne supposent aucune hypothèse sur la stationnarité du processus étudié. En outre, ils s'accompagnent d'une interprétation graphique, ce qui est appréciable.

Période	Indice	K_{CS}^{95}	K_{CS}^{99}	CvM_{CS}^{95}	CvM_{CS}^{99}
Période pré-crise :	Nasdaq	23	23	30	23
	Dow Jones	70	67	70	70
	SP500	41	37	44	37
Période de crise :	Nasdaq	83	83	83	83
	Dow Jones	76	62	66	66
	SP500	67	56	67	59
Période post-crise :	Nasdaq	23	20	23	23
	Dow Jones	34	31	28	28
	SP500	36	36	36	36

Période	Indice	K_{RS}^{95}	K_{RS}^{99}	CvM_{RS}^{95}	CvM_{RS}^{99}
Période pré-crise :	Nasdaq	23	23	30	23
	Dow Jones	70	67	70	70
	SP500	41	37	44	37
Période de crise :	Nasdaq	83	83	83	83
	Dow Jones	69	62	72	66
	SP500	67	59	67	59
Période post-crise :	Nasdaq	23	20	23	23
	Dow Jones	34	31	28	28
	SP500	36	36	36	36

TABLE 3.5 – Pourcentages de rejet de l'hypothèse nulle au risque 5% et 1% pour les prévisions à horizon $h = 1$ jour avec modèle GJR-GARCH(1,1) pour les résidus.

Période	Indice	K_{CS}^{95}	K_{CS}^{99}	CvM_{CS}^{95}	CvM_{CS}^{99}
Période pré-crise :	Nasdaq	20	20	20	20
	Dow Jones	67	67	67	67
	SP500	41	37	37	37
Période de crise :	Nasdaq	90	83	83	83
	Dow Jones	72	66	69	69
	SP500	74	63	67	63
Période post-crise :	Nasdaq	23	20	23	23
	Dow Jones	31	31	31	28
	SP500	33	30	37	30

Période	Indice	K_{RS}^{95}	K_{RS}^{99}	CvM_{RS}^{95}	CvM_{RS}^{99}
Période pré-crise :	Nasdaq	20	20	23	20
	Dow Jones	67	67	67	67
	SP500	37	37	37	37
Période de crise :	Nasdaq	93	86	90	83
	Dow Jones	72	66	69	69
	SP500	67	63	70	63
Période post-crise :	Nasdaq	23	23	23	23
	Dow Jones	38	31	31	28
	SP500	33	30	37	30

TABLE 3.6 – Pourcentages de rejet de l'hypothèse nulle au risque 5% et 1% pour les prévisions à horizon $h = 10$ jours avec modèle GJR-GARCH(1,1) pour les résidus.

Chapitre 4

A generic method for density forecast recalibration

4.1 Introduction

Due to the increasing need for risk management, forecasting is shifting from point forecasts to density forecasts. Density forecast is an estimate of the conditional probability distribution. Thus, it provides a complete estimate of uncertainty, in contrast to point forecast, which is not concerned whith uncertainty. Two alternative ways to evaluate density forecast exist:

— The first one was proposed by Gneiting et al. (2007, 2014): *Probabilistic forecasting aims to maximize the sharpness of the predictive distributions, subject to calibration, on the basis of the available information set.* Calibration means predictive distributions are consistent with observations, it is more formally defined in the article; sharpness refers to the concentration of the density forecast, and even in the survey paper of Gneiting and Katzfuss (2014), it is not formally defined. An important feature of this framework is that we face a multi-objective problem, which is difficult.

— The second way is the use of a scoring rule, which assesses simultaneously calibration and sharpness. Concerning the well-known *CRPS* scoring rule, Hersbach (2000) showed that it can be decomposed into three parts: reliability (or calibration) part, resolution (or sharpness) part, and uncertainty, which measures the intrinsic difficulty of the forecast. Bröcker (2009) generalized this result to any proper score, that is any score which is minimal if the forecasted probability distribution is the true one (w.r.t the available information). Recently, Wilks (2017) proposed to add an extra miscalibration penalty, in order to enforce calibration in ensemble postprocessing. Nevertheless, even if the score we use mixes calibration and sharpness, the framework is essentially different from the first one.

Besides these two alternative ways of evaluation, probabilistic forecast is mainly used in two different contexts: finance and economics, and weather forecast. In finance and economics, calibration is the unique objective, so a recent survey on "Predictive density evaluation" proposed by Corradi and Swanson (2006) is in fact entirely devoted to the validation of the calibration, without any hint of sharpness. In weather forecast, both ways of evaluation are used. For a quick view on forecasting methods in atmospheric sciences, one can look at the book of Wilks (2006). In the works of Gneiting et al. (2005 and 2007), and in the seminal work of Krzysztofowicz (2004), the goal is to improve sharpness, while preserving calibration. Nevertheless, one can state that there is no formal test of calibration in these works. In the article of Fortin et al. (2006), the only measure used is the *CRPS*, and Gogonel et al. (2013) addresses exclusively the calibration issue.

Here, we are interested in the first method of evaluation: calibration constraint and sharpness objective. Indeed, risk management involves many stakeholders and thus, calibration is a key feature of trust between stakeholders since it impacts all of them. For example, EDF also faces a regulatory constraint: the French technical system operator imposes that the probability of employing exceptional means (e.g.,

load shedding) to meet the demand for electricity must be lower than 1% for each week (RTE, 2004), so EDF has to prove the calibration of its forecasts. Even inside EDF, many different business units may be involved in the management of a given risk, so calibration is compulsory to obtain confidence between risk management stakeholders.

The consequence is that we face a multi-criterion problem, the goal of our contribution is to allow us to enforce the calibration constraint, in a generic way. Furthermore, we show that, even if the evaluation framework is the proper score use, recalibrating leads in many cases to an improvement, and to a very limited loss in other cases.
The remainder of this chapter will be organized as follows. The next section explains the principle of the method. The third part provides some theoretical results while the fourth is devoted to a case study.

4.2 Principle of the method

The Probability Integral Transform (*PIT*, Rosenblatt, 1952) is usually a measure of the calibration of density forecasts. Indeed, if $Y \sim F$ and is continue, the random variable $F(Y) \sim U[0,1]$. Thus, we can find in the literature many tests based on this transformation to evaluate the correct specification of a density forecast. In our case, it is used firstly to recalibrate the forecasts.

Let's look at the following case: let E be the set of all possible states of the world; for each forecasting time j the forecaster knows the current state of the world $e(j)$, and uses it to forecast. For example, in the case of a statistical regression model, E is the set of the possible values of the regressors, in the case of the post-processing of an weather forecasting model, E is the ensemble. The conditional **estimated** distribution is G_e, whereas the **true** one is F_e. So the *PIT* series is:

$$PIT \equiv \left(G_{e(j)}(Y_j)\right)_j .$$

— **A.2.1**: G_e is invertible $\forall\ e \in E$.

If E is discrete, we assume that the frequency of appearance of each state of the world e is p_e. Then, under the assumption **A.2.1**, the c.d.f of the *PIT* is:

$$C(y) \equiv \Pr(\, G(\boldsymbol{Y}) \leq y) \equiv \sum_e p_e\, F_e \circ G_e^{-1}(y).$$

Note that all the results obtained under the hypothesis that E is discrete are still valid in continuous case, even if we only treat the discrete case in this article.

— **A.2.2**: F is invertible.

We propose to use C to recalibrate the forecasts. For each quantile $\tau \in [0,1]$, we use the original model to forecast the quantile τ_C, such that $\Pr(\, G(\boldsymbol{Y}) \leq \tau_c) = \tau$. We remark that this implies $\tau_c = C^{-1}(\tau)$. This correction makes sense since under the assumptions **A.2.1** and **A.2.2**:

$$\begin{aligned} \Pr(\, C \circ G(\boldsymbol{Y}) \leq y) &= \Pr(\, G(\boldsymbol{Y}) \leq C^{-1}(y)) \\ &= C \circ C^{-1}(y) \\ &= y, \end{aligned}$$

which means that the recalibrated forecasts are uniformly distributed on the interval [0,1].

Note that this method is close to the quantile-quantile correction as in Michelangeli et al. (2013) but here, we are concerned by *PIT* recalibration, which allows us to consider the conditional case.

4.3 Impact on global score

If we evaluate our method on the basis of calibration, it ensures this constraint is enforced. But it is important to know if our method is still useful even if one of the probability forecasting users prefers to use scores, for example the Continuous Ranked Probability Score (*CRPS*).

The *CRPS*:

$$CRPS(G, x) = \int_{-\infty}^{+\infty} (G(y) - \mathbf{1}_{\{x \leq y\}})^2 \, dy \,,$$

with G a function and x the observation, is used to evaluate the whole distribution, since it is minimized by the true c.d.f of X.

However, since we have:

$$CRPS(G, x) = 2 \int_0^1 L_\tau(x, G^{-1}(\tau)) \, d\tau \,, \tag{4.1}$$

as shown in the article of Ben Taieb et al. (2016), with L_τ the *Pinball-Loss* function:

$$L_\tau(x, y) = \tau(x - y)\mathbf{1}_{\{x \geq y\}} + (y - x)(1 - \tau)\mathbf{1}_{\{x < y\}} \,,$$

with y the forecast, x the observation and $\tau \in [0, 1]$ a quantile level, and that L_τ is easier to work with, we use this scoring rule to obtain results on *CRPS*.

L_τ is used to evaluate quantile forecasts. Indeed, it is a proper scoring for the quantile of level τ, since its expectation is minimized by the true quantile of the distribution of X.

To begin with, we will prove that under some hypotheses, our correction improves systematically the quality of the forecasts in an infinite sample. Then we will show that under less restrictive hypotheses, our correction deteriorates only slightly—in the worst case—the quality of the forecasts in a more realistic case, e.g finite sample.

4.3.1 Impact on score: conditions for improvement

To assess conditions for improvements, we need to consider:

$$\mathrm{E}_Y[\, L_\tau - L_{\tau_c} \,] \equiv \mathrm{E}_{Y,e}[\, L_\tau(\boldsymbol{Y}, G_e^{-1}(\tau)) \,] - \mathrm{E}_{Y,e}[\, L_\tau(\boldsymbol{Y}, G_e^{-1}(\tau_c)) \,].$$

Here, under the assumption **A.2.1**, $G_e^{-1}(\tau)$ corresponds to the estimated conditional quantile of level $\tau \in [0, 1]$ and $G_e^{-1}(\tau_c)$ to the corrected conditional quantile. Denote: $\eta_e \equiv G_e - F_e$. Considering small errors of specification and regularity conditions on the estimated c.d.f G_e, the true one F_e and their derivatives g_e and f_e:

— **A.3.1.1**: G_e are $C^3 \;\; \forall \; e \in E$.

— **A.3.1.2**: F_e are C^3 and invertible $\;\forall \; e \in E$.

— **A.3.1.3**: η_e, f_e and their derivatives are bounded $\forall \; e \in E$ by a constant which doesn't depend on e.

— **A.3.1.4**: $\forall \; \tau \in [0, 1]$, $\forall \; e \in E$, η_e, its first, second and third derivatives are finite in $F_e^{-1}(\tau)$,

and using functional derivatives, directional derivatives and the implicit function theorem (proof in Appendix D.1.1) we can rewrite (adding the assumption **A.2.1**):

$$\begin{aligned} \mathrm{E}_Y[\, L_\tau - L_{\tau_c}\,] \quad \sim \quad & \left(\sum_e \frac{p_e \eta_e(F_e^{-1}(\tau))}{f_e(F_e^{-1}(\tau))}\right)\left(\sum_e p_e \eta_e(F_e^{-1}(\tau))\right) \\ & - \left(\sum_e \frac{p_e}{2 f_e(F_e^{-1}(\tau))}\right)\left(\sum_e p_e \eta_e(F_e^{-1}(\tau))\right)^2 \\ & \text{as } max\ \eta_e \to 0, \end{aligned} \tag{4.2}$$

with p_e the frequency of appearance of the state e.

This result allows us to find conditions for improvement of the expectation of the *Pinball-Loss* score, with additional following conditions:

— **A.3.1.5**: η or f^{-1} is a constant, or $max_e(\bullet)/min_e(\bullet) < 3 + 2\sqrt{2}$ for both η_e and f_e^{-1}, $\forall\ e \in E$.

— **A.3.1.6**: the correlation between η and f^{-1}, $\sigma_{f^{-1}}$ or σ_η is null. Here the correlation is used as a descriptive statistics notation, even if the series η and f^{-1} are deterministic. The null correlation means that the difference between the true probability distribution function and the model have the same magnitude in low and in high density regions.

Under the assumption **A.3.1.5** or **A.3.1.6**, if $\exists\ \nu \geq 0$ (sufficiently small) $\forall\ e \in E\ \ \forall\ y \in \mathbf{R}; |\eta_e(y)| \leq \nu$, we show that (proof in Appendix D.1.2):

$$0 \quad \leq \quad \mathrm{E}_Y[\, L_\tau - L_{\tau_c}\,] \text{ and} \tag{4.3}$$

$$0 \quad \leq \quad \mathrm{E}_Y[\, CRPS_{G,C\circ G}\,], \tag{4.4}$$

with $\mathrm{E}_Y[\, CRPS_{G,C\circ G}\,] \equiv \mathrm{E}_Y[\, CRPS(G, \mathbf{Y}) - CRPS(C \circ G, \mathbf{Y})\,]$.

In other words, with those restrictions, our recalibration systematically improves the quality of the forecasts. Indeed, remember that the expectation of the *Pinball-Loss* score is minimized by the true quantile of the distribution of $\mathbf{Y}$ and negatively oriented. Thus, the lower the expectation of the *Pinball-Loss* score, the better.

4.3.2 Impact on score: bounds on degradation

In reality, we cannot obtain the corrected probability level $\tau_c \in [0, 1]$, and we need to estimate it. If we want to upper bound the degradation, we can study the more realistic case of

$$\mathrm{E}_Y[L_{\widehat{\tau}_c} - L_\tau\,] \equiv \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n} L_\tau(y_j, G_j^{-1}(\widehat{\tau}_c)) - L_\tau(y_j, G_j^{-1}(\tau))\right], \tag{4.5}$$

with τ, $\widehat{\tau}_c \in [0, 1]$.

In our case study, $\widehat{\tau}_c$ is obtained empirically, on the basis of the available *PIT* values. Thus, we have a consistant estimator of τ_c and one can rewrite (4.5) such as $\mathrm{E}_Y[\, L_\tau(Y, G^{-1}(Q_\tau))\,] - \mathrm{E}_Y[\, L_\tau(Y, G^{-1}(\tau))\,]$, with Q_τ a random variable converging in distribution to a Normal distribution with mean τ_c and a variance decreasing at the rate $\frac{1}{n}$.

In such a case, it is still possible to obtain bounds concerning the error induced by our correction.

— **A.3.2.1**: F_e and G_e are $C^2 \ \ \forall\, e \in E$.

— **A.3.2.2**: $\forall\, y \in \mathbf{R}$, $\forall\, e \in E$, $|F_e(y) - G_e(y)| \leq \varepsilon$, with $\varepsilon \in [0,1]$.

— **A.3.2.3**: the derivatives of G_e are lower bounded $\forall\, e \in E$, $\forall\, \tau \in [0,1]$ by $1/\xi$, on the intervals $[G_e^{-1}(0 \vee (\tau - \varepsilon)), G_e^{-1}(1 \wedge (\tau + \varepsilon))]$, with $\xi \in]0, +\infty[$.

— **A.3.2.4**: $\forall\, e \in E$, $\forall\, \tau \in [0,1]$, $f_e(G_e^{-1}(\tau_c)) \ \leq \ \beta$, with $\beta \in]0, +\infty[$ and f_e the derivatives of F_e.

— **A.3.2.5**: f_e are continuous over the interval $[-\infty, G_e^{-1}(\tau_c)]$ $\forall\, e \in E$ and their derivatives are bounded, i.e $\forall y \in \mathbf{R}$, $\forall\, e \in E$, $|f_e^{'}(y)| \leq M$, with $M \in]0, +\infty[$ and f_e the derivative of F_e.

— **A.3.2.6**: the derivatives of g_e are bounded, i.e $\forall y \in \mathbf{R}$, $\forall\, e \in E$, $|g_e^{'}(y)| \leq \alpha$, with $\alpha \in]0, +\infty[$ and g_e the derivative of G_e.

Under the assumptions **A.2.1**, **A.3.2.1**, **A.3.2.2**, **A.3.2.3**, **A.3.2.4**, **A.3.2.5** and **A.3.2.6**, we prove (proof in Appendix D.2):

$$|\, \mathrm{E}_Y[L_{\widehat{\tau}_c} - L_\tau\,]\,| \ \ \leq \ \ 2\,\varepsilon^2\,\xi + \frac{C\,\lambda}{n} \ \text{ and} \tag{4.6}$$

$$|\, \mathrm{E}_Y[\,CRPS_{G,C\circ G}\,]| \ \ \leq \ \ 2\left(2\,\varepsilon^2\,\xi + \frac{C\,\lambda}{n}\right) \tag{4.7}$$

with $C = \frac{(1-\tau)\alpha\xi^3}{2} + C_{int} + C_{abs}$, $C_{int} = \frac{\xi^2\beta}{2}\left[1 + \alpha\xi^2 + \frac{\alpha^2\xi^4}{4}\right]$,

$C_{abs} = \frac{M\,\xi^3}{6}\left[1 + \frac{3\xi^3\alpha}{2} + \frac{3\xi^3\alpha^2}{4} + \frac{\xi^3\alpha^3}{8}\right]$ and $\frac{\lambda}{n}$, the variance of Q_τ

This inequality shows that our recalibration deteriorates only slightly the quality of the forecasts in the worst case. Obviously, it also shows that our method improves only slightly the quality, but remember that our goal is to enforce the validity constraint, which is achieved.

4.4 Case study

We use our method on ensemble forecasts data set from the European Centre for Medium-Range Weather Forecasts (ECMWF). One can see in the work of Gneiting (2014) that the statistical post-processing of the medium range ECMWF ensemble forecast has been addressed many times. The extended range (32 days instead of 10 days) has been addressed in some studies, but with the same methods and tools. We will show here that our recalibration method, despite its genericness, is competitive with a standard post-processing method. We dispose of temperature forecasts in a 3-dimensional array. The first one represents the date of forecasts delivery. The forecasts were made every Monday and Thursday from 11/02/13 to 02/02/17. Since 3 observations are missing, we have 413 dates of forecasts delivery. The second dimension is the number of the scenario in the ensemble member, and we have 51 scenarios. The third dimension is the forecast horizon. Since we have 32 days sampled with a forecast every 3 hours, it produces 256 horizons.

We study the calibration and compare the *CRPS* expectation using directly the ensemble forecast, the so-called Ensemble Model Output Statistics (*EMOS*) method and our recalibration method with a Cauchy Kernel dressing for the ensembles. We choose a Cauchy Kernel in order to address problems with the bounds of the ensembles. Indeed, a lot of observations were out of the bounds of the ensemble, which produces a lot of *PIT* with value 0 or 1. Thus, to avoid this problem, we need to use a Kernel with heavy tail.

During the last 12 years, the ECMWF has changed its models 27 times, which means a change every 162 days on average. Thus, it is important to use a train sample significally smaller than 162 days. However, it is also important to dispose of enough observations to obtain a consistant estimator of τ_c.

Our method obtains good results with 30 days used for the recalibration but the algorithm to minimize in order to find the parameters of the *EMOS* in the *R* package *EnsembleMOS* doesn't converge if we use less than 60 days (at least with our data set). Thus, we chose to use 60 days for the recalibration.

To recalibrate the forecasts for a particular forecasting day and a particular horizon (remember that we have 256 horizons), we use the forecasts made for the same horizon, over the 60 previous dates of forecast delivery for the two methods. However, with our method, we use a linear interpolation based on the *PIT* series formed by these 60 previous days to recalibrate the forecasts. The linear interpolation is also used to calculate the different quantile levels when we are not working with *EMOS* (in that case, for the recalibration or to calculate the quantile, we use the Normale distribution with the fitted parameters). Note that the hypotheses concerning only G_e are verified $\forall\ e \in E$. Besides, even if we cannot verifiy the other hypotheses, we show expected results

Let's start with the calibration property. We have calculated the *PIT* series for each horizon (256), and use 5% Kolmogorov-Smirnov test for each of them.

	Raw ensemble	*EMOS*	Our method
Succes rate in %	14	0.39	96

Table 4.1 – Success rate to 5% *K-S* test

In table 4.1, the success rate is the percentage of horizons passing the test. As expected, we can see that our method allows us the test of validity to be passed while the use of the raw ensemble fails. The *EMOS* also failed to pass the test. Clearly, our method is useful to ensure the calibration property. But how about the quality of the density forecast? In order to evaluate the impact of our correction on the forecast quality, we are interested in the *CRPS* expectation.

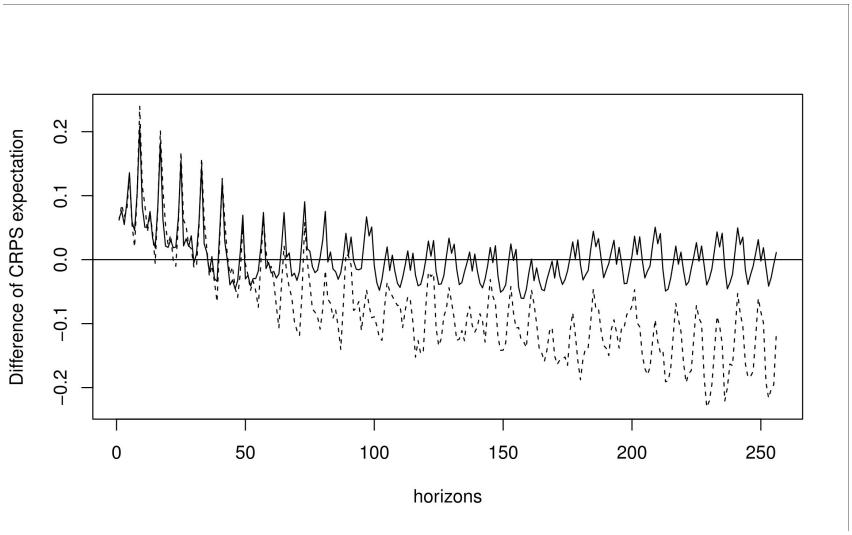

Figure 4.1 – Comparison of *EMOS* and our method *CRPS* expectation with that of raw ensemble The plain line corresponds to our method and the dashed one to the *EMOS*.

We can see in figure 4.1 that *EMOS* as well as our method are more efficient than the raw ensemble for little horizons. However, the *EMOS* deteriorates clearly the quality of the forecasts when the horizon grows, contrarily to our method which deteriorates only slightly the quality of the forecasts, when it is the case. Thus, this study highlights perfectly the usefulness of our method, which is very simple to use. Indeed, it shows that it allows us to ensure the validity constraint, with a limited negative impact on the quality.

Chapitre 5

Conclusion et perspectives

Dans le premier chapitre, nous avons étudié l'apport des méthodes de machine learning dans le cadre des prévisions de quantiles. L'objectif de cette partie est double. D'une part, nous avons comparé les différentes méthodes de machine learning afin de faire un benchmarking. D'autre part, nous avons comparé ces algorithmes de machine learning avec la régression quantile, qui est la première méthode proposée pour faire de la prévision de quantiles.

Pour cela, cinq algorithmes issues de cinq grandes familles de modèles de machine learning ont été implémentés. Il s'agit des méthodes de régression quantile par machine à vecteur de support, par réseaux de neurones et par forêts aléatoires, de la méthode de gradient boosting et de la méthode Extended Log-F (ELF) issue des modèles de type modèles additifs. Afin de comparer ces méthodes, nous avons utilisé trois jeux de données aux caractéristiques différentes. Il en ressort que la prévision de quantiles sur un jeu de données avec un seul régresseur et une variable à prédire avec beaucoup de variance est difficile, quelle que soit la méthode utilisée. Sur un jeu de données de série temporelle, deux algorithmes se démarquent, à savoir la méthode ELF et le gradient boosting. Les autres méthodes de machine learning donnent des résultats mitigés, quand la régression quantile semble ne pas être adaptée. Enfin, nous avons constaté que sur un jeu de données avec beaucoup de régresseurs, la régression quantile par forêts aléatoires semble être la plus adaptée. une nouvelle fois, les résultats pour les autres méthodes varient selon le quantile étudié, et l'on remarque que la régression quantile n'obtient pas systématiquement les moins bon résultats.

Cependant, ces conclusions ne sont que partielles, et plusieurs perspectives d'étude s'offrent donc à la suite de ce chapitre. En effet, nous n'avons étudié que six quantiles pour chaque jeu de données. Il serait intéressant de prédire tous les quantiles de 1% à 99% afin de pouvoir comparer la validité des prévisions issues de chaque méthode. De plus, un plus grand nombre de jeux de données permettrait de confirmer que certains algorithmes sont plus adaptés que d'autres selon le contexte dans lequel on évolue.

Dans le second chapitre, nous proposons une synthèse détaillée des différents tests de validation de prévisions de densité que l'on peut trouver dans la littérature économétrique. Les premiers tests n'étant plus applicables dans ce contexte de par leur construction, il a fallu repenser les tests d'évaluation. Ces tests sont en général des tests d'uniformité de type *Kolmogorv-Smirnov* basés sur la variable *PIT*. On trouve également des tests lissés de Neyman, ou encore des tests de normalité basés sur une transformation de la variable *PIT*.

Par ailleurs, il existe deux classes de tests : les tests de spécification correcte, qui reposent sur une évaluation absolue des prévisions et les tests basés sur une évaluation relative. Les tests de spécification correcte ont pour but de vérifier sous l'hypothèse nulle la validité du modèle. Les tests d'évaluation relative servent quant à eux à comparer différents modèles potentiellement tous mal spécifiés à un modèle de référence, éventuellement mal spécifié lui aussi. Au sein de ces deux classes, il faut distinguer deux types de tests. Les tests de spécification statiques ont pour vocation de vérifier que la forme paramétrique de la densité spécifiée appartient à la bonne famille de distributions. Les tests de spécification dynamiques s'attachent en plus de cela à vérifier que les variables *PIT* sont indépendantes, et ne concernent donc que des prévisions à horizon $t = 1$.

Dans ce deuxième chapitre, nous avons également appliqué deux des tests présentés. Il s'agit des tests de Corradi et Swanson (2006) et de Rossi et Sekhposyan (2013). Ces tests ont été appliqués à des prévisions de log-rendements de trois indices boursiers à horizons $h = 1$ jour et $h = 10$ jours, issues de modèles simples. Nous avons modélisé un processus ARMA(3,3) et des modèles GARCH(1,1) et GJR-GARCH(1,1) pour les résidus. Les prévisions ont ensuite été évaluées dans des contextes différents : la première période étudiée se situe avant la crise de 2008, la seconde durant la crise et la dernière après la crise. Pour les prévisions à horizons $h = 1$ jour, nous avons mis en évidence de bons résultats sur la période post-crise, quelle que soit la méthode utilisée. De plus, nous avons constaté que la modélisation GJR-GARCH(1,1) semblait plus adaptée. Pour les prévisions à horizon $h = 10$ jours, la modélisation GJR-GARCH(1,1) semble une nouvelle fois plus appropriée. En effet, les résultats obtenus avec le modèle GARCH(1,1) sont proches de ceux obtenus à horizon $h = 1$ jour. En revanche, avec le modèle GJR-GARCH(1,1), on obtient de bons résultats en période post-crise mais aussi pour deux des trois actifs en période pré-crise. Enfin, nous avons également remarqué que le taux de rejet de l'hypothèse nulle est moins élevé en période de crise avec avec ce modèle lorsque l'on utilise le test de Rossi et Sekhposyan (2013).

Dans le dernier chapitre de la thèse, nous proposons une méthode de recalibration de prévisions issues de modèles mal spécifiés. Nous avons montré que sous certaines hypothèses, cette méthode permet d'assurer la validité des prévisions tout en ne dégradant pas ou peu leur qualité, lorsque celle-ci n'est pas améliorée. Par la même occasion, cela permet de simplifier le choix entre plusieurs prévisions de densité. En effet, puisque la validité est assurée, seul le critère de qualité est nécessaire pour départager différentes prévisions.

Afin de mettre en évidence l'intérêt de notre méthode, nous l'avons appliquée sur deux jeux de données : un jeu de données composé de scénarios de températures et un autre composé de scénarios de prix. Sur le jeu de données météo, nous avons confronté notre méthode à la méthode *ensemble members output statistics* (*EMOS*) et nous avons mis en évidence l'impact positif de la recalibration sur la validité des prévisions. Nous avons également montré que la qualité des prévisions est améliorée dans la plupart des cas, et dégradée dans d'autres, mais seulement légèrement. La deuxième étude nous a permis de montrer l'apport potentiel de notre méthode pour EdF, dans la mesure où les scénarios utilisés sont générés par un outil de gestion de l'entreprise. En effet, nous avons observé, entre autres, que les cash-flows calculés à partir des scénarios recalibrés sont plus proches des vraies valeurs que celles calculées avec les scénarios bruts sur les périodes hors hivers. Cela représente donc un gain potentiel pour l'entreprise.

Plusieurs perspectives apparaissent suite à ce chapitre. Il serait intéressant de disposer de plus de données de scénarios de prix issues de EdF afin de pouvoir tester la validité des prévisions et vérifier que les conclusions quand à l'amélioration des valeurs de cash-flows persistent. De plus, travailler sur des rendements plutôt que des prix pourrait éventuellement donner de bons résultats. Enfin, nous pourrions étudier l'impact de notre méthode sur des prévisions issues de méthodes de machine learning.

Appendix A

Recalibration on electricity price ensemble forecasts

Here, we use our method to recalibrate electricity price ensemble forecasts. Our data are outputs of a management tool developed by EdF, called *OPUS*. This tool is intended to help manage balance between demand and generation for horizons between a month and some years, in French electric system. *OPUS* can use two models to produce the ensemble; the first one (called OPUS in the following) considers the foreign countries all together and the second one (OPUS-PAF) considers them individually. Thus, we have two datasets.

The price forecasts we will address are in a 3-dimensional array:

— First dimension is the date of forecasts delivery, we have 7 dates.

— Second dimension is member in the ensemble, we have 484 members.

— Third dimension is the forecast horizon, we have 385 days, each one sampled 6 times, which makes 2310 horizons.

The table A.1 shows the forecasting dates for each forecasts delivery.

Day of forecasts delivery	First forecasting day	Last forecasting day
1	05/09/16	24/09/17
2	24/10/16	12/11/17
3	09/01/17	28/01/18
4	13/03/17	01/04/18
5	15/05/17	03/06/18
6	10/07/17	29/07/18
7	09/10/17	28/10/18

Table A.1 – Forecasting dates for each output

Our goal is to use six first forecasts to recalibrate the seventh one. Our quality indicators will be the cash-flows for three commodities (thermal, hydraulic and nuclear) and the *Pinball-Loss*.

Since we only have 6 forecasts, we need to add some expertise, included in pre and post-processing, to improve the results of our non-parametric recalibration method. First, we transform the data, computing the logarithm of the prices, and we cap the prices, using a threshold δ defined by our experts. Since we can not manage a distribution with masses on the bounds, the cap is a soft cap (we replace $\min(p, T)$ by $\min(p, T) + \epsilon|p - T|$, where p is the price). However, since we have only 6 values by *PIT* series, we use linear interpolation to find τ_c.

The table A.2 highlights the need for recalibration, showing the percentages of raw *PIT* series (composed by the *PIT* values at each horizon issued from the 6 first dates of forecasts delivery) for OPUS and OPUS-PAF passing the Kolmogorov-Smirnov test at level 5%.

Raw OPUS	Raw OPUS-PAF
43	12

Table A.2 – Percentages of raw *PIT* series passing the Kolmogorov-Smirnov test at level 5%

Unfortunately, since we do not have enough data, we cannot compare the results with that obtained after recalibration.

The table A.3 shows the percentage of observations out the ensemble bounds issued from the 7^{th} date of forecast delivery for OPUS, OPUS-PAF and their recalibrated version.

Raw OPUS	Recalibrated OPUS	Raw OPUS-PAF	Recalibrated OPUS-PAF
12	7	28	21

Table A.3 – Percentages of obsrvations out of the ensemble bounds

We can see that after recalibration, the percentage of observations out of the bounds of the ensembles decrease.

The table A.4 shows the *Pinball-Loss* scores obtained with the raw scenarios and the recalibrated ones for quantile of level 1%, 4%, 50%, 96% and 99%. We chose those quantile levels because they are very useful to manage risk. Thus, it seems interesting to study those quantiles in price forecasting context.

Quantile level	Raw OPUS	Recalibrated OPUS	Raw OPUS-PAF	Recalibrated OPUS-PAF
1%	0.41	0.40	0.42	0.41
4%	1.32	1.31	1.34	1.31
50%	9.30	7.85	10.57	8.17
96%	5.24	3.20	9.93	5.02
99%	3.39	1.53	7.33	3.00

Table A.4 – *Pinball-Loss* score

For the low quantile levels, the results are similar with all the scenarios. However, we can see that the recalibration improves the results for the other quantile, since the *Pinball-Loss* score is negatively oriented.

We can see in figure A.1 that the cash-flows obtained with the recalibrated forecasts are often closer to the real cash-flows than the cash-flows calculated with the raw data.

Indeed, we can see an improvement for 68% of the day considered. We remark also that the recalibration is less efficient during winter, with an improvement for 18% of the day considered for the recalibrated OPUS data and 13% for the recalibrated OPUS-PAF data. The reason is that winters are difficult to forecast and that our method is very sensitive to the raw estimation. Indeed, the hypothesis $|G_e - F_e| < \epsilon$ with ϵ small is probably not verified during winter. In other words, the raw estimation is certainly too far

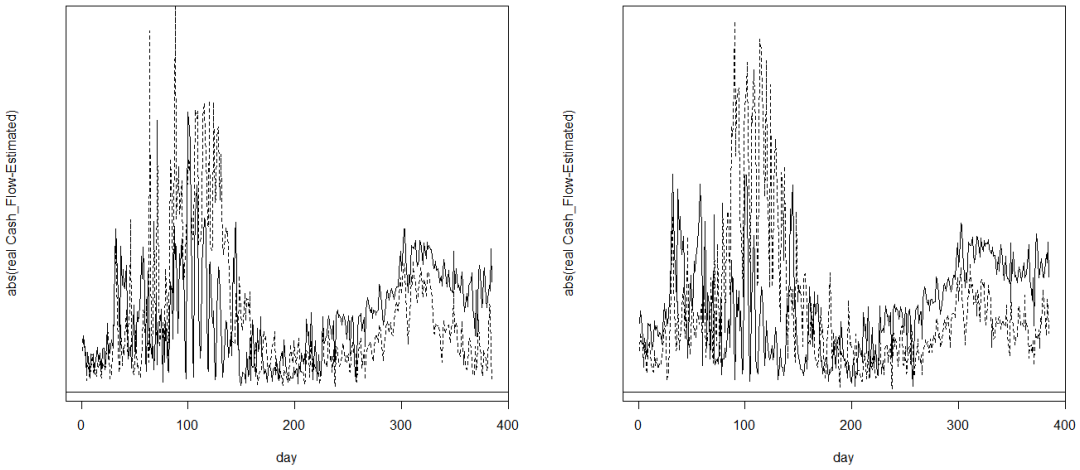

Figure A.1 – Comparison of raw OPUS ensembles and recalibrated ones cash-flows (left) and of raw OPUS-PAF ensembles and recalibrated ones cash-flows (right) with the real cash-flows, in absolute value. The plain line corresponds to the raw ensembles and the dashed one to our method.

to the true distribution, which explain the bad results for this period. However, acording to the results for the remainder of the period studied, we can see a potential gain for EdF using the recalibration.

Appendix B

Appendix chapter 2

The table B.1 shows the seed setting to select the hyper-parameters to create the grids search and the seed used before run the algorithms in order to obtain reproducible results. Note that to create the folds for validation procedure, we use a seed equal to 1 for "BostonHousing" and 125 for "mcycle".

Method	Dataset	Algorithm	Grid search
QRNN	mcycle	923	/
	BostonHousing	923	/
	GefCom2014	923	/
QRF	mcycle	12	/
	BostonHousing	12	/
	GefCom2014	587	1548
GBM	mcycle	45	/
	BostonHousing	45	/
	GefCom2014	45	17
ELF	mcycle	41241	/
	BostonHousing	41241	/
	GefCom2014	41241	/

Table B.1 – Seed setting.

Tables B.2, B.3 and B.4 show the grids search selected for "mcycle", "BostonHousing" and GEFCom2014 dataset respectively. Note that the notation s(c) a : b means that we have randomly selected c entire values on the interval [a,b] and that the notation a : b alone means that we select all the entire values in the interval [a,b]. Moreover, $1e^{-a:b}$ signify that we select the negative of all the entire values in the interval [a,b] as exponent.

Method	Parameters	Grid search
SVQR	C_τ	0.1, 1, 10, 15, 20, 100, 500, 750, 1000
	σ	$1e^{-(1:5)}$, 0.25, 0.5, 1
QRNN	nb neurons	1, 2
	λ_1	$1e^{-(1:4)}$
QRF	depth	2, 3, 4, 5
	nb obs leaf	10 : 18
	nb split var	1
	nb trees	100, 500, 1000
GBM	depth	2, 3, 4, 5
	nb obs leaf	10 : 18
	shrinkage	$1e^{-(1:4)}$
	nb trees	100, 500, 1000

Table B.2 – Grids search for mcycle.

Method	Parameters	Grid search
SVQR	C_τ	0.1, 1, 10, 15, 20, 100, 500, 750, 1000
	σ	$1e^{-(1:5)}$, 0.25, 0.5, 1
QRNN	nb neurons	1 : 5
	λ_1	$1e^{-(1:4)}$
QRF	depth	4 : 14
	nb obs leaf	12 : 25
	nb split var	1 : 13
	nb trees	100, 500, 1000
GBM	depth	4 : 14
	nb obs leaf	12 : 25
	shrinkage	$1e^{-(1:4)}$
	nb trees	100, 500, 1000

Table B.3 – Grids search for BostonHousing.

Method	Parameters	Grid search
SVQR	C_τ	0.1, 1, 10, 15, 20, 100, 500, 750, 1000
	σ	$1e^{-(1:5)}$, 0.25, 0.5, 1
QRNN	nb neurons	1 : 5
	λ_1	$1e^{-(1:4)}$
QRF	depth	s(8) 5 : 20
	nb obs leaf	s(10) 25 : 45
	nb split var	1 : 5
	nb trees	100, 500, 1000
GBM	depth	s(8) 5 : 20
	nb obs leaf	s(10) 25 : 45
	shrinkage	$1e^{-(1:4)}$
	nb trees	100, 500, 1000

Table B.4 – Grids search for GEFCom2014.

Annexe C

Annexe chapitre 3

Les tableaux C.1, C.2, C.3 et C.4 présentent respectivement les paramètres du processus ARMA(3,3) couplé au modèle GARCH(1,1), les paramètres du modèle GARCH(1,1), les écart-types des estimations des paramètres du processus ARMA(3,3) couplé au modèle GARCH(1,1) et ceux des estimations des paramètres du modèle GARCH(1,1). Les paramètres du processus ARMA(3,3) couplé au modèle GJR-GARCH(1,1), ceux du modèle GJR-GARCH(1,1) et les écart-types de leurs estimations sont présentés dans les tableaux C.5, C.6, C.7 et C.8.

Période	Indice	ar1	ar2	ar3	ma1	ma2	ma3
Période pré-crise :	Nasdaq	0.54	-0.76	-0.26	-0.59	0.85	0.18
du 28/05/04	Dow Jones	0.05	-0.07	0.94	-0.07	0.08	-1.00
au 24/05/05	SP500	-0.51	0.45	0.95	0.52	-0.49	-0.98
Période de crise :	Nasdaq	-0.78	-0.82	0.20	0.69	0.76	-0.33
du 27/08/07	Dow Jones	-1.11	0.39	0.72	1.11	-0.51	-0.85
au 21/08/08	SP500	0.51	-0.88	-0.14	-0.69	1.10	-0.04
Période post-crise :	Nasdaq	0.33	-0.30	0.93	-0.33	0.30	-1.00
du 11/01/10	Dow Jones	-1.08	-0.70	0.12	1.05	0.67	-0.18
au 05/01/11	SP500	-0.07	-0.97	-0.06	0.03	0.97	0.05

TABLE C.1 – Paramètres du processus ARMA(3,3) avec modèle GARCH(1,1) pour les résidus

Période	Indice	a0	a1	b1
Période pré-crise :	Nasdaq	$1.0\ e^{-05}$	$5.5\ e^{-02}$	$8.3\ e^{-01}$
du 28/05/04	Dow Jones	$5.9\ e^{-08}$	$2.6\ e^{-06}$	1.0
au 24/05/05	SP500	$5.7\ e^{-08}$	$6.6\ e^{-08}$	1.0
Période de crise :	Nasdaq	$1.2\ e^{-05}$	$4.7\ e^{-02}$	$8.9\ e^{-01}$
du 27/08/07	Dow Jones	$1.9\ e^{-07}$	$1.1\ e^{-04}$	1.0
au 21/08/08	SP500	$1.6\ e^{-07}$	$3.0\ e^{-14}$	1.0
Période post-crise :	Nasdaq	$2.3\ e^{-06}$	$1.1\ e^{-01}$	$8.7\ e^{-01}$
du 11/01/10	Dow Jones	$1.9\ e^{-06}$	$1.1\ e^{-01}$	$8.7\ e^{-01}$
au 05/01/11	SP500	$2.5\ e^{-06}$	$1.0\ e^{-01}$	$8.8\ e^{-01}$

TABLE C.2 – Paramètres du modèle GARCH(1,1)

Période	Indice	ar1	ar2	ar3	ma1	ma2	ma3
Période pré-crise :	Nasdaq	$6.4\ e^{-02}$	$5.2\ e^{-02}$	$6.1\ e^{-02}$	$6.6\ e^{-03}$	$7.8\ e^{-03}$	$1.2\ e^{-02}$
du 28/05/04	Dow Jones	$4.8\ e^{-02}$	$1.2\ e^{-01}$	$8.3\ e^{-02}$	$8.3\ e^{-02}$	$6.0\ e^{-02}$	$9.9\ e^{-02}$
au 24/05/05	SP500	$7.7\ e^{-02}$	$1.1\ e^{-01}$	$5.4\ e^{-02}$	$5.1\ e^{-02}$	$9.1\ e^{-02}$	$7.1\ e^{-02}$
Période de crise :	Nasdaq	$1.5\ e^{-03}$	$1.5\ e^{-03}$	$1.1\ e^{-03}$	$2.9\ e^{-03}$	$1.8\ e^{-03}$	$1.0\ e^{-03}$
du 27/08/07	Dow Jones	$8.9\ e^{-04}$	$3.9\ e^{-03}$	$4.8\ e^{-03}$	$8.7\ e^{-04}$	$3.6\ e^{-04}$	$7.4\ e^{-04}$
au 21/08/08	SP500	$1.3\ e^{-03}$	$2.4\ e^{-03}$	$1.3\ e^{-04}$	$1.6\ e^{-03}$	$3.3\ e^{-03}$	$7.6\ e^{-05}$
Période post-crise :	Nasdaq	$2.6\ e^{-02}$	$3.1\ e^{-02}$	$3.4\ e^{-0}$	$3.2\ e^{-03}$	$3.8\ e^{-03}$	$5.1\ e^{-03}$
du 11/01/10	Dow Jones	$5.4\ e^{-01}$	$6.2\ e^{-01}$	$5.0\ e^{-01}$	$5.4\ e^{-01}$	$6.3\ e^{-01}$	$4.6\ e^{-01}$
au 05/01/11	SP500	$0.15\ e^{01}$	$5.3\ e^{-02}$	$0.15\ e^{01}$	$0.15\ e^{01}$	$2.7\ e^{-02}$	$0.14\ e^{01}$

TABLE C.3 – Ecart-types des estimations des paramètres du processus ARMA(3,3) avec modèle GARCH(1,1) pour les résidus

Période	Indice	a0	a1	b1
Période pré-crise :	Nasdaq	$1.4\ e^{-07}$	$6.2\ e^{-03}$	$2.0\ e^{-02}$
du 28/05/04	Dow Jones	$1.1\ e^{-05}$	$8.0\ e^{-03}$	$6.2\ e^{-03}$
au 24/05/05	SP500	$9.8\ e^{-06}$	$1.3\ e^{-02}$	$9.2\ e^{-03}$
Période de crise :	Nasdaq	$3.0\ e^{-06}$	$2.9\ e^{-02}$	$6.2\ e^{-02}$
du 27/08/07	Dow Jones	$1.1\ e^{-05}$	$1.1\ e^{-02}$	$6.2\ e^{-03}$
au 21/08/08	SP500	$1.1\ e^{-05}$	$3.0\ e^{-03}$	$2.2\ e^{-03}$
Période post-crise :	Nasdaq	$1.0\ e^{-05}$	$6.4\ e^{-02}$	$7.0\ e^{-02}$
du 11/01/10	Dow Jones	$1.3\ e^{-05}$	$8.0\ e^{-02}$	$9.9\ e^{-02}$
au 05/01/11	SP500	$1.2\ e^{-05}$	$5.8\ e^{-02}$	$7.7\ e^{-02}$

TABLE C.4 – Ecarts-types des estimations des paramètres du modèle GARCH(1,1)

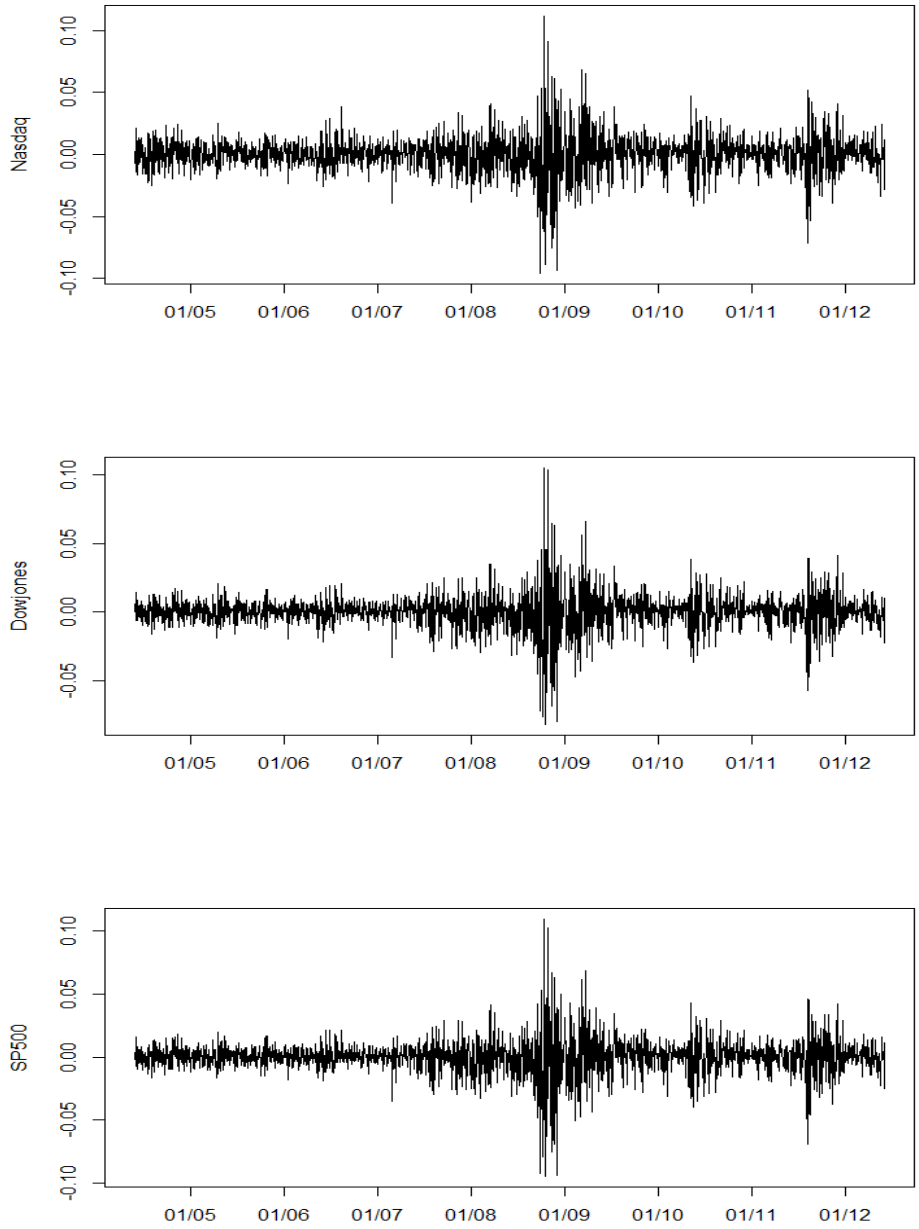

FIGURE C.1 – Rendements logarithmiques des prix à la clôture des trois indices sur la période 01/06/04-01/06/12

Période	Indice	ar1	ar2	ar3	ma1	ma2	ma3
Période pré-crise :	Nasdaq	0.31	-0.82	-0.26	-0.36	0.89	0.20
du 28/05/04	Dow Jones	0.10	-0.02	0.98	-0.10	0.05	-1.03
au 24/05/05	SP500	-0.50	0.46	0.96	0.52	-0.50	-0.99
Période de crise :	Nasdaq	0.75	-1.07	0.13	-0.91	1.25	-0.31
du 27/08/07	Dow Jones	-1.00	0.18	0.70	0.95	-0.26	-0.77
au 21/08/08	SP500	-0.97	0.66	0.85	1.00	-0.73	-0.97
Période post-crise :	Nasdaq	0.31	-0.32	0.93	-0.25	0.34	-0.95
du 11/01/10	Dow Jones	1.46	0.03	-0.51	-1.58	0.12	0.47
au 05/01/11	SP500	-0.59	0.56	0.93	0.66	-0.52	-0.95

TABLE C.5 – Paramètres du processus ARMA(3,3) avec modèle GJR-GARCH(1,1) pour les résidus

Période	Indice	ω	a1	b1	γ1
Période pré-crise :	Nasdaq	7.6 e^{-06}	9.4 e^{-12}	8.6 e^{-01}	1.0 e^{-01}
du 28/05/04	Dow Jones	2.2 e^{-06}	1.1 e^{-14}	9.0 e^{-01}	1.0 e^{-01}
au 24/05/05	SP500	3.8 e^{-06}	1.3 e^{-18}	8.7 e^{-01}	9.8 e^{-02}
Période de crise :	Nasdaq	1.1 e^{-05}	3.9 e^{-05}	8.5 e^{-01}	1.7 e^{-01}
du 27/08/07	Dow Jones	2.6 e^{-06}	3.9 e^{-08}	9.5 e^{-01}	5.3 e^{-02}
au 21/08/08	SP500	6.0 e^{-06}	7.3 e^{-15}	9.2 e^{-01}	7.1 e^{-02}
Période post-crise :	Nasdaq	3.9 e^{-06}	9.1 e^{-09}	8.7 e^{-01}	2.1 e^{-01}
du 11/01/10	Dow Jones	1.5 e^{-06}	2.0 e^{-10}	9.0 e^{-01}	2.0 e^{-01}
au 05/01/11	SP500	2.9 e^{-06}	3.9 e^{-10}	8.6 e^{-01}	2.6 e^{-01}

TABLE C.6 – Paramètres du modèle GJR-GARCH(1,1)

Période	Indice	ar1	ar2	ar3	ma1	ma2	ma3
Période pré-crise :	Nasdaq	6.6 e^{-02}	4.5 e^{-02}	6.6 e^{-02}	6.9 e^{-03}	7.3 e^{-03}	5.6 e^{-03}
du 28/05/04	Dow Jones	1.7 e^{-02}	1.4 e^{-02}	1.1 e^{-03}	7.2 e^{-03}	7.9 e^{-03}	7.1 e^{-07}
au 24/05/05	SP500	4.9 e^{-02}	2.1 e^{-02}	2.1 e^{-02}	4.4 e^{-02}	2.4 e^{-02}	2.3 e^{-02}
Période de crise :	Nasdaq	3.3 e^{-03}	4.0 e^{-03}	4.3 e^{-04}	4.8 e^{-03}	8.1 e^{-03}	5.3 e^{-04}
du 27/08/07	Dow Jones	2.1 e^{-02}	4.0 e^{-02}	3.4 e^{-02}	5.4 e^{-03}	5.9 e^{-03}	1.4 e^{-03}
au 21/08/08	SP500	1.7 e^{-03}	9.5 e^{-04}	1.9 e^{-03}	7.4 e^{-04}	1.6 e^{-03}	7.0 e^{-04}
Période post-crise :	Nasdaq	3.2 e^{-02}	3.8 e^{-02}	3.8 e^{-02}	1.3 e^{-02}	1.7 e^{-02}	1.3 e^{-02}
du 11/01/10	Dow Jones	3.7 e^{-03}	1.1 e^{-04}	8.8 e^{-04}	8.4 e^{-03}	3.3 e^{-04}	5.9 e^{-06}
au 05/01/11	SP500	2.3 e^{-02}	2.7 e^{-02}	2.6 e^{-02}	6.5 e^{-03}	8.1 e^{-03}	2.9 e^{-03}

TABLE C.7 – Ecart-types des estimations des paramètres du processus ARMA(3,3) avec modèle GJR-GARCH(1,1) pour les résidus

Période	Indice	ω	a1	b1	γ1
Période pré-crise :	Nasdaq	5.7 e^{-08}	1.2 e^{-02}	1.6 e^{-02}	4.4 e^{-02}
du 28/05/04	Dow Jones	2.2 e^{-06}	5.4 e^{-04}	9.9 e^{-03}	7.2 e^{-02}
au 24/05/05	SP500	3.3 e^{-08}	8.2 e^{-03}	1.6 e^{-02}	4.8 e^{-02}
Période de crise :	Nasdaq	4.1 e^{-06}	1.9 e^{-02}	4.4 e^{-01}	8.5 e^{-01}
du 27/08/07	Dow Jones	2.2 e^{-05}	9.4 e^{-02}	4.5 e^{-02}	6.9 e^{-02}
au 21/08/08	SP500	1.5 e^{-06}	2.5 e^{-02}	3.5 e^{-02}	5.6 e^{-02}
Période post-crise :	Nasdaq	5.6 e^{-07}	1.8 e^{-02}	2.4 e^{-02}	7.3 e^{-02}
du 11/01/10	Dow Jones	1.1 e^{-05}	6.3 e^{-04}	2.3 e^{-01}	4.7 e^{-01}
au 05/01/11	SP500	1.8 e^{-06}	7.0 e^{-02}	3.2 e^{-02}	1.1 e^{-01}

TABLE C.8 – Ecart-types des estimations des paramètres du modèle GJR-GARCH(1,1)

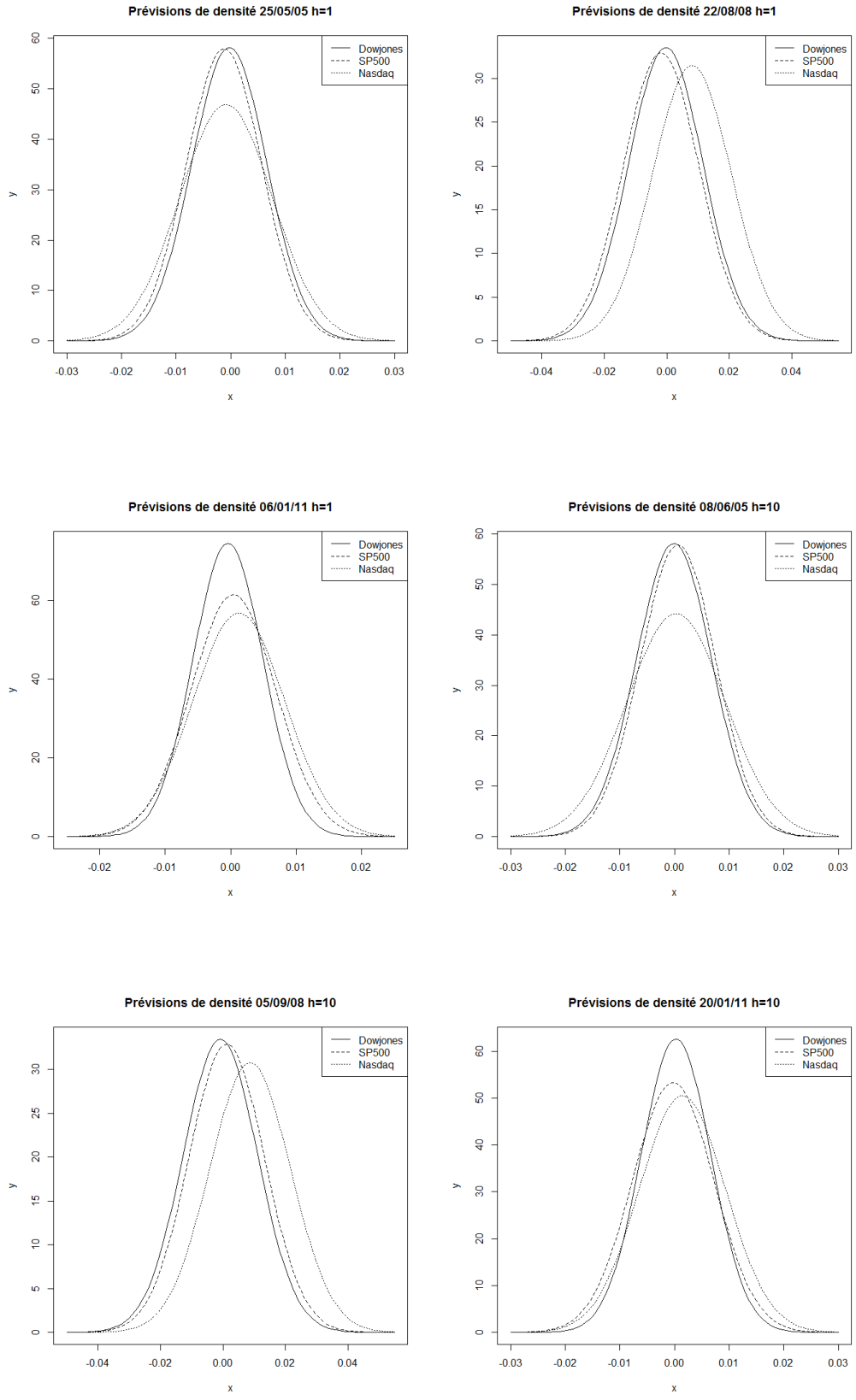

FIGURE C.2 – Premières prévisions de densité de chaque environnement avec une modélisation GARCH(1,1) pour les résidus

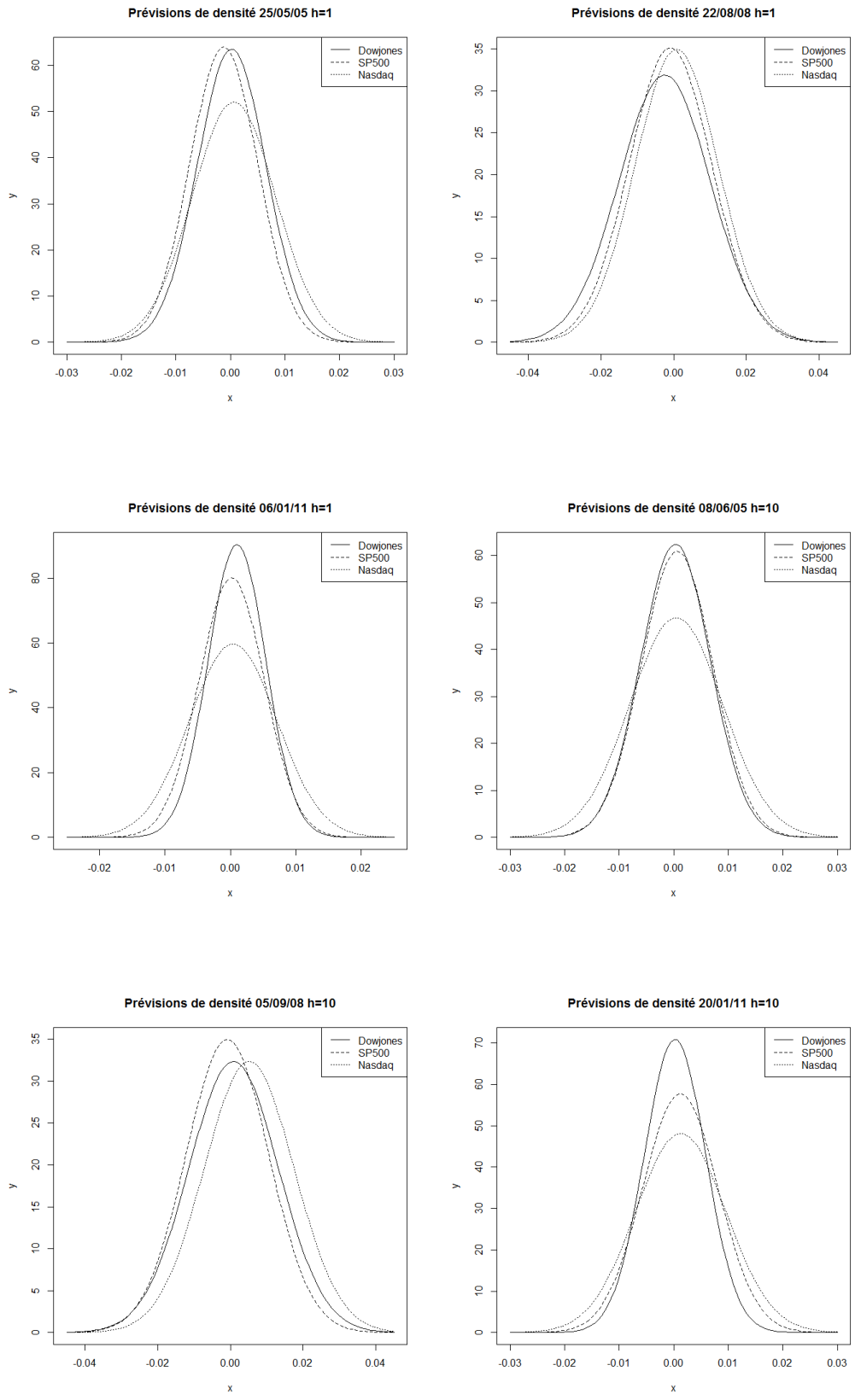

FIGURE C.3 – Premières prévisions de densité de chaque environnement avec une modélisation GJR-GARCH(1,1) pour les résidus

Appendix D

Appendix chapter 4

Here are gathered all the proofs concerning the results presented in the chapter. The first section is concerned by proofs of results in an infinite sample and the second by result in a finite sample.

Lemma 1.

$$\mathrm{E}_Y[\,L_\tau - L_{\tau_c}\,] = \sum_e p_e \int_{G_e^{-1}(\tau_c)}^{G_e^{-1}(\tau)} (F_e(y) - \tau)\, dy\,,$$

with τ,$\tau_c \in [0,1]$ and p_e the frequency of appearance of the state e.

Under the assumption **A.2.1**, we prove Lemma 1.

Proof. We have:

$$\mathrm{E}_Y[\,L_\tau - L_{\tau_c}\,] = \sum_e p_e \left(\mathrm{E}_Y[\,L_\tau(\boldsymbol{Y}, G_e^{-1}(\tau))\,] - \mathrm{E}_Y[\,L_\tau(\boldsymbol{Y}, G_e^{-1}(\tau_c)\,]\right). \tag{D.1}$$

First, we only focus on a particular e. Thus, we are interested in:

$$\mathrm{E}_Y[\,L_\tau(\boldsymbol{Y}, G_e^{-1}(\tau))\,] - \mathrm{E}_Y[\,L_\tau(\boldsymbol{Y}, G_e^{-1}(\tau_c)\,] \equiv \mathrm{E}_{Y,e}[\,L_{\tau,\tau_c}\,].$$

For ease of notation and comprehension, we suppress e in the notation since there is no confusion. Moreover, we suppose, for ease of notation again (and since we obtain the same result if we inverse the inequality) that $G^{-1}(\tau) \leq G^{-1}(\tau_c)$. So, we have:

$$\begin{aligned}
\mathrm{E}_Y[\,L_{\tau,\tau_c}\,] &= \int_{-\infty}^{+\infty} \left(\,[y - G^{-1}(\tau)]\,\tau + [G^{-1}(\tau) - y]\,\mathbf{1}_{\{y \leq G^{-1}(\tau)\}}\,\right) f_Y(y)\, dy \\
&\quad - \int_{-\infty}^{+\infty} \left(\,[y - G^{-1}(\tau_c)]\,\tau + [G^{-1}(\tau_c) - y]\,\mathbf{1}_{\{y \leq G^{-1}(\tau_c)\}}\,\right) f_Y(y)\, dy \\
&= [G^{-1}(\tau_c) - G^{-1}(\tau)]\,\tau + [G^{-1}(\tau) - G^{-1}(\tau_c)]\,F \circ G^{-1}(\tau) \\
&\quad - G^{-1}(\tau_c)\,[\,F \circ G^{-1}(\tau_c) - F \circ G^{-1}(\tau)\,] + \int_{y=G^{-1}(\tau)}^{G^{-1}(\tau_c)} \underbrace{y}_{v}\, \underbrace{f_Y(y)}_{u'}\, dy\,.
\end{aligned}$$

Using integral by parts, we have:

$$\begin{aligned} \mathrm{E}_Y[\, L_{\tau,\tau_c}\,] &= [G^{-1}(\tau_c) - G^{-1}(\tau)]\,\tau + \int_{y=G^{-1}(\tau_c)}^{G^{-1}(\tau)} F(y)\,dy \\ &= \int_{y=G^{-1}(\tau_c)}^{G^{-1}(\tau)} [\,F(y) - \tau\,]\,dy\,. \end{aligned}$$

Replacing it in (D.1) finishes the demonstration. □

D.1 Impact on score: conditions for improvement

In this section, the reader can find the proofs of results mentioned in Sect.4.3.1 of the chapter. We first demonstrate how to approximate the difference of L_τ expectation before showing that under some hypotheses, our correction improves systematically the quality of the forecasts.

D.1.1 Rewriting the difference of L_τ expectation

Under the assumptions **A.2.1**, **A.3.1.1**, **A.3.1.2**, **A.3.1.3** and **A.3.1.4** and using functional derivatives and the implicit function theorem, we prove (4.2).

Proof. Remember: Let H be a functional, h a function, α a scalar and δ an arbitrary function.

We can write the expression of the functional evaluated at $f + \delta\alpha$ as follow:

$$H[h + \delta\,\alpha] = H[h] + \frac{dH[h + \delta\,\alpha]}{d\alpha}|_{\alpha=0}\,\alpha + \frac{1}{2}\frac{d^2H[h + \delta\,\alpha]}{d\alpha^2}|_{\alpha=0}\,\alpha^2 + \cdots + \mathrm{Rem}(\alpha)\,,$$

with $\mathrm{Rem}(\alpha)$ the remainder.

Denote:

$$\begin{aligned} \Delta PL[h] &= \sum_e p_e \int_{h_e^{-1}(\tau_c)}^{h_e^{-1}(\tau)} (F_e(y) - \tau)dy \\ &= \sum_e p_e \Delta PL_e[h_e]\,. \end{aligned}$$

For ease of notation, denote $\Delta PL_e[F_e + \delta_e\alpha] \equiv \Delta PL_{F,\delta,e}$. Choosing $H = \Delta PL_e$, $h = F_e$ and $\eta_e = \alpha\delta_e$ (even if we use $\alpha\delta_e$ in the developpemnt in order to use functional derivatives, directional derivatives and the implicit function theorem), we have:

$$\Delta PL_{F,\delta,e} \sim \Delta PL_e[F_e] + \frac{d\Delta PL_{F,\delta,e}}{d\alpha}|_{\alpha=0}\,\alpha + \frac{1}{2}\frac{d^2\Delta PL_{F,\delta,e}}{d\alpha^2}|_{\alpha=0}\,\alpha^2 + \mathrm{Rem}_e(\alpha)$$

$$\begin{aligned} = &\left[\frac{\partial\Delta PL_{F,\delta,e}}{\partial\alpha}|_{\alpha=0\,,\,\tau_c=\tau} + \frac{\partial\Delta PL_{F,\delta,e}}{\partial\tau_c}|_{\alpha=0\,,\,\tau_c=\tau}\frac{d\tau_c}{d\alpha}\right]\alpha \\ &+ \left[\frac{\partial^2\Delta PL_{F,\delta,e}}{\partial\alpha^2}|_{\alpha=0\,,\,\tau_c=\tau} + 2\frac{\partial^2\Delta PL_{F,\delta,e}}{\partial\alpha\partial\tau_c}|_{\alpha=0\,,\,\tau_c=\tau}\frac{d\tau_c}{d\alpha}\right]\frac{\alpha^2}{2} \\ &+ \left[\frac{\partial^2\Delta PL_{F,\delta,e}}{\partial\tau_c^2}|_{\alpha=0\,,\,\tau_c=\tau}\left(\frac{d\tau_c}{d\alpha}\right)^2 + \frac{\partial\Delta PL_{F,\delta,e}}{\partial\tau_c}|_{\alpha=0\,,\,\tau_c=\tau}\frac{d^2\tau_c}{d\alpha^2}\right]\frac{\alpha^2}{2} \\ &+ \mathrm{Rem}_e(\alpha)\,. \end{aligned}$$

To calculate $\frac{d\tau_c}{d\alpha}$, we will use the equation which link τ_c and α:

$$\sum_e p_e F_e \circ (F_e + \delta_e \alpha)^{-1}(\tau_c) = \tau.$$

Using the implicit function theorem, we find:

$$\frac{d\tau_c}{d\alpha} = \sum_e p_e \delta_e \circ F_e^{-1}(\tau)$$

Now, we need to calculate partial derivatives:

$$\begin{aligned}
\frac{\partial \Delta PL_{F,\delta,e}}{\partial \alpha}|_{\alpha=0\,,\,\tau_c=\tau} &= \frac{\partial \left(\int_{(F_e+\delta_e\alpha)^{-1}(\tau_c)}^{(F_e+\delta_e\alpha)^{-1}(\tau)} (F_e(y)-\tau)dy \right)}{\partial \alpha}|_{\alpha=0\,,\,\tau_c=\tau} = 0\,; \\
\frac{\partial \Delta PL_{F,\delta,e}}{\partial \tau_c}|_{\alpha=0\,,\,\tau_c=\tau} &= 0 \;;\; \frac{\partial^2 \Delta PL_{F,\delta,e}}{\partial \tau_c^2}|_{\alpha=0\,,\,\tau_c=\tau} = -\frac{1}{f_e \circ F_e^{-1}(\tau)}; \\
\frac{\partial^2 \Delta PL_{F,\delta,e}}{\partial \alpha^2}|_{\alpha=0\,,\,\tau_c=\tau} &= 0 \;;\; \frac{\partial^2 \Delta PL_{F,\delta,e}}{\partial \alpha \partial \tau_c}|_{\alpha=0\,,\,\tau_c=\tau} = \frac{\delta_e \circ F_e^{-1}(\tau)}{f_e \circ F_e^{-1}(\tau)}\,.
\end{aligned}$$

Thus, we have:

$$\begin{aligned}
\Delta PL_e[F_e + \delta_e \alpha] \;\sim\; & \left[\left(\frac{\delta_e \circ F_e^{-1}(\tau)}{f_e \circ F_e^{-1}(\tau)} \right) \sum_e p_e \delta_e \circ F_e^{-1}(\tau) \right] \alpha^2 \\
& - \left[\frac{(\sum_e p_e \delta_e \circ F_e^{-1}(\tau))^2}{2 f_e \circ F_e^{-1}(\tau)} \right] \alpha^2 + \mathrm{Rem}_e(\alpha)\,,
\end{aligned}$$

and hence:

$$\begin{aligned}
\Delta PL[F + \delta\alpha] \;\sim\; & \left(\sum_e \frac{p_e \delta_e(F_e^{-1}(\tau))}{f_e(F_e^{-1}(\tau))} \right) \left(\sum_e p_e \delta_e(F_e^{-1}(\tau)) \right) \times \alpha^2 \\
& - \left(\sum_e \frac{p_e}{2 f_e(F_e^{-1}(\tau))} \right) \left(\sum_e p_e \delta_e(F_e^{-1}(\tau)) \right)^2 \times \alpha^2 \\
& + \sum_e p_e \mathrm{Rem}_e(\alpha).
\end{aligned}$$

Now, let's focus on the remainders. If M such that $\left| \frac{d^3 \Delta PL_{F,\delta,e}}{d\alpha^3} \right| \leq M$ exists, we have, according to the Taylor-Lagrange inequality, $|\,\mathrm{Rem}_e(\alpha)\,| \leq \frac{M|\alpha^3|}{3!}$. Let' s find conditions for the existence of M.

The third derivative is:

$$\begin{aligned}
\frac{d^3 \Delta PL_{F,\delta,e}}{d\alpha^3} &= \frac{\partial \Delta PL_{F,\delta,e}}{\partial \tau_c} \frac{d^3\tau_c}{\alpha^3} + 3 \frac{\partial^2 \Delta PL_{F,\delta,e}}{\partial \tau_c \partial \alpha} \frac{d^2\tau_c}{\alpha^2} + 3 \frac{\partial^2 \Delta PL_{F,\delta,e}}{\partial \tau_c^2} \frac{d^2\tau_c}{\alpha^2} \frac{\tau_c}{\alpha} \\
&+ \frac{\partial^3 \Delta PL_{F,\delta,e}}{\partial \alpha^3} + 3 \frac{\partial^3 \Delta PL_{F,\delta,e}}{\partial \tau_c \partial \alpha^2} \frac{\tau_c}{\alpha} + 3 \frac{\partial^3 \Delta PL_{F,\delta,e}}{\partial \tau_c^2 \partial \alpha} \left(\frac{\tau_c}{\alpha} \right)^2 \\
&+ \frac{\partial^3 \Delta PL_{F,\delta,e}}{\partial \tau_c^3} \left(\frac{\tau_c}{\alpha} \right)^3 .
\end{aligned}$$

Let's calculate the partial derivatives of order 3:

$$\begin{aligned}
\frac{\partial^3 \Delta PL_{F,\delta,e}}{\partial \alpha^3}|_{\alpha=0\,,\,\tau_c=\tau} &= 0 \;;\; \frac{\partial^3 \Delta PL_{F,\delta,e}}{\partial \tau_c^3}|_{\alpha=0\,,\,\tau_c=\tau} = 2\frac{f_e^{'} \circ F_e^{-1}(\tau)}{f_e \circ F_e^{-1}(\tau)}\,; \\
\frac{\partial^3 \Delta PL_{F,\delta,e}}{\partial \tau_c^2 \partial \alpha}|_{\alpha=0\,,\,\tau_c=\tau} &= -2\frac{f_e^{'} \circ F_e^{-1}(\tau)}{f_e \circ F_e^{-1}(\tau)^3}(\delta_e \circ F_e^{-1}(\tau)) - 2\frac{\delta_e^{'} \circ F_e^{-1}(\tau)}{f_e \circ F_e^{-1}(\tau)^2}\,; \\
\frac{\partial^3 \Delta PL_{F,\delta,e}}{\partial \tau_c \partial \alpha^2}|_{\alpha=0\,,\,\tau_c=\tau} &= \frac{f_e^{'} \circ F_e^{-1}(\tau)}{f_e \circ F_e^{-1}(\tau)^3}\left(\delta_e \circ F_e^{-1}(\tau)\right)^2 \\
&\quad -2\frac{\delta_e^{'} \circ F_e^{-1}(\tau)}{f_e \circ F_e^{-1}(\tau)^2}\left(\delta_e \circ F_e^{-1}(\tau)\right).
\end{aligned}$$

Moreover, we have:

$$\frac{d^2\tau_c}{d\alpha^2} = \sum_e p_e \left(\frac{2\delta_e^{'} \circ F_e^{-1}(\tau) - f_e^{'} \circ F_e^{-1}(\tau)}{f_e \circ F_e^{-1}(\tau)}\right) \delta_e \circ F_e^{-1}(\tau).$$

Since η_e, its first, second and third derivatives are finite in $F_e^{-1}(\tau)$, it is also the case for δ_e and the partial derivatives are finite. Furthermore, f_e, δ_e and their derivatives are bounded (since η_e and their derivatives are bounded), which implies that the second derivatives of $\Delta PL_e[F_e + \delta_e\alpha]$ are also bounded. Thus, under these conditions, M exists. Then, we can write $\frac{d^3 \Delta PL_{F,\delta,e}}{d\alpha^3} = M_1\delta_e^3$ and hence $|\mathrm{Rem}_e(\alpha)| \leq \frac{|M_1||\alpha\delta_e|^3}{3!}$ which implies that $\lim \frac{\mathrm{Rem}_e(\alpha)}{(\alpha\delta_e)^2} = 0$, $\alpha\delta_e \to 0$, which shows that $\mathrm{Rem}_e(\alpha)$ is negligible compared to $\frac{d^2 \Delta PL_{F,\delta,e}}{d\alpha^2}$.

Moreover, since $\forall\, e \in E$ the functions F_e are C^3 and the functions f_e and their derivatives are bounded by a constant which doesn't depend on e, $\forall\, e \in E$, the development is valid for all directions and thus, since $\eta_e = G_e - F_e$, we have:

$$\begin{aligned}
\mathrm{E}_Y[\, L_\tau - L_{\tau_c}\,] &\sim \left(\sum_e \frac{p_e \eta_e(F_e^{-1}(\tau))}{f_e(F_e^{-1}(\tau))}\right)\left(\sum_e p_e \eta_e(F_e^{-1}(\tau))\right) \\
&\quad - \left(\sum_e \frac{p_e}{2f_e(F_e^{-1}(\tau))}\right)\left(\sum_e p_e \eta_e(F_e^{-1}(\tau))\right)^2 \\
&\quad \text{as } max\ \eta_e \to 0.
\end{aligned}$$

To finish the demonstration, remark that Lemma 1 proves that:

$$\Delta PL[G] = \mathrm{E}_Y[\, L_\tau - L_{\tau_c}\,].$$

□

D.1.2 Systematic improvement of the quality

If $\exists\, \nu \geq 0$ (sufficiently small) $\forall\, e \in E \;\; \forall\, y \in \mathbf{R}; |\eta_e(y)| \leq \nu$, we show (4.3) and (4.4) under the assumption **A.3.1.5** or **A.3.1.6**.

Proof. Prove (4.3) is equivalent to show that $\Delta PL[G]$ is positive, and if we rewrite:

$$\Delta PL[G] \sim (2\mathrm{E}[f^{-1}\eta] - \mathrm{E}[f^{-1}]\mathrm{E}[\eta])\mathrm{E}[\eta]\,,$$

it is clear that the assumption **A.3.1.6** ensures the positivity of $\Delta PL[G]$.

However, we need more argumentation to understand the complete utility of the assumption **A.3.1.5**. Let's look at one of the two worst cases: only two states of the world, the correlation coefficient $\rho = -1$, $\eta > 0$ (the other case is when $\rho = 1$ and $\eta < 0$) and at each bound of the support of δ and f^{-1}, there is half of the probability mass. We also consider that the ratios between max and min of the supports are equal. If we define $max_e = M$ and $min_e = \frac{M}{r}$, one has the following equation:

$$\frac{1}{2} = \frac{2(r^2+1)}{(r+1)^2} - 1.$$

Solving this equation in r produces the expected result concerning the ratio between max and min values of η and f^{-1}.

Now, let's prove (4.4). According to (4.1), we have:

$$\mathrm{E}_Y[\,CRPS_{G,C\circ G}\,] \;=\; 2\int_{-\infty}^{+\infty}\left(\int_0^1 L_\tau(y,G^{-1}(\tau)) - L_\tau(y,G^{-1}\circ C^{-1}(\tau))d\tau\right) f_Y(y)dy.$$

We can rewrite:

$$\begin{aligned}\mathrm{E}_Y[\,CRPS_{G,C\circ G}\,] \;=\;& 2\int_{-\infty}^{+\infty}\int_0^1 L_\tau(y,G^{-1}(\tau))\, f_Y(y)\, d\tau\, dy \\ & -2\int_{-\infty}^{+\infty}\int_0^1 L_\tau(y,G^{-1}\circ C^{-1}(\tau))\, f_Y(y)\, d\tau dy\,,\end{aligned}$$

and using the Fubini-Tonelli theorem, one obtains:

$$\begin{aligned}\mathrm{E}_Y[\,CRPS_{G,C\circ G}\,] \;&=\; 2\int_0^1 \mathrm{E}_Y[\,L_\tau - L_{\tau_c}\,]d\tau \\ &\geq\; 0.\end{aligned} \tag{D.2}$$

□

D.2 Impact on score: bounds on degradation

Under the assumptions **A.2.1**, **A.3.2.1**, **A.3.2.2**, **A.3.2.3**, **A.3.2.4**, **A.3.2.5** and **A.3.2.6** we prove (4.6) and (4.7).

Proof. Adding and substracting $\mathrm{E}_Y[\,L_\tau(Y,G^{-1}(\tau_c))\,]$ to $\mathrm{E}_Y[L_{\widehat{\tau}_c} - L_\tau\,]$, we obtain:

$$\begin{aligned}\mathrm{E}_Y[L_{\widehat{\tau}_c} - L_\tau\,] \;=\;& \mathrm{E}_Y[\,L_\tau(Y,G^{-1}(Q_\tau))\,] - \mathrm{E}_Y[\,L_\tau(Y,G^{-1}(\tau_c))\,] \\ & +\mathrm{E}_Y[\,L_\tau(Y,G^{-1}(\tau_c))\,] - \mathrm{E}_Y[\,L_\tau(Y,G^{-1}(\tau))\,],\end{aligned}$$

and finally:

$$\mathrm{E}_Y[L_{\widehat{\tau}_c} - L_\tau\,] = \mathrm{E}_{Y,e}[\,L_\tau(Y,G_e^{-1}(Q_\tau))\,] - \mathrm{E}_{Y,e}[\,L_\tau(Y,G_e^{-1}(\tau_c))\,] - \mathrm{E}_Y[\,L_\tau - L_{\tau_c}\,].$$

To begin with, we treat the third term on the right side. We have:

$$\mathrm{E}_{Y,e}[\,L_{\tau,\tau_c}\,] = \int_{y=G_e^{-1}(\tau_c)}^{G_e^{-1}(\tau)} [\,F_e(y) - \tau\,]\, dy\,.$$

Using the change of variable $y = G_e^{-1}(z)$ and taking the absolute value, we find:

$$|\mathrm{E}_{Y,e}[\, L_{\tau,\tau_c}\,]| = \left| \int_{z=\tau_c}^{\tau} (F_e \circ G_e^{-1}(z) - \tau)\, \frac{1}{g_e(G_e^{-1}(z))}\, dz \right| .$$

Now, one needs to distinguish two cases.

If $\tau > \tau_c$, one has:

$$\begin{aligned} |\, \mathrm{E}_{Y,e}[\, L_{\tau,\tau_c}\,]\,| &= \int_{z=\tau_c}^{\tau} \left| (F_e \circ G_e^{-1}(z) - \tau)\, \frac{1}{g_e(G_e^{-1}(z))} \right| dz \\ &\leq \int_{z=\tau_c}^{\tau} \left| (F_e \circ G_e^{-1}(z) - \tau) \right| \xi\, dz\,. \end{aligned}$$

Since $|\, F_e(z) - G_e(z)\,| \leq \varepsilon$, $\forall z \in \mathbf{R}$, $\forall e \in E$, one obtains $|\, F_e \circ G_e^{-1}(z) - z\,| \leq \varepsilon$, $\forall z \in [0,1]$, $\forall e \in E$ and then:

— if $z = \tau$, one has $\left|\, F_e \circ G_e^{-1}(\tau) - \tau\,\right| \leq \varepsilon$,

— if $z = \tau_c$, $\left|\, F_e \circ G_e^{-1}(\tau_c) - \tau)\,\right| = \left|\, F_e \circ G_e^{-1}(\tau_c) - \tau_c + \tau_c - \tau\,\right|$.

Moreover, one has:

$$\begin{aligned} |\, \tau_c - \tau\,| &= \left| \sum_e p_e\, \left(\, \tau_c - F_e \circ G_e^{-1}(\tau_c)\,\right) \right| \\ &\leq \sum_e p_e\, \left|\, F_e \circ G_e^{-1}(\tau_c) - \tau_c\,\right| \\ &\leq \varepsilon\,, \end{aligned}$$

and finally:

$$\begin{aligned} \left|\, F_e \circ G_e^{-1}(\tau_c - \tau)\,\right| &\leq \left|\, F_e \circ G_e^{-1}(\tau_c) - \tau_c\,\right| + |\, \tau_c - \tau\,| \\ &\leq 2\,\varepsilon\,. \end{aligned}$$

One deduces, when $\tau > \tau_c$:

$$|\, \mathrm{E}_{Y,e}[\, L_{\tau,\tau_c}\,]\,| \leq 2\,(\,\tau - \tau_c\,)\,\varepsilon\,\xi.$$

When $\tau < \tau_c$, one obtains:

$$|\, \mathrm{E}_{Y,e}[\, L_{\tau,\tau_c}\,]\,| \leq \int_{z=\tau}^{\tau_c} \left|\, (F_e \circ G_e^{-1}(z) - \tau)\,\right| \xi\, dz\,,$$

and using the same arguments as previously:

$$|\, \mathrm{E}_{Y,e}[\, L_{\tau,\tau_c}\,]\,| \leq 2\,(\,\tau_c - \tau\,)\,\varepsilon\,\xi.$$

Hence, one concludes that:

$$|\, \mathrm{E}_{Y,e}[\, L_{\tau,\tau_c}\,]\,| \leq 2\,|\,\tau - \tau_c\,|\,\varepsilon\,\xi.$$

To finish, replacing $\mathrm{E}_{Y,e}[\,L_{\tau,\tau_c}\,]$ in (D.1), we have:

$$|\,\mathrm{E}_Y[\,L_\tau - L_{\tau_c}\,]\,| \leq 2\,\varepsilon^2\,\xi.$$

Now let's focus on the remainder on the right side. First, we only focus on a particular e. Thus, we are interested in:

$$\mathrm{E}_Y[\,L_\tau(Y, G_e^{-1}(Q_\tau))\,] - \mathrm{E}_Y[\,L_\tau(Y, G_e^{-1}(\tau_c))\,] \equiv \mathrm{E}_{Y,e}[L_{\widehat{\tau}_c} - L_{\tau_c}\,].$$

For ease of notation and comprehension, we suppress e in the notation since there is no confusion. So, we have:

$$\begin{aligned}\mathrm{E}_Y[L_{\widehat{\tau}_c} - L_{\tau_c}\,] &= \left(\frac{1}{2} - \tau\right)\,\mathrm{E}_Y\left[G^{-1}(Q_\tau) - G^{-1}(\tau_c)\right] \\ &\quad + \frac{1}{2}\,\mathrm{E}_Y\left[\,|Y - G^{-1}(Q_\tau)| - |Y - G^{-1}(\tau_c)|\right].\end{aligned}$$

We find:

$$\begin{aligned}|\,\mathrm{E}_Y[L_{\widehat{\tau}_c} - L_{\tau_c}\,]\,| &\leq \left|\frac{1}{2}\,\mathrm{E}_Y\left[\,|Y - G^{-1}(Q_\tau)| - |Y - G^{-1}(\tau_c)| - G^{-1}(Q_\tau) + G^{-1}(\tau_c)\,\right]\right| \\ &\quad + (1-\tau)\left|\,\mathrm{E}_Y\left[G^{-1}(Q_\tau) - G^{-1}(\tau_c)\right]\,\right|.\end{aligned}$$

Let's focus on the second term on the right side. Using a Taylor series approximation around $\tau_c \in [0,1]$ and the Taylor-Lagrange formula for the remainder, one has:

$$G^{-1}(Q_\tau) = G^{-1}(\tau_c) + \frac{1}{g(G^{-1}(\tau_c))}\,(Q_\tau - \tau_c) + \frac{g^{'}(\gamma)}{g(\gamma)^3}\,\frac{(Q_\tau - \tau_c)^2}{2}\,,$$

with $\gamma = \tau_c + (Q_\tau - \tau_c)\,\theta$, and $0 < \theta < 1$.

And so

$$(1-\tau)\left|\,\mathrm{E}_Y\left[G^{-1}(Q_\tau) - G^{-1}(\tau_c)\right]\,\right| \leq \frac{(1-\tau)\,\alpha\,\xi^3}{2}\;\frac{\lambda}{n}.$$

Now, one can study the first term on the right side. Some useful remarks before the next: one can easily see that the study of such a function can be restricted to a study on the interval $I_y :=]-\infty, G^{-1}(\tau_c)]$, since we can find results on the interval $[G^{-1}(\tau_c), \infty[$ using the same arguments.

Let's define $G^{-1}(Q_\tau) \equiv Z_\tau$, $G^{-1}_{\tau_c} \equiv G^{-1}(\tau_c)$ and $f_Y^{\,G^{-1}_{\tau_c}} \equiv f_Y(G^{-1}(\tau_c))$, for ease of notation.

Thus, we are interested in calculating:

$$\frac{1}{2}\int_{y=-\infty}^{G^{-1}_{\tau_c}} f_Y(y)\,\underbrace{\left(\,\mathrm{E}_{Z_\tau}[\,|G^{-1}_{\tau_c} - Z_\tau|\,+|Z_\tau - y|\,] - G^{-1}_{\tau_c} + y\right)}_{=\mathrm{E}_{Z_\tau}[\,|Z_\tau - y| - Z_\tau\,] + y}\,dy\,. \tag{D.3}$$

However, the function studied in the integral is complicated to work with. So, one will prefer to use its integral version, that is,

$$\mathrm{E}_{Z_\tau}[\,|Z_\tau - y| - Z_\tau\,] + y \;=\; \int_{u=-\infty}^{y} \frac{d}{du}\left(\mathrm{E}_{Z_\tau}[\,|Z_\tau - u| - Z_\tau\,] + u\right)du\,.$$

For the bounds of the integral, the upper one is obvious. To justify the lower one, it is important to note that $\lim\;\mathrm{E}_{Z_\tau}[\,|Z_\tau - y| - Z_\tau\,] + y = 0$, $y \to -\infty$.

Indeed, one has:

$$\begin{aligned}
\mathrm{E}_{Z_\tau}[\,|Z_\tau - y| - Z_\tau\,] + y &= \int_{z=-\infty}^{y} (y-z)\,h(z)\,dz + \int_{z=y}^{\infty} (z-y)\,h(z)\,dz \\
&\quad + \int_{z=-\infty}^{\infty} (y-z)\,h(z)\,dz \\
&= \int_{z=-\infty}^{y} 2(y-z)\,h(z)\,dz \\
&= 2y\,H(y) - \int_{z=-\infty}^{y} 2z\,h(z)\,dz\,,
\end{aligned}$$

with h and H the p.d.f and the c.d.f of the variable Z_τ, respectively.

If the variable Z_τ has a finite mean, $\lim h(y) = 0$, $y \to -\infty$, and thus it is clear that the choice $-\infty$ for the lower bound of the integral is the good one.

At this stage, it is not easy to see the usefulness of the transformation, but it will be after the following calculus:

$$\begin{aligned}
\frac{d}{du}(\mathrm{E}_{Z_\tau}[\,|Z_\tau - u| - Z_\tau\,] + u) &= 1 + \frac{d}{du}\left(\int_{z=-\infty}^{u} (u-z)\,h(z)\,dz\right) \\
&\quad + \frac{d}{du}\left(\int_{z=u}^{\infty} (z-u)\,h(z)\,dz\right).
\end{aligned}$$

Finally, we have:

$$\begin{aligned}
\frac{d}{du}(\mathrm{E}_{Z_\tau}[\,|Z_\tau - u| - Z_\tau\,] + u) &= \int_{z=-\infty}^{u} h(z)\,dz - \int_{z=u}^{\infty} h(z)\,dz + 1 \\
&= H(u) - (\,1 - H(u)\,) + 1 \\
&= 2\,H(u).
\end{aligned}$$

Now, it is clear that this transformation could help us for the calculus of (D.3) since it is equivalent to study:

$$\int_{y=-\infty}^{G_{\tau_c}^{-1}} f_Y(y) \left(\int_{u=-\infty}^{y} H(u)\,du\right) dy \equiv \text{Half Int.}$$

A difficulty remains, though. Indeed, f_Y in unknown, and in consequence, not easy to work with. That's why, at first, one will use $f_Y^{G_{\tau_c}^{-1}}$ for our calculus, and then we will study the impact of such a manipulation.

Let's start with the first task. Using an integral by part on Half Int:

$$\int_{y=-\infty}^{G_{\tau_c}^{-1}} \underbrace{f_Y^{G_{\tau_c}^{-1}}}_{u'} \underbrace{\left(\int_{u=-\infty}^{y} H(u)\,du\right)}_{v} dy\,.$$

One obtains:

$$\begin{aligned}
\text{Half Int} &= \left[y f_Y^{G_{\tau_c}^{-1}} \left(\int_{u=-\infty}^{y} H(u)\, du \right) \right]_{y=-\infty}^{G_{\tau_c}^{-1}} - \int_{y=-\infty}^{G_{\tau_c}^{-1}} y f_Y^{G_{\tau_c}^{-1}} H(y)\, dy \\
&= \int_{u=-\infty}^{G_{\tau_c}^{-1}} f_Y^{G_{\tau_c}^{-1}} \underbrace{[G_{\tau_c}^{-1} - u]}_{u'} \underbrace{H(u)}_{v} du \\
&= \left[f_Y^{G_{\tau_c}^{-1}} \left(u\, G_{\tau_c}^{-1} - \frac{u^2}{2} \right) H(u) \right]_{u=-\infty}^{G_{\tau_c}^{-1}} \\
&\quad - \int_{u=-\infty}^{G_{\tau_c}^{-1}} f_Y^{G_{\tau_c}^{-1}} \left(u\, G_{\tau_c}^{-1} - \frac{u^2}{2} \right) h(u)\, du\,.
\end{aligned}$$

Since $\left(u\, G_{\tau_c}^{-1} - \frac{u^2}{2} \right) = \left(\frac{(u - G_{\tau_c}^{-1})^2}{2} - \frac{(G_{\tau_c}^{-1})^2}{2} \right)$, we have:

$$\text{Half Int} = f_Y^{G_{\tau_c}^{-1}} \left(\int_{u=-\infty}^{G_{\tau_c}^{-1}} \frac{(u - G_{\tau_c}^{-1})^2}{2} h(u)\, du \right).$$

Now, using the change of variable $G(u) = z$, a Taylor series approximation around τ_c and the Taylor-Lagrange formula, one has the following approximation for Half Int:

$$\frac{f_Y^{G_{\tau_c}^{-1}}}{2} \int_{z=0}^{\tau_c} \left[\frac{1}{g(G_{\tau_c}^{-1})^2} (z - \tau_c)^2 + \frac{g^{'}(\gamma)}{g(G_{\tau_c}^{-1})\, g(\gamma)^3} (z - \tau_c)^3 + \frac{g^{'}(\gamma)^2}{4 g(\gamma)^6} (z - \tau_c)^4 \right] \phi(y) dy\,,$$

with ϕ the p.d.f of the random variable Q_τ.

Using the Jensen inequality and since $0 \leq z \leq \tau_c$, we find:

$$\begin{aligned}
|\text{Half Int}| &\leq \frac{f_Y^{G_{\tau_c}^{-1}}}{2} \left[\frac{\xi^2}{2} \frac{\lambda}{n} + \frac{\alpha \xi^4}{2} \frac{\lambda}{n} + \frac{\alpha^2 \xi^6}{8} \frac{\lambda}{n} \right] \\
&\leq \frac{1}{2} \frac{C_{int}\, \lambda}{n}.
\end{aligned}$$

Since $\frac{\lambda}{n}$, which is the variance of the random variable Q_τ, is decreasing with n, let's study:

$$\Delta_{\int f} \equiv \left| \int_{u=-\infty}^{G_{\tau_c}^{-1}} (f_Y(y) - f_Y^{G_{\tau_c}^{-1}}) \left(\int_{u=-\infty}^{y} H(u)\, du \right) dy \right|.$$

Since one supports the hypothesis that $f_Y^{'}$ is bounded, using the mean value theorem, one has:

$$\begin{aligned}\Delta_{\int f} &\leq \int_{y=-\infty}^{G_{\tau_c}^{-1}} |\, f_Y(y) - f_Y^{G_{\tau_c}^{-1}} \,| \left(\int_{u=-\infty}^{y} H(u)\, du \right) dy \\ &\leq \int_{y=-\infty}^{G_{\tau_c}^{-1}} M \underbrace{(\, G_{\tau_c}^{-1} - y\,)}_{u'} \underbrace{\left(\int_{u=-\infty}^{y} H(u)\, du \right)}_{v} dy\,,\end{aligned}$$

and thus,

$$\begin{aligned}\Delta_{\int f} &\leq M \left(\left[\left(y\, G_{\tau_c}^{-1} - \frac{y^2}{2} \right) \int_{u=-\infty}^{y} H(u)\, du \right]_{y=-\infty}^{G_{\tau_c}^{-1}} - \int_{y=-\infty}^{G_{\tau_c}^{-1}} \left(y\, G_{\tau_c}^{-1} - \frac{y^2}{2} \right) H(y) dy \right) \\ &= M \left(\frac{(G_{\tau_c}^{-1})^2}{2} \int_{u=-\infty}^{G_{\tau_c}^{-1}} H(u)\, du + \int_{u=-\infty}^{G_{\tau_c}^{-1}} \left(\frac{(u - G_{\tau_c}^{-1})^2}{2} - \frac{(G_{\tau_c}^{-1})^2}{2} \right) H(u)\, du \right) \\ &= M \int_{u=-\infty}^{G_{\tau_c}^{-1}} \underbrace{H(u)}_{v} \underbrace{\frac{(u - G_{\tau_c}^{-1})^2}{2}}_{u'} du \\ &= M \left(\left[\frac{(u - G_{\tau_c}^{-1})^3}{6} H(u) \right]_{u=-\infty}^{G_{\tau_c}^{-1}} - \int_{u=-\infty}^{G_{\tau_c}^{-1}} \frac{(u - G_{\tau_c}^{-1})^3}{6} h(u)\, du \right).\end{aligned}$$

Finally, we obtain with the same change of variable and Taylor approximation as previously:

$$\begin{aligned}\Delta_{\int f} &\leq \frac{M}{6} \int_{z=0}^{\tau_c} \left[\frac{1}{g(G_{\tau_c}^{-1}))} (\tau_c - z) + \frac{g^{'}(\gamma)}{2 g(\gamma)^3} (\tau_c - z)^2 \right]^3 \phi(z)\, dz \\ &\leq \frac{M}{6} \left[\frac{\xi^3}{2} \frac{\lambda}{n} + \frac{3\xi^3 \alpha}{4} \frac{\lambda}{n} + \frac{3\xi^3 \alpha^2}{8} \frac{\lambda}{n} + \frac{\xi^3 \alpha^3}{16} \frac{\lambda}{n} \right] \\ &\leq \frac{1}{2} \frac{C_s\, \lambda}{n}.\end{aligned}$$

Thus, one has $|\, \mathrm{E}_{Y,e}[L_{\widehat{\tau}_c} - L_{\tau_c}\,]\,| \leq \frac{(C_{int} + C_s)\lambda}{n}$. Since C_{int} and C_s do not depend on e, this result remains meaningful when we are interested in the conditional expectation with respect to the random variable E and so $|\, \mathrm{E}_Y[L_{\widehat{\tau}_c} - L_{\tau}\,]\,| \leq 2\, \varepsilon^2\, \xi + \frac{C\, \lambda}{n}$.

Moreover, using (D.2), we prove (4.7). □

Bibliography

[1] Nesreen K Ahmed, Amir F Atiya, Neamat El Gayar et Hisham El-Shishiny : An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, 29(5-6):594–621, 2010.

[2] Donald WK Andrews : A conditional Kolmogorov test. *Econometrica: Journal of the Econometric Society*, pages 1097–1128, 1997.

[3] Susan Athey : The impact of machine learning on economics. *In The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, 2018.

[4] Jushan Bai : Testing parametric conditional distributions of dynamic models. *The Review of Economics and Statistics*, 85(3):531–549, 2003.

[5] Gilbert Basset et R Koenker : Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[6] Pier Giovanni Bissiri, Chris C Holmes et Stephen G Walker : A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

[7] Leo Breiman : Random forests. *Machine learning*, 45(1):5–32, 2001.

[8] Glenn W Brier : Verification of forecasts expressed in terms of probability. *Monthey Weather Review*, 78(1):1–3, 1950.

[9] Jochen Bröcker : Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, 2009.

[10] Alex J Cannon : Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic Environmental Research and Risk Assessment*, 32(11):3207–3225, 2018.

[11] Arthur Charpentier, Emmanuel Flachaire et Antoine Ly : Econom\'etrie et machine learning. *arXiv preprint arXiv:1708.06992*, 2017.

[12] Yebin Cheng, Jan G De Gooijer et Dawit Zerom : Efficient estimation of an additive quantile regression model. *Scandinavian Journal of Statistics*, 38(1):46–62, 2011.

[13] Michael P Clements et Jeremy Smith : Evaluating forecasts from SETAR models of exchange rates. *Journal of International Money and Finance*, 20(1):133–148, 2001.

[14] Gilbert Colletaz et Christophe Hurlin : Modèles non linéaires et prévisions. 2007.

[15] Valentina Corradi et Norman R Swanson : Predictive density evaluation. *Handbook of economic forecasting*, 1:197–284, 2006.

[16] Jan G De Gooijer et Dawit Zerom : On additive conditional quantiles with high-dimensional covariates. *Journal of the American Statistical Association*, 98(461):135–146, 2003.

[17] Holger Dette et Regine Scheder : Estimation of additive quantile regression. *Annals of the Institute of Statistical Mathematics*, 63(2):245–265, 2011.

[18] Francis X Diebold, Todd A Gunther et Anthony S Tay : Evaluating Density Forecasts with Applications to Financial Risk Management. International Economic Review, vol. 39, no. 4. *In Symposium on Forecasting and Empirical Methods in Macroeconomics and Finance*, page 863, 1998.

[19] Francis X Diebold et Robert S Mariano : Comparing predictive accuracy. *Journal of Business & economic statistics*, 13(3):253–263, 1995.

[20] Kjell Doksum et Ja-Yong Koo : On spline estimators and prediction intervals in nonparametric regression. *Computational Statistics & Data Analysis*, 35(1):67–82, 2000.

[21] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola et Vladimir Vapnik : Support vector regression machines. *In Advances in neural information processing systems*, pages 155–161, 1997.

[22] Robert F Engle : Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007, 1982.

[23] BWAC Farley et W Clark : Simulation of self-organizing systems by digital computer. *Transactions of the IRE Professional Group on Information Theory*, 4(4):76–84, 1954.

[24] Matteo Fasiolo, Yannig Goude, Raphael Nedellec et Simon N Wood : Fast calibrated additive quantile regression. *arXiv preprint arXiv:1707.03307*, 2017.

[25] Vincent Fortin, Anne-catherine Favre et Mériem Saïd : Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. *Quarterly Journal of the Royal Meteorological Society*, 132(617):1349–1369, 2006.

[26] Jerome H Friedman et Werner Stuetzle : Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823, 1981.

[27] JH Friedman : Greedy function approximation: A gradient boosting machine. 1999. *DOI= http://www-stat. stanford. edu/˜ jhf/ftp/trebst. pdf.*

[28] Pierre Gaillard, Yannig Goude et Raphaël Nedellec : Additive models and robust aggregation for gefcom2014 probabilistic electric load and electricity price forecasting. *International Journal of forecasting*, 32(3):1038–1050, 2016.

[29] John W Galbraith et Simon van Norden : Assessing gross domestic product and inflation probability forecasts derived from Bank of England fan charts. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(3):713–727, 2012.

[30] Tilmann Gneiting : *Calibration of medium-range weather forecasts*. European Centre for Medium-Range Weather Forecasts, 2014.

[31] Tilmann Gneiting, Fadoua Balabdaoui et Adrian E Raftery : Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[32] Tilmann Gneiting et Matthias Katzfuss : Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

[33] Tilmann Gneiting, Adrian E Raftery, Anton H Westveld III et Tom Goldman : Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.

[34] A Gogonel, J Collet et A Bar-Hen : Improving the calibration of the best member method using quantile regression to forecast extreme temperatures. *Natural Hazards and Earth System Sciences*, 13(5):1161–1168, 2013.

[35] Gloria González-Rivera, Zeynep Senyuz et Emre Yoldas : Autocontours: dynamic specification testing. *Journal of Business & Economic Statistics*, 29(1):186–200, 2011.

[36] Gloria González-Rivera et Yingying Sun : Generalized autocontours: Evaluation of multivariate density models. *International Journal of Forecasting*, 31(3):799–814, 2015.

[37] Gloria González-Rivera et Yingying Sun : Density forecast evaluation in unstable environments. *International Journal of Forecasting*, 33(2):416–432, 2017.

[38] Peter Reinhard Hansen et Allan Timmermann : Choice of sample split in out-of-sample forecast evaluation. 2012.

[39] Kostas Hatalis et Shalinee Kishore : A composite quantile fourier neural network for multi-horizon probabilistic forecasting. *arXiv preprint arXiv:1712.09641*, 2017.

[40] Hans Hersbach : Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.

[41] Yongmiao Hong : Evaluation of out of sample probability density forecasts with applications to s&p 500 stock prices. Rapport technique, Working Paper, Cornell University, 2001.

[42] Yongmiao Hong et Haitao Li : Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. *The Review of Financial Studies*, 18(1):37–84, 2004.

[43] Yongmiao Hong, Haitao Li et Feng Zhao : Can the random walk model be beaten in out-of-sample density forecasts? Evidence from intraday foreign exchange rates. *Journal of Econometrics*, 141(2):736–776, 2007.

[44] Joel L Horowitz et Sokbae Lee : Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, 100(472):1238–1249, 2005.

[45] JRM Hosking : Equivalent forms of the multivariate portmanteau statistic. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 261–262, 1981.

[46] Changha Hwang et Jooyong Shim : A simple quantile regression via support vector machine. *In International Conference on Natural Computation*, pages 512–520. Springer, 2005.

[47] Tadeusz Inglot et Teresa Ledwina : Towards data driven selection of a penalty function for data driven Neyman tests. *Linear algebra and its applications*, 417(1):124–133, 2006.

[48] Carlos M Jarque et Anil K Bera : A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, pages 163–172, 1987.

[49] Leena Kalliovirta : Misspecification tests based on quantile residuals. *The Econometrics Journal*, 15(2):358–393, 2012.

[50] Estate V Khmaladze : Martingale approach in the theory of goodness-of-fit tests. *Theory of Probability & Its Applications*, 26(2):240–257, 1982.

[51] Malte Knüppel : Evaluating the calibration of multi-step-ahead density forecasts using raw moments. 2011.

[52] Roger Koenker *et al.* : Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3):239–262, 2011.

[53] Roman Krzysztofowicz : Bayesian processor of output: a new technique for probabilistic weather forecasting. *In 17th Conference on Probability and Statistics in the Atmospheric Sciences*, volume 4, 2004.

[54] Piotr Kulczycki et H Schioler : Estimating conditional distributions by neural networks. *In Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, volume 2, pages 1344–1349. IEEE, 1998.

[55] Mark Landry, Thomas P Erlinger, David Patschke et Craig Varrichio : Probabilistic gradient boosting machines for gefcom2014 wind forecasting. *International Journal of Forecasting*, 32(3): 1061–1066, 2016.

[56] Breiman Leo, Jerome H Friedman, Richard A Olshen et Charles J Stone : Classification and regression trees. *Wadsworth International Group*, 1984.

[57] Juan Lin et Ximing Wu : A sequential test for the specification of predictive densities. *The Econometrics Journal*, 2017.

[58] Nicolai Meinshausen : Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun): 983–999, 2006.

[59] P-A Michelangeli, Matthieu Vrac et H Loukos : Probabilistic downscaling approaches: Application to wind cumulative distribution functions. *Geophysical Research Letters*, 36(11), 2009.

[60] Leslie H Miller : Table of percentage points of Kolmogorov statistics. *Journal of the American Statistical Association*, 51(273):111–121, 1956.

[61] James Mitchell et Stephen G Hall : Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR 'fan'charts of inflation. *Oxford bulletin of economics and statistics*, 67(s1):995–1033, 2005.

[62] James Mitchell et Kenneth F Wallis : Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. *Journal of Applied Econometrics*, 26(6):1023–1040, 2011.

[63] Sendhil Mullainathan et Jann Spiess : Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

[64] Allan H Murphy : A new vector partition of the probability score. *Journal of applied Meteorology*, 12(4):595–600, 1973.

[65] John Ashworth Nelder et Robert William Maclagan Wedderburn : *Generalized linear models*, volume 135. Wiley Online Library, 1972.

[66] Jerzy Neyman : « smooth test» for goodness of fit. *Scandinavian Actuarial Journal*, 1937(3-4):149–199, 1937.

[67] Sung Yong Park et Yupeng Zhang : Density forecast evaluation using data-driven smooth test, 2010.

[68] Murray Rosenblatt : Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472, 1952.

[69] Barbara Rossi et Tatevik Sekhposyan : Conditional predictive density evaluation in the presence of instabilities. *Journal of Econometrics*, 177(2):199–212, 2013.

[70] Bernhard Schölkopf, Alexander J Smola, Francis Bach *et al.* : *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[71] Kyung Ha Seok, Daehyeon Cho, Changha Hwang et Jooyong Shim : Support vector quantile regression using asymmetric e-insensitive loss function. *In Education Technology and Computer (ICETC), 2010 2nd International Conference on*, volume 1, pages V1–438. IEEE, 2010.

[72] Ben Sherwood, Lan Wang *et al.* : Partially linear additive quantile regression in ultra-high dimension. *The Annals of Statistics*, 44(1):288–317, 2016.

[73] Joo-Yong Shim et Chang-Ha Hwang : Support vector quantile regression with weighted quadratic loss function. *Communications for Statistical Applications and Methods*, 17(2):183–191, 2010.

[74] Souhaib Ben Taieb, Raphaël Huser, Rob J Hyndman et Marc G Genton : Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression. *IEEE Transactions on Smart Grid*, 7(5):2448–2455, 2016.

[75] Souhaib Ben Taieb, Raphael Huser, Rob J Hyndman, Marc G Genton *et al.* : Probabilistic time series forecasting with boosted additive models: an application to smart meter data. Rapport technique, Monash University, Department of Econometrics and Business Statistics, 2015.

[76] Ichiro Takeuchi, Quoc V Le, Timothy D Sears et Alexander J Smola : Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7(Jul):1231–1264, 2006.

[77] Leonard J Tashman : Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting*, 16(4):437–450, 2000.

[78] James W Taylor : A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, 19(4):299–311, 2000.

[79] Timo Teräsvirta : Forecasting economic variables with nonlinear models. *Handbook of economic forecasting*, 1:413–457, 2006.

[80] V Vapnik et A Chervonenkis : On a perceptron class. *Automation and Remote Control*, 25:112–120, 1964.

[81] Kenneth F Wallis : Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts. *International Journal of Forecasting*, 19(2):165–175, 2003.

[82] Wen-Chuan Wang, Kwok-Wing Chau, Chun-Tian Cheng et Lin Qiu : A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *Journal of hydrology*, 374(3-4):294–306, 2009.

[83] Kenneth D West et Michael W McCracken : Regression-based tests of predictive ability, 1998.

[84] Daniel S Wilks : *Statistical Methods in the Atmospheric Sciences (International Geophysics Series; V. 91)*. Academic Press, 2006.

[85] Daniel S Wilks : Enforcing calibration in ensemble postprocessing. *Quarterly Journal of the Royal Meteorological Society*, 144(710):76–84, 2018.

[86] Simon N Wood, Natalya Pya et Benjamin Säfken : Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563, 2016.

[87] Qifa Xu, Kai Deng, Cuixia Jiang, Fang Sun et Xue Huang : Composite quantile regression neural network with applications. *Expert Systems with Applications*, 76:129–139, 2017.

[88] Qifa Xu, Jinxiu Zhang, Cuixia Jiang, Xue Huang et Yaoyao He : Weighted quantile regression via support vector machine. *Expert Systems with Applications*, 42(13):5441–5451, 2015.

[89] Hong ZHANG et Zhen ZHANG : Feedforward networks with monotone constraints. *In Neural Networks, 1999. IJCNN'99. International Joint Conference on*, volume 3, pages 1820–1823. IEEE, 1999.

[90] Songfeng ZHENG : Qboost: Predicting quantiles with boosting for regression and binary classification. *Expert Systems with Applications*, 39(2):1687–1697, 2012.

Michael RICHARD

Évaluation et validation de prévisions en loi

Résumé : Cette thèse porte sur l'évaluation et la validation de prévisions en loi.

Dans la première partie, nous nous intéressons à l'apport du machine learning vis à vis des prévisions quantile et des prévisions en loi. Pour cela, nous avons testé différents algorithmes de machine learning dans un cadre de prévisions de quantiles sur données réelles. Nous tentons ainsi de mettre en évidence l'intérêt de certaines méthodes selon le type de données auxquelles nous sommes confrontés.

Dans la seconde partie, nous exposons quelques tests de validation de prévisions en loi présents dans la littérature. Certains de ces tests sont ensuite appliqués sur données réelles relatives aux log-rendements d'indices boursiers.

Dans la troisième, nous proposons une méthode de recalibration permettant de simplifier le choix d'une prévision de densité en particulier par rapport à d'autres. Cette recalibration permet d'obtenir des prévisions valides à partir d'un modèle mal spécifié. Nous mettons également en évidence des conditions sous lesquelles la qualité des prévisions recalibrées, évaluée à l'aide du CRPS, est systématiquement améliorée, ou très légèrement dégradée. Ces résultats sont illustrés par le biais d'applications sur des scénarios de températures et de prix.

Mots clés : Prévision en loi, prévision de quantile, machine learning, tests de validité, calibration, correction de biais, transformation de Rosenblatt, Pinball-Loss, CRPS.

Evaluation and validation of predictive densities

Abstract : In this thesis, we study the evaluation and validation of predictive densities.

In a first part, we are interested in the contribution of machine learning in the field of quantile and density forecasting. We use some machine learning algorithms in quantile forecasting framework with real data, in order to highlight the efficiency of particular method varying with nature of the data.

In a second part, we expose some validation tests of predictive densities present in the literature. As illustration, we use two of the mentionned tests on real data concerned about stock indexes log-returns.

In the third part, we address the calibration constraint of probability forecasting. We propose a generic method for recalibration, which allows us to enforce this constraint. Thus, it permits to simplify the choice between some density forecasts. It remains to be known the impact on forecast quality, measured by predictive distributions sharpness, or specific scores. We show that the impact on the Continuous Ranked Probability Score (CRPS) is weak under some hypotheses and that it is positive under more restrictive ones. We use our method on weather and electricity price ensemble forecasts.

Keywords : Density forecasting, quantile forecasting, machine learning, validity tests, calibration, bias correction, PIT series , Pinball-Loss, CRPS.

Laboratoire d'Économie d'Orléans FRE CNRS 2014
11 Rue de Blois – BP 6739
45067 Orléans Cedex 2