Précisions importantes sur le backtesting comparatif de la VaR

Samir Saissi Hassani
Chaire de recherche du Canada en gestion des risques
HEC Montréal
samir.saissi-hassani@hec.ca

24 février 2022

Résumé

La VaR garde toute son importance dans la gestion du risque de marché vu que Bâle garde l'essentiel du backtesting basé sur la VaR à 1%. Le backtesting comparatif tel que pratiqué dans la littérature actuelle souffre d'un problème majeur double. D'une part, les fonctions de score, bien que strictement consistantes, peuvent attribuer de très bons scores, voire les meilleurs, à des modèles clairement défaillants. D'autre part, le test DM (Diebold et Mariano, 1995), tel que largement utilisé pour valider les fonctions de score, échoue dans la détection de ces "faux meilleurs" modèles. Nous documentons ces faits par des cas concrets. Pour corriger cela, nous proposons de construire le test DM avec les fonctions d'identification associées aux fonctions de score plutôt que sur les fonctions de score elles-mêmes.

De plus, nous apportons une amélioration clé au test de calibration conditionnelle de Nolde et Ziegel (2017). Ceci permet une validation complémentaire des modèles via les fonctions de score, tout en offrant une façon nouvelle de tester l'hypothèse de couverture conditionnelle du côté du backtesting standard. La démarche parallèle et collaborative des deux volets du backtesting de la VaR, le standard et celui comparatif, permet une richesse conceptuelle accrue et une plus grande robustesse du backtesting global. Enfin, vu les similitudes conceptuelles, les points traités quant à la VaR devraient également concerner le backtesting comparatif de la CVaR.

Mots clés : Valeur à risque, risque de marché, accords de Bâle, backtesting standard, standard comparatif, fonctions de score, couverture inconditionnelle, couverture conditionnelle.

Codes JEL : C44, C46, C52, G21, G24, G28, G32.

Important facts on comparative backtesting of Value at Risk

Samir Saissi Hassani
Canada Research Chair in Risk Management
HEC Montréal
samir.saissi-hassani@hec.ca

24 February 2022

Abstract

VaR remains important in market risk management as Basel keeps most of the backtesting based on 1% VaR. Comparative backtesting as practiced in the current literature suffers from a major double problem. On the one hand, the score functions, although strictly consistent, may assign very good or even the best scores to clearly failing models. On the other hand, the DM test (Diebold and Mariano, 1995), as widely used to validate the score functions, fails to detect these "false best" models. We document these facts with concrete cases. To correct this issue, we propose to build the DM test using the identification functions associated with the score functions rather than on the score functions themselves.

In addition, we provide a key improvement to the conditional calibration test of Nolde and Ziegel (2017). This allows for additional model validation via score functions, while providing a novel way to test the conditional coverage assumption on the standard backtesting side. The parallel and collaborative approach of the two VaR backtesting components, the standard and the comparative one, allows for increased conceptual richness and robustness of the overall backtesting. Finally, given the conceptual similarities, the points addressed regarding VaR should also concern the comparative backtesting of CVaR.

Keywords: Value at Risk, market risk, Basel settlements, standard backtesting, comparative backtesting, score functions, unconditional coverage, conditional coverage.

JEL codes: C44, C46, C52, G21, G24, G28, G32.

Précisions importantes sur le backtesting comparatif de la VaR

1. Introduction

La VaR garde toute son importance comme mesure du risque de marché vu que l'essentiel du backtesting reste basé sur la VaR 1 % à la suite de la décision de Bâle de déterminer le capital de couverture par la CVaR à 2,5% (BCBS, 2016, 2019). Cette innovation dans le calcul de couverture est motivée par la queue de la distribution des pertes, de plus en plus épaisse. Lorsque les rendements pouvaient être supposés normaux, cela aurait représenté une augmentation du capital d'environ 0,5%[1]. L'écart (CVaR 2,5 % – VaR 1 %) devient de plus en plus conséquent à mesure que la queue gauche des rendements s'épaissit. Autrement dit, si Bâle augmente la couverture à CVaR 2,5 % et ne demande pas de valider les violations par rapport au nouveau niveau du capital c'est que c'est plus exigeant de laisser les backtests des violations sur VaR 1 %. En effet, les violations atteignent plut tôt et plus fréquemment les niveaux les plus bas. En conséquence, plus la queue de la distribution à modéliser est épaisse, plus exigeant est le backtesting sur la VaR 1 %, ce qui nous pousse à croire que le backtesting de la VaR 1 % est là pour rester, et qu'il sera très probablement maintenu en parallèle après que l'un ou l'autre des backtests de la CVaR aura été suffisamment fiabilisé pour devenir un standard de Bâle. D'où, encore une fois, l'importance capitale que garderont la VaR et le backtesting de la VaR dans la gestion du risque de marché.

La validation des modèles du risque de marché aura été principalement basée sur les backtests dits standards ou conventionnels, composés de tests inconditionnels (Unconditional Coverage) supposant l'indépendance des dépassements de VaR et ceux conditionnels (Conditional Coverage) tenant compte explicitement de la dépendance des violations de VaR. La décision de Bâle de passer à la CVaR a lancé des débats sur la non-élicitabilité de cette mesure de risque et a propulsé en avant plan le concept de backtesting comparatif par fonctions de score. En effet, la preuve de non-élicitabilité de la CVaR de

[1] En effet, la VaR 1 % d'une N(0, 1) est $\Phi^{-1}(0{,}01) = 2{,}326$ et la CVaR 2,5% est $\phi[\Phi^{-1}(0{,}025)]/0{,}025 = 2{,}338$ d'où une variation relative de $2{,}338/2{,}326 - 1 = 0{,}51\%$ (ϕ et Φ sont la densité et la cumulative de la loi normale standard).

Gneiting (2011) est bâtie sur la non-existence d'une fonction de score consistante pour cette mesure. La popularité grandissante des fonctions de score, aussi bien pour la VaR que la CVaR, les rend incontournables depuis une série d'articles sur les fondements théoriques dont ceux de Fissler, Ziegel et Gneiting (2016), Fissler et Ziegel (2016) et Nolde et Ziegel (2017).

Intuitivement, une fonction de score, S, permet d'attribuer à une statistique X une pénalité $S(X)$ selon l'erreur de l'estimé de X par rapport à sa "vraie" valeur. Concrètement, c'est un outil pour évaluer et classer en expost des modèles concurrents quant à leur pouvoir prédictif (*forecasting*) calculé en ex-ante. L'idée remonte à Mincer et Zarnowitz (1969) et Thomson (1979), entre autres. La popularité grandissante des fonctions de score est également motivée par certaines lacunes des backtests standards. La critique principale adressée à ces derniers serait le manque de puissance des tests dans certains cas (voir par exemple Berkowitz et al., 2011). Aussi, les conclusions des backtests standards sont binaires (rejet/non-rejet). Elles ne délivrent pas d'information permettant d'identifier les modèles concurrents les plus performants pour représenter le risque incorporé aux données par rapport aux modèles les moins performants. Les fonctions de score peuvent fournir ce genre d'information sous certaines conditions.

La notion de consistance stricte est fondamentale pour établir un classement de modèles concurrents. Les travaux de Thomson (1979), Osband (1985) et Saerens (2000) sont pionniers dans les développements autour de cette propriété. Les fonctions de score strictement consistantes pour la VaR peuvent s'écrire selon une formulation générale décrite et discutée dans Gneiting et Raftery (2007, éq. (40)).

Nous mettons le doigt sur des faits et des résultats importants à prendre en compte concernant le backtesting comparatif de la VaR. Le premier est que, dans des conditions plutôt habituelles, les fonctions, bien que strictement consistantes, de score peuvent attribuer de très bons scores, voire les meilleurs, à des modèles définitivement défaillants. De plus, le test DM (Diebold et Mariano, 1995), tel que construit et largement utilisé pour valider les fonctions de score, échoue inévitablement dans la détection de ces "faux meilleurs" modèles !

Nous montrons en nous référant aux fondements théoriques du test DM que les fonctions de score ne sont pas conçues pour convenir et ne devraient pas être utilisées pour calculer les *loss differentials* du test DM. D'autre part, nous démontrons comment plutôt bâtir les *loss differentials* du test DM à partir des fonctions d'identification associées aux fonctions de score au lieu que ce soit avec ces dernières directement.

Les contributions de la présente recherche se situent en grande partie dans ce contexte précis que nous présentons à la littérature pour la première fois à notre connaissance. De plus, nous documentons nos raisonnements théoriques avec des cas concrets à partir de nos données et nos modèles qui étayent la justesse des démarches proposées.

Dans une seconde partie de la contribution, nous attirons l'attention sur l'importance du choix des fonctions de score à utiliser. Les fonctions ne sont pas équivalentes, ni interchangeables. Entre autres, elles n'assurent pas la même puissance du test DM. Patton et Sheppard (2009) signalent que les fonctions de score non 0-homogènes[2] peuvent offrir une puissance insuffisante du test DM pouvant affecter drastiquement les résultats du test. Nous montrons, par un cas concret, que les fonctions d'identification associées à des fonctions non 0-homogènes occasionnent également une puissance non optimale du test DM. La fonction *tick loss,* n'étant pas 0-homogène, souffre de ce fait. La puissance du test DM en est si faible que les conclusions du test sont inutilisables. Malgré sa popularité dans la littérature, où elle est souvent utilisée seule pour conclure quant aux modèles de VaR à retenir, nous suggérons fortement de remplacer avantageusement la fonction *tick loss* par la fonction de score *log los*s, strictement consistante et 0-homogène, comme il convient (voir Nolde et Ziegel (2017, équation (2.2)).

Nous utilisons six modèles, dont deux non paramétriques, tous de caractéristiques différentes, que nous mettons à rude épreuve avec des données provenant d'une période de turbulence extrême des marchés. L'un des modèles est RiskMetrics de J.P. Morgan (1996). Dans sa mouture la plus simple, ce modèle va jouer un rôle central dans tout ce

[2] Une fonction de score S est dite b-homogène, ou homogène de degré b, si pour tout $c > 0$ on a $S(cv_t, cy_t) = c^b S(v_t, y_t)$ où $b \in \mathbb{R}$ est le degré d'homogénéité de S. Si $b = 0$, la fonction est alors 0-homogène. Voir les détails et effet sur le test DM dans Nolde et Ziegel (2017) et Patton et al. (2009).

travail. Entre autres, les fonctions de score vont le désigner unanimement comme meilleur modèle, alors qu'il est le plus mauvais modèle vu les conclusions des backtests standards ! Dans la suite, nous appellerons ce phénomène un faux meilleur modèle.

Une troisième contribution de la présente note consiste à apporter une amélioration déterminante afin de mettre à profit le concept du test de calibration conditionnelle de Nolde et Ziegel (2017) pour contrôler plus la détection des faux meilleurs modèles. L'amélioration que nous apportons serait assez importante pour renverser l'avis plutôt mitigé de ses propres concepteurs.[3]

L'une des clés de voûte de cette recherche est la conduite en parallèle du backtesting de la VaR dans ses deux volets, standard et comparatif. Le document est organisé de la façon suivante. La section 2 présente les données de l'échantillon et les modèles concurrents que nous utilisons. Un rappel théorique des backtests standards est dressé dans la section 3.1. Les résultats correspondants sont exposés à la section 3.2. Les fonctions de score et le test DM occupent les sections 4.1 et 4.2, respectivement. Suivra une critique de l'utilisation du test DM appliqué sur les fonctions de score dans la section 4.3, puis une proposition d'utiliser le test DM plutôt sur les fonctions d'identification associées aux scores. Les preuves et les résultats des calculs sont dans la section 4.4. Une façon améliorée de mettre en oeuvre le test de calibration conditionnelle est exposée dans la section 4.5; cela inclut le rôle de cette construction pour à la fois détecter les faux meilleurs modèles du côté backtesting comparatif tout en créant une nouvelle technique d'effectuer le backtest cc standard (Conditional Coverage). Enfin, la section 5 propose une conclusion.

2. Données et modèles

Les données correspondent aux rendements quotidiens du S&P 500. La période étudiée est du 08 avril 1994 au 10 janvier 2005, ce qui englobe des turbulences extrêmes des marchés pendant les années 1990 (crise asiatique en 1997, crise de la dette russe en 1998), la récession de 2001 (NBER: de mars 2001 à novembre 2001) et, enfin, la crise *dot.com* dont

[3] Voir aussi Patton et al. (2019).

les pertes culminent en automne 2002. La figure 1 montre l'amplitude et la fréquence des pertes quotidiennes enregistrées (traits gris verticaux). On peut y voir les pertes qui dépassent la VaR à 1 % calculée en appliquant le modèle normal (trait fin en bleu). Sur les 2710 observations, il y a 46 dépassements de cette VaR. Ceci représente un taux de 1,697 %, ce qui est clairement au-dessus de ce que suppose la couverture théorique à 1 %. De même, la VaR à 1 % du modèle *t*-Student (trait épais rouge) est dépassée 38 fois, d'où une fréquence de hit de 1,402 %. Les turbulences semblent être très importantes et devraient exiger des modèles de VaR plus performants, comme on verra par la suite.

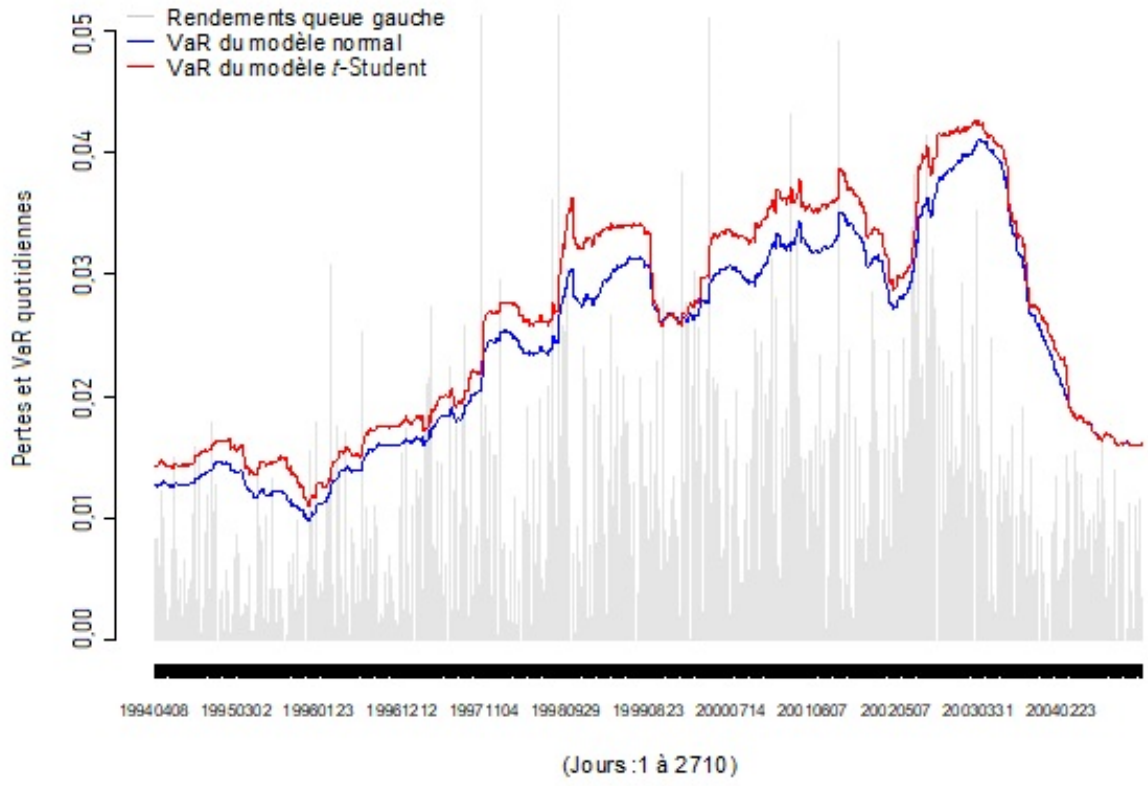

Figure 1. Tracé des pertes quotidiennes du S&P 500 (du 8 avril 1994 au 10 janvier 2005) et des VaR quotidiennes des modèles normal et *t*-Student

Tableau 1 : Statistiques descriptives des rendements quotidiens

							——– Quantiles ——–			
Min.	1er quartile	Médiane	Moyenne	3e quartile	Max.	Écart type	1 %	5 %	95 %	99 %
-5,11878	-0,53550	0,04748	0,04332	0,62326	5,20685	3,28575	-4,88961	-3,97296	4,06095	4,97767

Valeurs des rendements en %

Pour assurer le recul de 250 jours aux backtests en hors-échantillon (*Out-of-Sample*), le calcul des rendements est effectué à partir du 14 avril 1993. Le présent échantillon contient 2710 observations. Trois valeurs ont été considérées aberrantes et sont remplacées par 4,5 fois l'écart type global. Les statistiques descriptives sont résumées au Tableau 1.

2.1. Modèles de VaR concurrents

On passe en revue 6 modèles concurrents : simulation historique (SimHist, *Historical Simulation*), *RiskMetrics* de J.P. Morgan (1996) (RMetrics), modèle normal (NO), modèle *t*-Student (TF), le *Skewed Exponential Power* type 3 (SEP3) et le Skewed-t type 4 (ST4); voir définitions et détails des 4 derniers modèles paramétriques dans Rigby et al. (2019), Stasinopoulos et al. (2017) et Saissi Hassani et Dionne (2021). Ce dernier document fournit les expressions analytiques pour calculer les VaR.

La simulation historique, bien que très rudimentaire, serait utilisée par 75 % de banques américaines, selon Mehta et al. (2012). Plus récemment, selon Woodall (2021), elle est utilisée pour le backtesting réglementaire sur la VaR 1% par 72 % des banques européennes selon les résultats d'une enquête dirigée en 2020 par la EBA (European Banking Authority). Par ailleurs, la simulation historique, telle quelle ou avec des variantes, est souvent employée dans la littérature du backtesting comparatif pour servir de repère minimal (benchmark) afin de baliser les performances des modèles concurrents d'intérêt. Le modèle SimHist que nous employons utilise une fenêtre de 250 jours concernant les procédures out-of-sample (voir, par exemple, Nolde et Ziegel (2017), Taylor (2019) et Patton et al. (2019)). De plus, comme on le verra, malgré sa simplicité, ce modèle performe plutôt très bien malgré les difficultés qu'imposent les turbulences incorporées dans l'échantillon. Cela semble confirmer le choix des banques, selon les chiffres des enquêtes citées. Enfin, nous suivons Rockafellar et Uryasev (2002) pour estimer les VaR empiriques du modèle SimHist.

Le modèle RiskMetrics de J.P. Morgan (1996), non paramétrique, est utilisé tel quel dans certains cas (voir Liu et al., 2020) ou en tant que cas particulier de la famille de modèles plus élaborés sous EWMA (Exponential Weighted Moving Average) ou, plus généralement encore, dans les modèles récents dits GAS (Generalized Autoregressive Score Models). Il

semble continuer à susciter de l'intérêt pour les praticiens - voir par exemple Lucas et Zhang (2016) - ou du côté académique comme dans Liu et al. (2020). Dans sa parure la plus simple, avec $\lambda = 0{,}96$, ce modèle va jouer un rôle important tout au long de la présente recherche.

Le modèle normal (NO) et celui *t*-Student (TF) sont des classiques afin d'en visualiser les comportements par rapport aux autres modèles et aux données utilisées. Le modèle SEP3 représente un intérêt particulier. En effet, des travaux récents sur les fonctions de score de la CVaR ont montré que les modèles bâtis autour de la distribution de Laplace asymétrique ont une fonction de vraisemblance qui, elle-même, peut être naturellement transformée en une fonction de score de la CVaR strictement consistante (voir détails dans Taylor (2019) et Patton et al. (2019)).

Quant au modèle ST4, c'est une extension de la distribution Skewed-t conçue par Fernandez et Steel (1998). Nolde et Ziegel (2017) utilisent ce modèle pour illustrer leurs développements. La distribution ST4 a un paramètre de plus pour modéliser l'épaisseur des queues gauche et droite d'une façon indépendante. Voir Saissi Hassani et Dionne (2021) pour plus de détails sur ce modèle qui performe bien pour modéliser le risque de marché.

2.2. Notations et définitions

Le vecteur des rendements est $Y = \{y_t\}_t, t = 1, \dots, T$. Pour un modèle donné, $V = \{v_t\}_{t=1}^{T}$ désigne le vecteur des T VaR quotidiennes couvrant les jours $t = 1, \dots, T$ à un degré de confiance $(1 - \alpha)$. Par convention, les VaR sont positives, $v_t > 0, \forall t$ alors que les rendements y_t sont négatifs (queue gauche des rendements). Dans certaines dérivations on posera $u_t = -v_t$ qui correspond à la VaR algébrique, $u_t < 0, \forall t$. On parlera de hit ou de violation quand y_t dépasse en valeur absolue la couverture v_t du jour t. On définit : la fonction indicatrice $I_t = 1_{y_t < -v_t}$; T_1: le nombre de hits; T_0: le nombre de non-hits, $T_0 = T - T_1$. La fréquence observée des violations est notée $\pi; \pi = P(I_t = 1) = \frac{T_1}{T} = \frac{1}{T}\Sigma_t I_t \Rightarrow \pi = E(I_t)$.

On aura également besoin d'étudier les violations qui se suivent.

- T_{11} : nombre de hits qui se suivent :$I_t = 1$ sachant $I_{t-1} = 1$; $\pi_{11} = P(I_t = 1|I_{t-1} = 1)$; $\pi_{11} = T_{11}/(T_{01} + T_{11})$
- T_{01}: nombre de hits qui suivent un non-hit : $I_t = 1$ sachant $I_{t-1} = 0$; $\pi_{11} = P(I_t = 1|I_{t-1} = 0)$; $\pi_{01} = T_{01}/(T_{00} + T_{01})$
- T_{10} : nombre de non-hits qui suivent un hit : $I_t = 0$ sachant $I_{t-1} = 1$; $\pi_{11} = P(I_t = 0|I_{t-1} = 1)$; $\pi_{10} = T_{10}/(T_{00} + T_{10})$
- T_{00} : nombre de non-hits qui se suivent : $I_t = 0$ sachant $I_{t-1} = 0$; $\pi_{11} = P(I_t = 0|I_{t-1} = 0)$; $\pi_{00} = T_{00}/(T_{01} + T_{00})$

3. Présentation des backtests de la VaR

3.1. Backtests standards

Globalement, le modèle qui convient doit permettre d'éviter à la fois une insuffisance de couverture, comme une couverture trop grande (two-side tests). Les backtests standards travaillent directement sur le processus des hits $\{I_t\}$ en tant que suite de tirages successifs sur une distribution Bernoulli (un tirage par jour $t = 1, \dots, T$). Si ces tirages sont considérés indépendants, alors le backtest est dit à couverture inconditionnelle (Uncoditional Coverage, uc en abrégé). D'autre part, les backtests qui testent simultanément la suffisance de couverture en présence de la dépendance des hits sont appelés backtests conditionnels (Conditional Coverage, cc en abrégé). On présentera le backtest uc de Kupiec (1995), le test d'indépendance des hits, désigné par ind, et le backtest cc de Christoffersen (1998). On verra aussi le backtest cc conçu par Engle et Manganelli (2004). D'autres backtests standards sont proposés dans la littérature, dont une revue est présentée dans Berkowitz et al. (2011).

On rappelle rapidement le fonctionnement du test de rapport de vraisemblance (Likelihood Ratio test) qui est à la base des backtests standards utilisés. Soit un modèle contraint, M_c, sous une hypothèse nulle H_0, imbriqué à un modèle non contraint, M_{nc}. Si H_0 est vraie alors la statistique LR doit suivre une loi χ^2 de degré de liberté (ddl) égal au nombre de contraintes imposées par H_0 :

$$LR = -2\log\left[\frac{L(M_c)}{L(M_{nc})}\right] \sim \chi^2_{ddl=\#c}$$

où $L(.)$ est la vraisemblance du modèle concerné, ddl désigne les degrés de liberté de la loi χ^2, $\#c$ est le nombre de contraintes de l'hypothèse nulle H_0. On se souviendra que la vraisemblance du modèle contraint (non contraint) est toujours au numérateur (dénominateur). Dans la suite on fixe le seuil critique de rejet à 5 %.

3.1.1. Backtest inconditionnel *uc*

On veut comparer statistiquement la fréquence observée des hits, π, par rapport à la fréquence théorique des hits, α. L'hypothèse nulle est :

$$H_0: \pi = \alpha. \tag{3.1}$$

À remarquer que le nombre total des violations $T_1 = \Sigma_{t=1}^{T} I_t$, d'où $\pi = T_1/T = E(I_t)$. L'hypothèse nulle peut s'écrire autrement :

$$H_0: E(I_t) = \alpha. \tag{3.2}$$

On reverra l'écriture (3.2) plus loin. Pour le moment, revenons à (3.1) que Kupiec (1995) propose de tester avec la fonction de masse de la loi binomiale. En effet, la somme des succès de Bernoulli indépendants suit une binomiale. La probabilité de succès de la binomiale est $\pi = T_1/T$. La vraisemblance $L_{Binom}(\pi)$ de la binomiale est

$$L_{Binom}(\pi) = \binom{T}{T_1} (1-\pi)^{T-T_1} \pi^{T_1} \tag{3.3}$$

où le terme $\binom{T}{T_1} = \frac{T!}{T_1!(T-T_1)!}$ est le nombre de combinaisons de T_1 éléments parmi T éléments. Pour évaluer H_0 on compare les vraisemblances $L_{Binom}(\pi)$ et $L_{Binom}(\alpha)$ par ratio de vraisemblance (*Likelihood Ratio*). Si H_0 est vraie la statistique du backtest *uc*, LR_{uc}, doit suivre une χ^2 de degré de liberté ddl=1 (H_0 contient une contrainte). Le modèle contraint (non contraint, au numérateur) est celui relatif à α (π, au dénominateur). On calcule

$$LR_{uc} = -2 \times log[L_{binom}(\alpha) \,/\, L_{Binom}(\pi)] = -2 \times log\left[\left(\frac{1-\alpha}{1-\pi}\right)^{T-T_1} \left(\frac{\alpha}{\pi}\right)^{T_1}\right] \sim \chi^2_{ddl=1}. \tag{3.4}$$

Au passage, Christoffersen (1998) arrive à la même expression (3.4) avec un autre raisonnement. Il écrit directement la vraisemblance de la série des hits $\{I_t\}$ comme un processus de tirages Bernoulli indépendants. D'où :

$$L_{Bernoulli}(\pi) = Pr(I_1, I_2, \dots, I_T) = \prod_{t=1}^{T}(1-\pi)^{1-I_t}\,\pi^{I_t} = (1-\pi)^{T-T_1}\pi^{T_1}. \qquad (3.5)$$

La différence entre les équations (3.3) et (3.5) est que le terme $\binom{T}{T_1}$ est éliminé dans le rapport de vraisemblance. La p-value du test est obtenue en calculant $1 - p\chi^2(LR_{uc}, ddl = 1)$ où $p\chi^2$ est la cumulative[4] de la loi χ^2.

3.1.2. Backtest d'indépendance des hits

On veut évaluer statistiquement l'indépendance des hits. En effet, advenant un hit à une date t, la banque doit avoir du temps pour reconstituer son capital de couverture. Si deux hits ou plus se suivent trop souvent il y aurait alors un problème sérieux. L'idéal est d'avoir des hits espacés et bien répartis sur toute la période $t = 1, \dots, T$. Statistiquement, il y a indépendance des hits si ces derniers sont répartis i.i.d. ou encore, si la probabilité d'avoir un hit à un jour t ne dépend pas du passé. Si on considère que le passé est incorporé dans l'état de la veille et si on conditionne par rapport à la veille, les probabilités doivent être égales $P(I_t = 1|I_{t-1} = 1) = P(I_t = 1|I_{t-1} = 0,)$, ce qui est finalement égal à la probabilité inconditionnelle $P(I_t = 1)$. Cela donne l'hypothèse nulle

$$H_0: \pi_{11} = \pi_{01}(= \pi) \qquad (3.6)$$

Cette fois, Christoffersen (1998) considère que les hits suivent un processus Markovien d'ordre 1 (*First-Order Markov Process*). L'état hit (1) ou non-hit (0) relativement à un jour t dépend uniquement de l'état de la veille qui contient l'information du passé. Dans le cas général, la matrice de transition du processus est décrite par les deux probabilités π_{01} et π_{11} tel que :

$$\Pi_{\pi_{01},\pi_{11}} = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix}. \qquad (3.7)$$

La vraisemblance de ce processus, $L_{Markov1}$ s'écrit :

$$L_{Markov1}\left(\Pi_{\pi_{01},\pi_{11}}\right) = (1-\pi_{01})^{T_{00}}\pi_{01}^{T_{01}}(1-\pi_{11})^{T_{10}}\pi_{11}^{T_{11}}. \qquad (3.8)$$

Si indépendance, alors $\pi_{01} = \pi_{11} = \pi$ et l'expression (3.8) devient :

---

[4] Dans Excel, la fonction de probabilité de la χ^2 donne directement $1 - p\chi^2(.)$. La formule à utiliser est *p-value(uc) = LOI.KHIDEUX.DROITE(LR_{uc}, ddl=1).*

$$
\begin{aligned}
L_{Markov1}\left(\Pi_{\pi_{01}=\pi_{11}=\pi}\right) &= (1-\pi)^{T_{00}}\pi^{T_{01}}(1-\pi)^{T_{10}}\pi^{T_{11}} \\
&= (1-\pi)^{T_{00}+T_{10}}\pi^{T_{01}+T_{11}} \\
&= L_{Bernoulli}(\pi). \qquad (3.9)
\end{aligned}
$$

On rappelle que $T_{00} + T_{10} = T_0 = T - T_1$ et $T_{01} + T_{11} = T_1$ pour obtenir l'égalité (3.9). Le modèle contraint, au numérateur, est celui relatif à l'hypothèse nulle. L'expression du ratio de vraisemblance du test *ind* s'écrit :

$$
\begin{aligned}
LR_{ind} &= -2 \times log\left[\frac{L_{Markov1}\left(\Pi_{\pi_{01}=\pi_{11}=\pi}\right)}{L_{Markov1}\left(\Pi_{\pi_{01},\pi_{11}}\right)}\right] \qquad (3.10) \\
&= -2log\left[\frac{L_{Bernoulli}(\pi)}{L_{Markov1}\left(\Pi_{\pi_{01},\pi_{11}}\right)}\right] \qquad (3.11) \\
&= -2log\frac{(1-\pi)^{T_{00}+T_{10}}\pi^{T_{01}+T_{11}}}{(1-\pi_{01})^{T_{00}}\pi_{01}^{T_{01}}(1-\pi_{11})^{T_{10}}\pi_{11}^{T_{11}}}. \qquad (3.12)
\end{aligned}
$$

Dans la pratique, on calcule la statistique LR_{ind} avec les estimés de π_{01}, π_{11}et π. La p-value est obtenue en calculant $1 - p\chi^2(LR_{ind}, ddl = 1)$ où $p\chi^2$ est la cumulative de la loi χ^2.

À noter, enfin, que l'hypothèse nulle de (3.6) peut être exprimée en espérances mathématiques conditionnelles : $E_{t-1}(I_t) = \Sigma_t I_t \times P(I_t|I_{t-1}) = 0 \times P(I_t = 0|I_{t-1}) + 1 \times P(I_t = 1|I_{t-1})$. D'où une deuxième façon d'écrire l'hypothèse nulle d'indépendance qu'on utilisera dans la suite :

$$
H_0: E_{t-1}(I_t) = \pi;\ \forall t. \qquad (3.13)
$$

3.1.3. Backtests de couverture conditionnelle

L'objectif est de pouvoir tester simultanément les deux hypothèses nulles de suffisance de couverture plus l'indépendance des hits. On conduira ce type de backtest à l'aide de deux techniques, Christoffersen (1998) et Engle et Manganelli (2004) dont on utilisera les abréviations cc et DQ, respectivement.

Test cc de Christoffersen (1998)

Le travail est presque déjà fait pour *cc* de Christoffersen (1998). La fusion des hypothèses (3.1) et (3.6) donne :

$$
H_0: \pi_{11} = \pi_{01} = \alpha = (\pi).
$$

L'expression analytique du ratio de vraisemblance de ce test est :

$$LR_{cc} = -2 \times log\left[\frac{L_{Markov1}\left(\Pi_{\pi_{01}=\pi_{11}=\alpha}\right)}{L_{Markov1}\left(\Pi_{\pi_{01},\pi_{11}}\right)}\right] = -2 \times log\left[\frac{L_{Bernoulli}(\alpha)}{L_{Markov1}\left(\Pi_{\pi_{01},\pi_{11}}\right)}\right] \quad (3.14)$$

$$= -2 \times log\left[\frac{L_{Bernoulli}(\alpha)}{L_{Bernoulli}(\pi)} \times \frac{L_{Bernoulli}(\pi)}{L_{Markov1}\left(\Pi_{\pi_{01},\pi_{11}}\right)}\right]$$

$$= -2 \times log\left[\frac{L_{Bernoulli}(\alpha)}{L_{Bernoulli}(\pi)}\right] - 2 \times log\left[\frac{L_{Bernoulli}(\pi)}{L_{Markov1}\left(\Pi_{\pi_{01},\pi_{11}}\right)}\right] \quad (3.15)$$

$$= LR_{uc} + LR_{ind} \sim \chi^2_{ddl=2}. \quad (3.16)$$

En décomposant le ratio de vraisemblance (3.14), on aboutit à l'égalité (3.16) qui simplifie le travail. La statistique de *cc* est la somme de celles de *uc* et *ind*. D'autre part, on a bien deux contraintes simultanées pour ce test, à savoir $\pi_{01} = \pi_{11}$ et $\pi = \alpha$.

Dans la pratique, le backtest d'indépendance des hits est peu utilisé seul. Il est plutôt implicitement inclus dans *cc* de Christoffersen (1998).

Test DQ de Engle et Manganelli (2004)

Le test simultané d'hypothèse nulle de suffisance de couverture (3.2) et de celle d'indépendance des hits (3.13) revient à tester

$$H_0 : E_{t-1}(I_t) = \alpha;\ \forall t \quad (3.17)$$

En posant $\gamma_t = I_t - \alpha$, la variable aléatoire γ_t satisfait $E[\gamma_t \mid \mathcal{F}_{t-1}] = 0$ quand H_0 est vraie[5]. Si on considère que toute l'information jusqu'à la veille, $\mathcal{F}_{t-1}$, est incluse dans γ_{t-1} alors les coefficients de la régression suivante doivent être simultanément nuls :

$$\gamma_t = \xi_0 + \xi_1 \gamma_{t-1} + \xi_2 v_t + \epsilon_t$$

où $\Theta = (\xi_0, \xi_1, \xi_2)$ est le vecteur des coefficients de la régression et ϵ_t est le terme d'erreur. Engle et Manganelli (2004) montrent que la statistique DQ devrait suivre une χ^2 avec ddl = 3 (trois contraintes = trois coefficients simultanément nuls) :

$$DQ = \frac{H'X(X'X)^{-1}X'H}{\alpha(1-\alpha)} \sim \chi^2(ddl = 3).$$

[5] $E[\gamma_t \mid \mathcal{F}_{t-1}] = E[I_t - \alpha \mid \mathcal{F}_{t-1}] = E[I_t \mid \mathcal{F}_{t-1}] - \alpha = 0.$

Tableau 2: Résultats des backtests

			—————— Backtests standards ——————-								—— Backtests comparatifs — tick loss —			
	Modèles		—— uc ——		—— ind ——		—— cc ——		—— DQ ——		— tick score —		— log score —	
α		Hits	Stat.	*p*-value	Stat.	*p*-value	Stat.	*p*-value	Stat.	*p*-value	Score	Rank	Score	Rank
en %		en %	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
1	SimHist	1,292	2,1303	0,1444	2,9933	0,0836	5,1235	0,0772	3,3670	0,3384	0,033969	3	-0,034245	3
1	RMetrics	1,550	7,0863	0,0078	1,8806	0,1703	8,9669	0,0113	13,4720	0,0037	0,033311	1	-0,034457	1
1	NO	1,697	11,0114	0,0009	7,1025	0,0077	18,1138	0,0001	13,2963	0,0040	0,035019	6	-0,033513	6
1	TF	1,402	3,9363	0,0473	2,4725	0,1159	6,4089	0,0406	10,4507	0,0151	0,034621	5	-0,033919	5
1	SEP3	1,365	3,2790	0,0702	2,6383	0,1043	5,9173	0,0519	7,0776	0,0695	0,034145	4	-0,034067	4
1	ST4	1,107	0,3030	0,5820	4,0370	0,0445	4,3399	0,1142	3,2881	0,3493	0,033890	2	-0,034270	2
2,5	SimHist	2,989	2,5038	0,1136	8,7855	0,0030	11,2892	0,0035	8,7815	0,0323	0,068986	6	-0,091019	4
2,5	RMetrics	3,247	5,6818	0,0171	1,4187	0,2336	7,1004	0,0287	11,5222	0,0092	0,067445	1	-0,091865	1
2,5	NO	2,989	2,5038	0,1136	6,2550	0,0124	8,7588	0,0125	8,5335	0,0362	0,068828	4	-0,090980	6
2,5	TF	2,878	1,5178	0,2180	4,6677	0,0307	6,1855	0,0454	8,2295	0,0415	0,068888	5	-0,091005	5
2,5	SEP3	2,694	0,4071	0,5234	5,7896	0,0161	6,1967	0,0451	5,8572	0,1188	0,068294	3	-0,091257	3
2,5	ST4	2,768	0,7695	0,3804	5,3222	0,0211	6,0917	0,0476	5,9695	0,1131	0,068259	2	-0,091333	2
5	SimHist	5,314	0,5505	0,4581	2,3875	0,1223	2,9380	0,2302	5,4109	0,1441	0,115357	6	-0,191705	6
5	RMetrics	4,945	0,0175	0,8946	1,6661	0,1968	1,6836	0,4309	9,0565	0,0285	0,112988	1	-0,193417	1
5	NO	5,055	0,0174	0,8950	2,2911	0,1301	2,3085	0,3153	6,6512	0,0839	0,114824	4	-0,192220	4
5	TF	5,683	2,5514	0,1102	0,6376	0,4246	3,1890	0,2030	8,3292	0,0397	0,115294	5	-0,191910	5
5	SEP3	4,982	0,0019	0,9648	2,5674	0,1091	2,5693	0,2767	5,7326	0,1254	0,114375	2	-0,192351	2
5	ST4	5,498	1,3734	0,2412	2,7280	0,0986	4,1013	0,1286	5,3569	0,1475	0,114664	3	-0,192238	3

Les valeurs des scores de la *tick loss* sont en %.

où H est le vecteur contenant la série des hits. Dans la présente note, on utilise plutôt le test de Wald pour évaluer la probabilité que les trois coefficients soient simultanément nuls. L'intérêt est qu'on estimera en HAC (*Heteroskedasticity and Autocorrelation Consistent*) la matrice variance-covariance utilisée dans le test de Wald. Cela contribue grandement à mieux tenir compte des autocorrélations très fortes dans les variables de cette régression. On reviendra sur ce point et sur le backtest *DQ* vers la fin du développement de la partie du backtesting comparatif.

3.2. Résultats des backtests standards

Le seuil critique de rejet est fixé à 5 % pour tous les tests statistiques. Les abréviations *uc*, ind, *cc* et *DQ* désignent respectivement les backtests inconditionnel, d'indépendance des hits, le conditionnel de Christoffersen (1998) et enfin celui de Engle et Manganelli (2004). D'autre part, on analyse les performances des modèles dans l'ordre décroissant du degré de confiance des VaR. D'abord, $\alpha = 1\,\%$ est le niveau prioritaire exigé par la réglementation. Les modèles rejetés à ce niveau seront éliminés. Ensuite, le niveau $\alpha =$ 2,5 % acquiert une deuxième importance d'autant plus plus que toute erreur sur la VaR à 2,5 % peut affecter le calcul de la couverture qui correspond à la CVaR à 2,5 %[6]. Enfin, le backtest portera aussi sur le niveau $\alpha = 5\,\%$ à titre complémentaire. L'objectif est de s'assurer davantage de l'adéquation des modèles à ce niveau traditionnel de la queue de distribution.

Quant aux résultats des backtests, la priorité est donnée à *uc* vu que pour le moment la réglementation de Bâle exige explicitement un backtest statuant sur la fréquence des hits. Ensuite, dans cet ordre, les résultats des backtests *cc* et *DQ* puis ceux de *ind* complèteront la décision finale quant au choix de modèle de VaR.

Le Tableau 2 affiche les résultats. À $\alpha = 1\%$, trois modèles sur six sont rejetés par le backtest *uc* : NO, RMetrics et TF. On peut s'attendre au rejet de NO, mais pour TF son rejet est surprenant malgré l'épaisseur de queue de sa distribution. Cela étant, sa *p*-value 0,0473 n'est pas loin du seuil critique 5 %. Les modèles RMetrics et NO ont clairement les résultats

[6] Depuis janvier 2019, le capital de couverture du risque de marché est défini selon la CVaR à 2,5 %; le backtesting minimal exigé reste essentiellement basé sur la VaR à 1 % (BCBS, 2016, 2019).

les plus médiocres ne passant pas, non plus, ni *cc* (0,0113 et 0,0001) ni *DQ* (0,0037 et 0,004). Les trois modèles non rejetés ST4, SEP3 et SimHist passent *uc* avec des *p*-values respectives de 0,582, 0,0702 et 0,1444 toutes > 5 %. Les backtests conditionnels *cc* et *DQ* valident les trois modèles (*p*-values respectives de *cc :* 0,1142, 0,0519 et 0,0772; celles de *DQ :* 0,3493, 0,0695 et 0,3384, toutes > 5 %). À noter que ST4 présente une certaine faiblesse quant au backtest *ind* dont la *p*-value = 0,0445, qui finalement n'est pas loin de 5 %. C'est sans conséquence sur la décision du backtest *cc* qui valide convenablement ST4.

Pour $\alpha = 2{,}5\%$, seul RMetrics est rejeté par *uc* avec *p*-value = 0,0171 clairement < 5 %. Quant à *cc*, les modèles ST4, SEP3 et TF ne sont pas loin du seuil critique (*p*-values 0,0476, 0,0451 et 0,0454 $\approx$ 5 %). Ceci semble indiquer la présence d'un probable problème de clustering de hautes pertes qui provoquent des violations en groupe pour tous les modèles pour $\alpha = 2{,}5\%$. Les *p*-values de *DQ* valident malgré tout ST4 et SEP3 (0,0476 et 0,0451 > 5 %). Rien de spécial concernant $\alpha = 5$ %. Aucun modèle n'est rejeté par les backtests *uc*, *ind, cc* et *DQ*, sauf RMetrics, encore lui, clairement rejeté par *DQ* (*p*-value 0,0285 < 5 %).

En résumé, le plus mauvais modèle serait RMetrics étant sévèrement rejeté par *uc* aux niveaux $\alpha = 1$ % et 2,5 %. Le modèle ST4 serait clairement un premier choix pour couvrir la VaR des données de l'échantillon. Puis, un deuxième choix serait SEP3. A noter que SimHist performe plutôt bien également, si bien qu'il pourrait être un deuxième choix plutôt que SEP3, sauf que SimHist est le modèle qui a la plus petite *p*-value du backtest cc à $\alpha = 2{,}5$ % (*p*-value $= 0{,}0035 \ll 0{,}01$) tout en étant rejeté par DQ (*p*-value 0,0323 < 5 %). L'un des problèmes de la simulation historique est qu'elle suppose la stationnarité temporelle des données, condition qui n'est pas facile à garantir, surtout pendant les périodes de turbulence extrême des marchés, comme c'est le cas pour le présent échantillon. Malgré cela, et comme déjà mentionné, la simulation historique, bien que très rudimentaire, serait utilisée par 75 % des banques américaines selon Mehta et al. (2012). Plus encore et plus récemment, elle est utilisée pour le backtesting réglementaire sur la VaR 1 % par 72 % des banques européennes selon une enquête dirigée en 2020 par la EBA (European Banking Authority) selon Woodall (2021). D'autre part, la simulation historique, telle quelle ou avec des variantes, est souvent employée dans la littérature pour donner un repère (benchmark) afin de baliser les performances des modèles concurrents d'intérêt.

C'est une question qui se pose pour le backtesting comparatif. On passe, justement, à la partie suivante qui traite du backtesting comparatif.

4. Backtesting comparatif de la VaR

4.1. Concept de fonction de score

Intuitivement, une fonction de score, S, permet d'attribuer à une statistique X une pénalité $S(X)$ selon l'erreur de l'estimé de X par rapport à sa "vraie" valeur. Concrètement, cet outil permet d'évaluer et de classer en ex-post des modèles concurrents quant à leur pouvoir prédictif (*forecasting*) déterminé en ex-ante. La popularité grandissante des fonctions de score est motivée par certaines lacunes des backtests standards, dont le manque de puissance des tests dans certains cas (voir par exemple Berkowitz et al., 2011). Aussi, les backtests standards se limitent à une conclusion binaire, rejet ou non-rejet. Ils n'offrent pas en standard la possibilité d'ordonner les modèles concurrents, des plus pertinents pour les données jusqu'aux moins performants. Les fonctions de score le peuvent sous certaines conditions.

La notion de consistance stricte est fondamentale pour établir un classement des modèles concurrents. Les fonctions de score strictement consistantes pour la VaR peuvent prendre une formulation générale telle que celle décrite dans Gneiting et Raftery (2007, éq. (40)) :

$$S(v_t) = (I_t - \alpha) \times G(-v_t) - I_t \times G(y_t) + h(y_t) \qquad (4.1)$$

où $I_t = 1_{y_t<-v_t}$, G et h sont des fonctions. La fonction h doit être intégrable dans son domaine de définition et G doit être continue, différentiable et strictement croissante. À noter qu'en toute rigueur, S est une fonction à deux variables $S(v_t, y_t)$. Cela étant, ce qui varie d'un modèle à un autre est le vecteur $\{v_t\}$. On écrit alors S en fonction de la variable v_t seule pour simplification, sans risque de confusion. Le score du jour t est $S(v_t)$, v_t étant la couverture VaR du jour. La moyenne $E[S(v_t)]$ est le score du modèle en question sur toute la période $t = 1, \dots, T$. En quelques mots, si la fonction score S est strictement consistante alors $E[S(v_t)]$ possède un et un seul minimum qui est global. Ce minimum absolu correspond au score d'un modèle optimal idéal (a priori inconnu). Les scores sont des pénalités : plus un modèle s'éloigne du modèle optimal plus grand est son score.

Autrement dit, moins le score est grand meilleur est le modèle et vice-versa. C'est cette fonctionnalité qui permet de classer les modèles concurrents.

Depuis l'article de Gneiting (2011), qui montre définitivement que la CVaR n'est pas élicitable, une série de travaux dont Fissler, Ziegel et Gneiting (2016), Fissler et Ziegel (2016) et Nolde et Ziegel (2017) reprennent et établissent les bases théoriques récentes du backtesting comparatif par fonction de score consistante. D'autre part, il existe deux formulations fréquemment utilisées pour la VaR qui sont, en fait, des cas particuliers de l'expression (4.1). Les voici :

$$S(v_t) = (I_t - \alpha) \times \big(G(-v_t) - G(y_t)\big) \tag{4.2}$$

$$S(v_t) = (I_t - \alpha) \times G(-v_t) - I_t \times G(y_t). \tag{4.3}$$

Pour la première expression, (4.2), voir, par exemple, Fissler, Ziegel et Gneiting (2016, éq. (1)). Concernant la deuxième expression, (4.3), voir Nolde et Ziegel (2017, éq. (2.2)). Dans l'annexe A.2 on démontre l'équivalence des trois formulations (4.1), (4.2) et (4.3). Pour tous nos calculs de score, nous allons utiliser la forme (4.3).

Dans la suite, nous allons explorer le backtesting comparatif par deux fonctions de score : *tick loss*[7] où $G(x) = x$ et une fonction de score basée sur le log naturel qu'on nommera *log loss* où $G(x) = -log(-x)$. Ces deux fonctions sont strictement croissantes. En effet, dans la queue gauche $x < 0$, les dérivées respectives de ces fonctions sont $G'(x) = 1 > 0$ et $G'(x) = -1/x > 0$. Les fonctions score S_{tick} et S_{log} sont donc strictement consistantes (voir Nolde et Ziegel, 2017, page 1851, éq. (2.19) et (2.20)).

Pour un jour t, si $y_t < -v_t \Rightarrow I_t = 1$ (il y a hit) et $G(y_t) < G(-v_t)$ (G strictement croissante). Dans ce cas $S(v_t) = (1-\alpha)G(-v_t) - G(y_t) \approx G(-v_t) - G(y_t) > 0$, si on néglige α devant 1. À l'opposé, s'il n'y a pas hit au jour t alors $I_t = 0$ et $-v_t < y_t \Rightarrow G(-v_t) < G(y_t)$, d'où $S(v_t) = -\alpha \times G(-v_t)$ qui est du signe inverse de $G(-v_t)$. Il y a une asymétrie importante du score d'un hit par rapport à celui d'un non-hit. On va prendre quelques exemples chiffrés afin d'en visualiser l'ordre de grandeur. On suppose les

[7] La *tick loss* porte plusieurs noms dans la littérature comme *check, linlin, hinge, pinball loss*; elle est traditionnellement formulée par l'équation (A.3) où $G(x) = x$, ce qui est équivalent à l'équation (A.4) comme montré dans l'annexe A.2.

rendements de trois jours $t = 1\,à\,3$ tels que $y_1 = -0{,}026$, $y_2 = -0{,}011$ et $y_3 = -0{,}035$. La VaR de couverture quodienne est supposée la même pour $v_{1\,à\,3} = 0{,}025$. Il y a hit à $t = 1$: $S_{tick}^{t=1} = (1 - 0{,}01) \times -0{,}025 + 0{,}026 = 0{,}00125$, non-hit à $t = 2$: $S_{tick}^{t=2} = -\alpha G(-v_t) = -0{,}01 \times -0{,}025 = 0{,}00025$. L'écart relatif dû au hit par rapport au non-hit est $\Delta S/S = \left|S_{tick}^{t=1} - S_{tick}^{t=2}\right| / S_{tick}^{t=2} = 4$, ce qui confirme une asymétrie importante défavorisant les hits. Cela dit, avec la fonction de score S_{tick} les jours non-hit contribuent avec un score positif, donc une pénalité, ce qui demeure contre intuitif même si la pénalité est moindre. Au jour $t = 3$ il y a dépassement de la v_t plus important qu'au jour $t = 1$. On calcule $S_{tick}^{t=3} = 0{,}01025$, ce qui correspond à un écart relatif de $(0{,}01025 - 0{,}00025)/0{,}00025 = 40$. Les pénalités deviennent de plus en plus accentuées à mesure que y_t s'éloigne de la couverture du jour.

On refait les mêmes calculs pour la log loss: $S_{\log}^{t=1} = (1 - 0{,}01)(-\log(0{,}025)) + \log(0{,}026) = 0{,}00233$, $S_{\log}^{t=2} = -\alpha G(-v_t) = -0{,}03689$ et $S_{\log}^{t=3} = (1 - 0{,}01)(-\log(0{,}025)) + \log(0{,}035) = 0{,}29958$. Les écarts relatifs des jours $t = 1$ et 3 sont respectivement 1.1 et 9.1. Ces écarts semblent moins importants que ceux avec S_{tick}. Par contre, les contributions des jours non-hits sont plus proches de l'intuition puisqu'elles sont négatives avec $S_{\log}$.

Il reste à mentionner deux remarques importantes. La première est que les scores S_{tick} des modèles concurrents vont être tous positifs. D'autre part, quand $\alpha = 1\,\%$, par exemple, environ 99 % des observations ne dépassent pas les VaR quotidiennes et donc les scores des modèles, $E[S(v_t)]$ proviennent des 99 % nonhits. Par conséquent, l'ordre de grandeur des scores $E[S(v_t)]$ est obtenu en calculant $(1 - \alpha)$ fois la moyenne $E\left[G\left(-\alpha G(-v_t)\right)\right]$. L'ordre de grandeur de S_{tick} à $\alpha = 1\,\%$ est $1\,\% \times 0{,}025 = 0{,}00025$, ce qui est très petit. On va multiplier par 100 les scores S_{tick} pour éviter de manipuler de très petits chiffres. Concernant $S_{\log}$ l'ordre de grandeur est $\approx (1 - \alpha)\alpha \times E[-\log(0{,}025)] = -0{,}0365$. Les scores S_{log} seront négatifs. Pour mettre les scores des deux fonctions S_{tick} et S_{log} à un même niveau on va multiplier par 100 les scores S_{tick}. Enfin, les scores des modèles concurrents vont être proches les uns des autres. D'ailleurs, dans la littérature, il est souvent nécessaire de garder plus de 3 décimales pour différencier les modèles. À la

lumière de cette discussion, on est maintenant prêt à examiner les résultats des scores des modèles.

Les 4 dernières colonnes du Tableau 2 montrent les calculs relatifs aux deux fonctions de score. Les colonnes (12) et (13) donnent les scores et le classement de S_{tick}, les colonnes (14) et (15) montrent les scores et les classements de $S_{\log}$. Les valeurs des scores de S_{tick} sont affichées en % dans la colonne (12), car très petites[8], comme prévu. Par contre, on discutera les scores de S_{tick} sans ajouter explicitement le signe % afin de simplifier le texte. Contrairement aux valeurs des statistiques des backtests standards (colonnes (4), (6), (8) et (10)), les valeurs des scores sont plus proches les unes des autres pour chaque niveau de α. D'une façon globale, le classement des deux fonctions est presque le même. L'exception est que pour $\alpha = 2{,}5\%$ les modèles SimHist et NO s'alternent la 4e ou 6e place selon la fonction. Techniquement, il est normal que les classements puissent différer, car les fonctions n'ont pas le même rapport relatif entre les scores des hits versus ceux des non-hits, comme discuté plus haut.

Pour $\alpha = 1$ %, le classement des modèles déterminé par les deux fonctions, du premier au dernier, est : RMetrics, ST4, SimHist, SEP3, TF, NO. La dernière place n'est pas surprenante pour NO (pénalités $S_{tick} = 0{,}035019$ et $S_{log} = -0{,}033513$). Ce qui est frappant est que RMetrics serait le meilleur modèle avec le plus petit score $S_{\text{tick}} = 0{,}033311$ et $S_{\log} = -0{,}034457$. Ceci contredit les résultats des backtests standards, surtout que uc rejette sévèrement ce modèle avec une p-value très petite aussi bien pour le niveau de première importance $\alpha = 1$ % (p-value $= 0{,}0078$) que pour celui de deuxième importance $\alpha = 2{,}5\%$ (p-value $= 0{,}0171$). Hormis RMetrics, les cinq autres modèles ont en gros un classement cohérent avec le backtest uc.

Le fait que le classement des deux fonctions est essentiellement le même ne constitue pas une confirmation de la performance des modèles. De plus, un modèle classé premier aux trois niveaux de $\alpha = 1$ %, 2,5 % et 5 % n'est pas une preuve non plus de la supériorité du modèle en question. RMetrics est un excellent contre-exemple dans ce sens. En fait, une

[8] Pour mieux ressortir les valeurs des scores de la *tick loss*, les auteurs les affichent en % comme Taylor (2019); d'autres auteurs les divisent par α avant de les afficher comme dans Nolde et Ziegel (2017).

partie du problème des scores est que ce sont des moyennes des vecteurs $\{S(v_t)\}$. Des caractéristiques importantes de ces vecteurs, dont la dispersion et les corrélations, ne sont pas prises en compte. Cela peut influencer beaucoup les résultats. Pour commencer à résoudre l'énigme du modèle RMetrics on fait appel au test Diebold et Mariano (1995), désigné par DM dans la suite.

4.2. Test Diebold et Mariano (1995)

On présente dans cette section le test DM tel qu'il est conçu en théorie. On verra dans la section suivante les problèmes qui se posent pour transposer ce test DM au backtesting comparatif tel que pratiqué actuellement dans la littérature.

En théorie, le test DM ne compare pas des modèles entre eux, mais des vecteurs de prédictions, $\{u_t\}$, estimées par un moyen quelconque[9] pour une entité aléatoire X. Plus précisément, le test compare les erreurs des prédictions, $\{u_t\}$, par rapport aux réalisations ex-post, $\{x_t\}$, de l'entité aléatoire X. La mesure d'erreur se fait à l'aide d'une fonction, $g: \mathbb{R}^2 \to \mathbb{R}, (u_t, x_t) \mapsto g(u_t, x_t)$. La fonction g peut prendre l'appellation de *loss function* sans que cela ait forcément un lien avec les fonctions de score de VaR ou de CVaR. Nous l'appellerons mesure d'erreur pour éviter une confusion. Comme mesure d'erreur, la fonction g peut prendre des formes diverses pour autant qu'elle soit positive et strictement croissante en respect à l'écart $|u_t - x_t|$. Le cas le plus simple est la valeur absolue des erreurs, $g(u_t, x_t) = |x_t - u_t|$ (*Absolute Errors*) et la mesure quadratique des erreurs, $g(u_t, x_t) = (x_t - u_t)^2$ (*Squared Errors*). Le test fonctionne par paire de modèles, M_i et M_j. L'hypothèse nulle du test DM comparant les prédictions $\{u_t^{M_i}\}$ versus $\{u_t^{M_j}\}$ s'écrit :

$$H_0: E\left[g\left(u_t^{M_i}, x_t\right) - g\left(u_t^{M_j}, x_t\right)\right] = 0. \qquad (4.4)$$

La statistique DM est déterminée par la moyenne du vecteur $D_{ij} = \{g\left(u_t^{M_i}, x_t\right) - g\left(u_t^{M_j}, x_t\right)\}$, normalisée par son écart type estimé avec correction HAC (*Heteroskedasticity and Autocorrelation Consistent*). On fait également appel à la correction de Harvey et al.

[9] La distinction entre les modèles et les prédictions est importante dans le sens que les prédictions qui peuvent être testées peuvent provenir de modèles, certes, mais aussi de toutes sources dont les enquêtes, les algorithmes, voire par postulat ou tout autre moyen.

(1997) moyennant un lag $h = 5$ pour tenir compte de la forte persistance dans les autocorrélations (h est le lag à partir duquel les autocorrélations sont supposées négligeables). Avec cette correction la statistique DM suit sous H_0 une t-Student de degré de liberté $T - 1$, T étant la taille de l'échantillon. Bien entendu, quand les prédictions proviennent de modèles, on peut utiliser un abus de langage courant parlant de meilleur modèle plutôt que de meilleure prédiction. On conclura que le modèle M_i est meilleur (moins performant) que M_j quand la statistique du test est significativement négative (positive) au seuil critique fixé. Si la statistique n'est pas significative alors les deux modèles sont jugés statistiquement comparables (non distinguables).

Le vecteur $D_{ij} = \{g(u_t^{M_i}, x_t) - g\left(u_t^{M_j}, x_t\right)\}$ prend le nom de *loss differential* ou *loss differences,* étant la différence de deux vecteurs d'erreur. Chacun est censé minimiser l'erreur, $g(u_t, x_t)$ dont la moyenne devrait être la plus petite, voire tendre vers 0, pour le meilleur des deux modèles comparés.

4.3. Critique majeure de l'emploi des fonctions de score en tant que *loss differentials* du test DM

Dans l'optique de transposer le test DM aux fonctions de score il y a trois problèmes sérieux qui se posent. Le premier est que, déjà, les valeurs ex-post y_t, ne sont évidemment pas des réalisations des prédictions quotidiennes de la VaR $\{u_t\}$. Les prédictions de VaR ne cherchent pas à égaler mais à couvrir les réalisations des rendements $\{y_t\}$. Le deuxième est que, par construction, une fonction de score n'évalue pas une erreur entre des prédictions $\{u_t\}$ avec une quelconque réalisation en ex-post. L'objectif d'une fonction de score est plutôt d'attribuer un score et le meilleur modèle est celui ayant le moindre score. Une fonction de score strictement consistante tend vers un minimum absolu, un score idéal $S^* = \min_u E[S(u_t, y_t)]$ qui n'est pas une erreur ni forcément nul. Un troisième problème est que les fonctions de score peuvent ne pas être positives. La fonction *log loss* est négative pour la VaR. De plus, il ne sera pas utile d'ajouter une constante pour rendre les scores positifs, car cette constante s'annulerait immédiatement dans l'expression du *loss differential*. En somme, les fonctions de score de VaR (et de CVaR) n'ont quasiment de commun que le nom de *loss functions* avec la fonction de mesure d'erreur g, tel que le test

DM l'exige en théorie. Par conséquent, malgré son utilisation devenue standard dans la littérature du backtesting comparatif, les fonctions de score ne peuvent pas et ne devraient pas être utilisées telles qu'elles sont, c'est-à-dire conçues comme *loss differential* $\{S(v_t^{M_i})\} - \{S\left(v_t^{M_j}\right)\}$ du test DM. Voyons au Tableau 3 un petit exemple, simple mais éloquent, pour documenter clairement ce fait.

Le degré de confiance des VaR de l'exemple du Tableau 3 est fixé à 90 % ($\alpha = 10\ \%$). Les données de l'exemple sont constituées d'une série de rendements, y_t, de 10 jours consécutifs (colonne (2)). Trois modèles M_1, M_2 et M_3 engendrent chacun un vecteur couverture de VaR quotidiennes, u_1, u_2 et u_3 (colonnes (3), (6) et (9), respectivement). Le modèle M_1, ayant un hit sur 10 jours, représente une couverture idéale de VaR 10 %. En revanche, M_3 a cinq hits, d'où une importante fréquence des violations de 50 %, ce qui fait que ce modèle est clairement défaillant en termes de couverture de VaR. Le modèle M_2 subit trois hits, ce qui est beaucoup, mais moins grave que M_3. Les scores sont calculés avec la *tick loss* (en %) et paraissent dans les colonnes (5), (8) et (11) respectivement. La ligne des moyennes délivre les scores finaux des modèles. On constate que $score_{M1} > score_{M2} > score_{M3}$ (0,1915 > 0,1821 > 0,1138).

À priori, on devrait normalement avoir l'inverse puisque la fonction *tick loss* est strictement consistante. M_1 devrait avoir le plus petit score, M_3 la plus grande pénalité. Dans cet exemple, M_3 est clairement un faux meilleur modèle. Ces scores classés à l'envers constituent un premier problème. Le plus étonnant est que le recours au test DM appliqué sur les fonctions de score ne va pas détecter cette anomalie. Au contraire, en comparant $M_1 - M_2$ (M_1 versus M_2), il délivre une statistique positive non significative au seuil critique de 5% stipulant que M_1 et M_2 ne sont pas distinguables (stat = 2,0339 > 0 et p-value = 0,0725 > 5 %). Plus encore, en comparant $M_1 - M_3$ la conclusion du test DM est significativement en faveur de M_3 (stat = 2,4902 > 0 et p-value = 0,0344 < 5 %). D'où un deuxième problème confirmant et documentant un dysfonctionnement majeur du test DM s'il est utilisé sur les fonctions de score, comme pratiqué durant plusieurs années dans la littérature de la VaR (et de la CVaR).

Tableau 3: Exemple de calcul du test DM

		—— Modèle M_1——			—— Modèle M_2——			—— Modèle M_3——			Fonctions d'identification			Mesures d'erreurs		
Jours	y_t	u_t	$hits_t$	$Score_t$	u_t	$hits_t$	$Score_t$	u_t	$hits_t$	$Score_t$	M_1	M_2	M_3	M_1	M_2	M_3
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
1	-0,025	-0,0250	1	0,2700	-0,0251	1	0,2610	-0,0251	1	0,2610	0,9000	0,9000	0,9000	0,9000	0,9000	0,9000
2	-0,025	-0,0255	0	0,2550	-0,0250	1	0,2600	-0,0250	1	0,2600	-0,1000	0,9000	0,9000	0,1000	0,9000	0,9000
3	-0,019	-0,0230	0	0,2300	-0,0190	1	0,2000	-0,0170	1	0,3800	-0,1000	0,9000	0,9000	0,1000	0,9000	0,9000
4	-0,008	-0,0220	0	0,2200	-0,0190	0	0,1900	-0,0080	1	0,0900	-0,1000	-0,1000	0,9000	0,1000	0,1000	0,9000
5	-0,005	-0,0210	0	0,2100	-0,0180	0	0,1800	-0,0050	0	0,0500	-0,1000	-0,1000	-0,1000	0,1000	0,1000	0,1000
6	-0,002	-0,0150	0	0,1500	-0,0150	0	0,1500	-0,0021	0	0,0210	-0,1000	-0,1000	-0,1000	0,1000	0,1000	0,1000
7	-0,001	-0,0160	0	0,1600	-0,0160	0	0,1600	-0,0011	0	0,0110	-0,1000	-0,1000	-0,1000	0,1000	0,1000	0,1000
8	-0,001	-0,0150	0	0,1500	-0,0150	0	0,1500	-0,0009	0	0,0091	-0,1000	-0,1000	-0,1000	0,1000	0,1000	0,1000
9	-0,001	-0,0140	0	0,1400	-0,0140	0	0,1400	-0,0008	0	0,0081	-0,1000	-0,1000	-0,1000	0,1000	0,1000	0,1000
10	-0,001	-0,0130	0	0,1300	-0,0130	0	0,1300	-0,0008	1	0,0480	-0,1000	-0,1000	0,9000	0,1000	0,1000	0,9000
Moyennes/tot hits			1	0,1915		3	0,1821		5	0,1138				0,1800	0,3400	0,5000
Test DM H0 : $M_1 - M_i = 0$																
Statistiques							2,0339			2,4902					-1,5000	-2,4495
p-values							0,0725			0,0344					0,1679	0,0368

Les valeurs des scores de la *tick loss* en %.

En fait, le signe de la statistique DM est le même que pour le numérateur de la statistique étant égal à $E\left[S(v_t^{M_i}) - S\left(v_t^{M_j}\right)\right] = E\left[S(v_t^{M_i})\right] - E\left[S\left(v_t^{M_j}\right)\right]$ ce qui correspond à $score(M_i) - score(M_j)$. De ce fait, en présence d'un faux meilleur modèle, le test DM ne pourrait pas renverser le classement des scores. En effet, dans le cas où la statistique est significative, la conclusion du test DM confirmerait la supériorité du faux meilleur modèle, ce qui est faux, comme pour la comparaison de M_1 versus M_3 de l'exemple présenté. Si la statistique du test DM est non significative alors la conclusion du test serait que les modèles seraient statistiquement non distinguables, ce qui revient à confirmer que le faux arrive ex aequo avec le bon modèle, comme pour M_1 versus M_2. Encore, une fois, ceci est relié à la nature des fonctions de score suffisamment différentes de celle des fonctions de mesure d'erreur telles que requises selon la théorie du test DM.

Nous reviendrons au Tableau 3 pour étudier les informations des colonnes (12) à (17) qui concernent la fonction d'identification associée à une fonction de score. Pour le moment, reprenons nos données réelles et les six modèles concurrents proposés. Nous allons retrouver une situation similaire où RMetrics se comporte aussi en tant que faux meilleur modèle, sans que ce soit détecté par le test DM, toujours pour les mêmes raisons !

Les résultats du test DM appliqué sur les scores sont présentés au Tableau 4. Dans le panel de la *tick loss,* pour $\alpha = 1$ %, à la ligne du supposé meilleur modèle, RMetrics, toute la ligne est négative (les statistiques respectives : -0,7949, -1,7047, -1,4893, -1,0546 et -0,7351 toutes < 0), ce qui est cohérent vu les scores des modèles. D'autre part, aucune des cinq statistiques de cette ligne n'est significative à 5 %. La conclusion directe dès la lecture de la ligne du supposé être meilleur modèle serait que les six modèles ne seraient pas distinguables statistiquement du point de vue de la *tick loss*; idem pour $\alpha = 2{,}5\%$, aucun des six modèles concurrents n'est distinguable des autres. Au panel de la *log loss*, à l'exception du modèle NO qui serait moins performant que RMetrics à 5 %, pour $\alpha = 1$ %, on retrouve la même conclusion : les modèles ne seraient pas distinguables entre eux pour le niveau réglementaire $\alpha = 1$ % et pour le niveau de seconde priorité $\alpha = 2{,}5\%$. Nous ne pouvons tirer aucune conclusion du backtesting comparatif ainsi construit !

La figure 2 trace, pour $\alpha = 1\ \%$, les scores quotidiens calculés pour les modèles RMetrics, SimHist, ST4 et le modèle normal NO. Le vecteur des scores de RMetrics, en noir, est le plus volatil. Il varie nettement plus vite que ceux des autres modèles selon les rendements de la queue gauche. Lors des périodes de hautes pertes (voir l'amplitude des pertes en traits gris verticaux), il est capable de réagir vite, mais sans assurer une couverture suffisante, puis de baisser au plus tôt dès la fin des turbulences. En conséquence, les scores de RMetrics sont la plupart du temps au-dessous des autres. Ce fait est encore plus clair dans la moitié droite de la figure. Or, la couverture de RMetrics est sévèrement rejetée par les backtests standards. RMetrics obtient le meilleur score uniquement parce qu'il réalise beaucoup d'économies de capital en basses pertes, au point que les mauvais scores lors des jours de violation de la VaR, bien que comptabilisés avec une forte asymétrie défavorable, n'apparaissent pas suffisamment pour affecter le score global de ce modèle.

Maintenant que nous voyons que le test DM ne fonctionne pas sur les fonctions de score, nous allons présenter une solution qui permet au test DM de retrouver les éléments conformes nécessaires à son bon fonctionnement.

Tableau 4: Résultats du test DM pour *tick loss* et *log loss*

			Modèles concurrents					
			SimHist	RMetrics	NO	TF	SEP3	ST4
α	Modèles	Rank	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)
— *tick loss*								
1	SimHist	3		0,7949	-1,2178	-0,9312	-0,3264	0,1786
1				(0,4267)	(0,2234)	(0,3518)	(0,7441)	(0,8582)
1	RMetrics	1	-0,7949		-1,7047*	-1,4893	-1,0546	-0,7351
1			(0,4267)		(0,0884)	(0,1365)	(0,2917)	(0,4624)
1	NO	6	1,2178	1,7047*		0,8644	1,7956*	1,8379*
1			(0,2234)	(0,0884)		(0,3875)	(0,0727)	(0,0662)
1	TF	5	0,9312	1,4893	-0,8644		2,0355**	
1			(0,3518)	(0,1365)	(0,3875)		(0,0419)	(0,0607)
1	SEP3	4	0,3264	1,0546	-1,7956*	-2,0355**		1,1423
1			(0,7441)	(0,2917)	(0,0727)	(0,0419)		(0,2534)
1	ST4	2	-0,1786	0,7351	-1,8379*	-1,8767*	-1,1423	
1			(0,8582)	(0,4624)	(0,0662)	(0,0607)	(0,2534)	
2,5	SimHist	6		1,1341	0,2575	0,1461	1,1944	1,6251
2,5				(0,2569)	(0,7968)	(0,8839)	(0,2324)	(0,1043)
2,5	RMetrics	1	-1,1341		-1,0226	-1,0922	-0,7023	-0,6430
2,5			(0,2569)		(0,3066)	(0,2749)	(0,4825)	(0,5203)
2,5	NO	4	-0,2575	1,0226		-0,4025	1,4722	1,9178*
2,5			(0,7968)	(0,3066)		(0,6873)	(0,1411)	(0,0552)
2,5	TF	5	-0,1461	1,0922	0,4025		1,7431*	1,8879*
2,5			(0,8839)	(0,2749)	(0,6873)		(0,0814)	(0,0591)
2,5	SEP3	3	-1,1944	0,7023	-1,4722	-1,7431*		0,1554
2,5			(0,2324)	(0,4825)	(0,1411)	(0,0814)		(0,8765)
2,5	ST4	2	-1,6251	0,6430	-1, 9178*	-1,8879*	-0,1554	
2,5			(0,1043)	(0,5203)	(0,0552)	(0,0591)	(0,8765)	
— *log score*								
1	SimHist	3		0,5471	-1,7194*	-1,1677	-0,7851	0,1452
1				(0,5843)	(0,0857)	(0,2430)	(0,4325)	(0,8846)
1	RMetrics	1	-0,5471		-1,9682**	-1,3165	-0,9846	-0,4755
1			(0,5843)		(0,0491)	(0,1881)	(0,3249)	(0,6345)
1	NO	6	1,7194*	1,9682**		1,7717*	2,1674**	2,1931**
1			(0,0857)	(0,0491)		(0,0766)	(0,0303)	(0,0284)
1	TF	5	1,1677	1,3165	-1,7717*		1,6793*	2,0341**
1			(0,2430)	(0,1881)	(0,0766)		(0,0932)	(0,0420)
1	SEP3	4	0,7851	0,9846	-2,1931**	-1,6793*		1,7847*
1			(0,4325)	(0,3249)	(0,0303)	(0,0932)		(0,0744)
1	ST4	2	-0,1452	0,4755	-2,1931**	-2,0341**	-1,7847*	
1			(0,8846)	(0,6345)	(0,0284)	(0,0420)	(0,0744)	
2,5	SimHist	4		1,3113	-0,1055	-0,0360	0,8313	1,3493

			Modèles concurrents					
			SimHist	RMetrics	NO	TF	SEP3	ST4
α	Modèles	Rank	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)
2,5				(0,1899)	(0,9160)	(0,9713)	(0,4059)	(0,1774)
2,5	RMetrics	1	-1,3113		-1,2833	-1,2852	-0,9714	-0,8538
2,5			(0,1899)		(0,1995)	(0,1988)	(0,3314)	(0,3933)
2,5	NO	6	0,1055	1,2833		0,3389	1,4066	1,8282*
2,5			(0,9160)	(0,1995)		(0,7347)	(0,1597)	(0,0676)
2,5	TF	5	0,0360	1,2852	-0,3389		1,4176	1,7968*
2,5			(0,9713)	(0,1988)	(0,7347)		(0,1564)	(0,0725)
2,5	SEP3	3	-0,8313	0,9714	-1,4066	-1,4176		0,7867
2,5			(0,4059)	(0,3314)	(0,1597)	(0,1564)		(0,4315)
2,5	ST4	2	-1,3493	0,8538	-1,8282*	-1,7968*	-0,7867	
2,5			(0,1774)	(0,3933)	(0,0676)	(0,0725)	(0,4315)	

***p<0,01 **p<0,05 *p<0,1

Les p-values sont entre parenthèses au-dessous des statistiques; les valeurs de α sont en %,

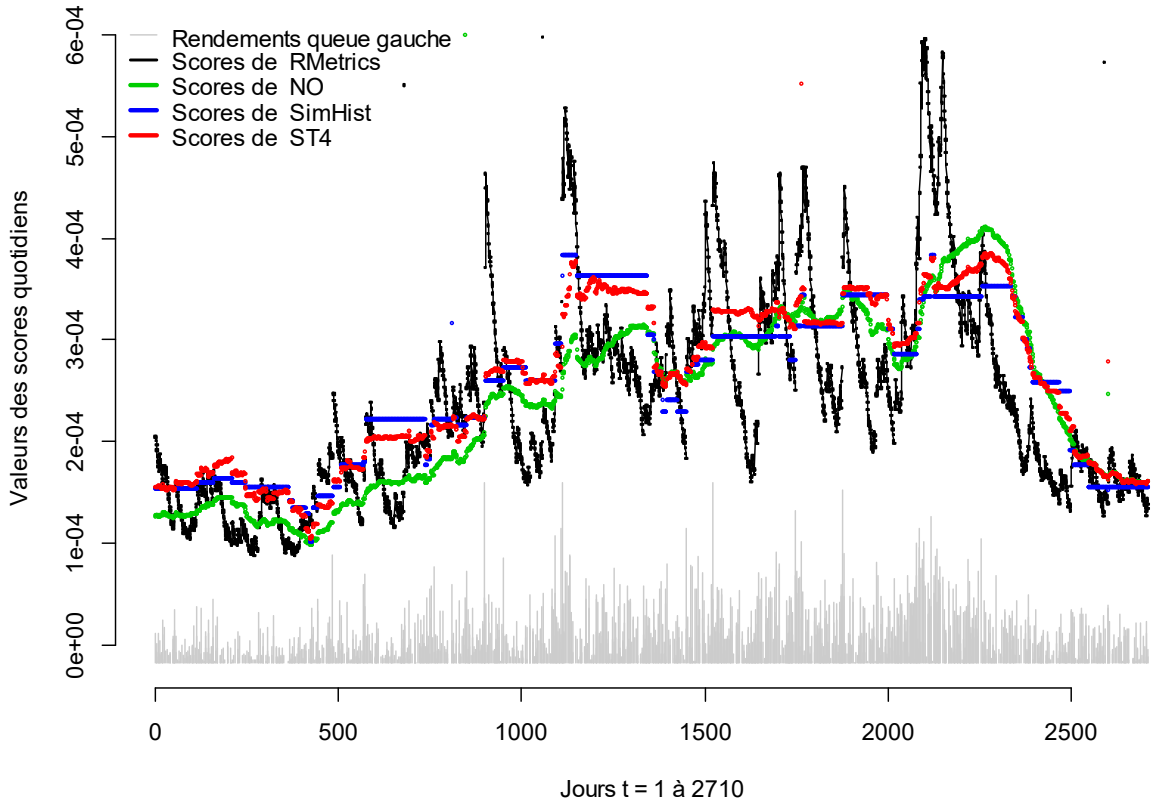

Figure 2. Tracé des scores de la *tick loss*

4.4. Fonctions d'identification associées aux fonctions de score

On sait qu'une fonction de score strictement consistante atteint un minimum absolu pour un modèle optimal. Ce point minimum est caractérisé par la condition de premier ordre qui doit y être nulle et où la couverture VaR est idéale. Cette dernière constatation est équivalente à dire que la condition de premier ordre devrait nous donner une mesure d'erreur concernant les prédictions des VaR $\{u_t\}$. C'est un premier résultat de ce que nous cherchons. Le point suivant consiste à préciser maintenant que la fonction de score est bel et bien continue et différentiable malgré la présence de la fonction discontinue I_t dans son expression, et ce, à condition que la fonction G à l'intérieur de son expression soit continue et différentiable (voir une démonstration en annexe A.3). La fonction d'identification, désormais notée $V(.)$, associée à une fonction de score strictement consistante est alors tout simplement la dérivée $\partial S(u_t)/\,\partial u_t$ où on pose $u_t = -v_t$ pour éviter une confusion au niveau des signes dans la queue gauche des rendements. On obtient :

$$V(u_t, y_t) = \frac{\partial S(u_t)}{\partial u_t} \quad = (I_t - \alpha)\frac{\partial G(u_t)}{\partial u_t} \tag{4.5}$$

où $G'(u_t) = \partial G(u_t)/\,\partial u_t$ est la dérivée première de $G(.)$ évaluée à $u_t = -v_t$. La condition de premier ordre de la minimisation de la fonction de score est :

$$E[V(u_t, y_t)] = E\left[(I_t - \alpha)\frac{\partial G(u_t)}{\partial u_t}\right] = 0. \tag{4.6}$$

À remarquer que cette dernière expression établit clairement un écart entre $I_t = 1_{y_t<u_t}$ et α et que pour un modèle optimal on a $E[I_t] = \alpha$. D'autre part, la fonction I_t lie la prédiction u_t à la réalisation ex-post y_t. La quantité $\frac{\partial G(u_t)}{\partial u_t}$ apporte une pondération dans l'espérance, laquelle pondération est cohérente avec la fonction de score primitive. À partir de l'expression (4.6), on est capable de comparer les performances d'une paire de modèles, $M_i$et M_j, à l'aide du test DM, quitte à s'assurer que les valeurs calculées par $V(u_t, y_t)$ soient positives. Nous utiliserons pour cela la fonction valeur absolue, $|V(u_t, y_t)|$ (nous aurions pu aussi choisir la forme quadratique, $V(u_t, y_t)^2$). L'hypothèse nulle du test DM appliqué à la fonction d'identification s'écrit finalement

$$E\left[\left|(I_t(M_i)-\alpha)\frac{\partial G(u_t(M_i))}{\partial u_t(M_i)}\right|-\left|\left(I_t(M_j)-\alpha\right)\frac{\partial G\left(u_t(M_j)\right)}{\partial u_t(M_j)}\right|\right]=0. \quad (4.7)$$

La formulation de l'hypothèse nulle (4.7), relative au test DM sur la fonction d'identification, est ainsi conceptuellement conforme à l'exigence test DM. À notre connaissance, nous proposons en primeur d'effectuer ainsi le test DM sur les fonctions d'identification plutôt que sur les fonctions de score elles-mêmes. On est prêt maintenant à passer à l'application numérique afin de documenter la justesse de notre démarche. Pour cela, on revient d'abord au petit exemple du Tableau 3, colonnes (12) à (16). On y voit que les valeurs de la fonction d'identification associée à *tick loss* (colonnes (12) et (13)) peuvent être positives ou négatives. L'application de la valeur absolue les rend toutes positives dans les colonnes mesurant l'erreur[10] (15) et (16). D'autre part, la statistique du test DM montrée en bas de la colonne (16) est bien négative, cette fois, et est significative au seuil 10%, avec une *p*-value égale à 0,17 (on se contente du seuil critique de 10% vu le petit nombre de jours de cet exemple illustratif). Comme attendu, la conclusion du test DM utilisant la fonction d'identification associée à la *tick loss* est en faveur du modèle M_1, détectant et corrigeant le problème du score du phénomène du faux meilleur modèle.

On passe maintenant à l'analyse des résultats du test DM sur fonctions d'identification quant à nos six modèles concurrents du Tableau 5. Une dernière surprise nous y attend : une faiblesse concernant la non 0-homogénéité de la *tick loss* dont nous mettons les conséquences en exergue pour la première fois. Voyons comment, progressivement.

La fonction d'identification associée à la *tick loss* est $V(u_t, y_t) = I_t - \alpha$ puisque dans son cas $G(x) = x \Rightarrow \partial G(x)/\partial x = 1$. L'hypothèse nulle de la minimisation de la *tick loss* est :

$$E[V(u_t, y_t)] = E[I_t] - \alpha = 0. \quad (4.8)$$

[10] Il est crucial de souligner l'importance que les mesures d'erreurs doivent être positives. On illustre ceci par un exemple simple. Si pour un jour t donné l'écart qui correspond à un modèle M_i vaut -1 et vaut $+1$ pour un modèle M_j, dans ces deux cas l'erreur résultante est 1 et les deux modèles M_i et M_j sont équivalents pour ce jour t. Dans le cas où le signe $-$ n'est pas neutralisé, l'écart $V(M_i) - V(M_j) = -1 - (+1) = -2$, ce qui serait faux, les erreurs doivent être absolues autour de 0 (erreur nulle).

Tableau 5 : Résultats du test DM utilisant la fonction d'identification de *tick loss*

			Modèles concurrents					
			SimHist	RMetrics	NO	TF	SEP3	ST4
α	Modèles	Rang	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)
1	SimHist	2		-1,4028	-2,6861***	-0,9054	-0,6328	2,2469**
1				(0,1608)	(0,0073)	(0,3653)	(0,5269)	(0,0247)
1	RMetrics	5	1,4028		-0,7076	0,8953	1,0925	2,5745**
1			(0,1608)		(0,4793)	(0,3707)	(0,2747)	(0,0101)
1	NO	6	2,6861***	0,7076		2,5454**	2,7327***	3,8225***
1			(0,0073)	(0,4793)		(0,0110)	(0,0063)	(0,0001)
1	TF	4	0,9054	-0,8953	-2,5454**		0,5776	2,5454**
1			(0,3653)	(0,3707)	(0,0110)		(0,5636)	(0,0110)
1	SEP3	3	0,6328	-1,0925	-2,7327***	-0,5776		2,3456**
1			(0,5269)	(0,2747)	(0,0063)	(0,5636)		(0,0191)
1	ST4	1	-2,2469**	-2,5745**	-3,8225***	-2,5454**	-2,3456**	
1			(0,0247)	(0,0101)	(0,0001)	(0,0110)	(0,0191)	
2,5	SimHist	5		-0,9122	0,0000	0,7281	2,3213**	1,7372*
2,5				(0,3617)	(1,0000)	(0,4666)	(0,0203)	(0,0825)
2,5	RMetrics	6	0,9122		0,8690	1,2932	1,8964*	1,6691*
2,5			(0,3617)		(0,3849)	(0,1960)	(0,0580)	(0,0952)
2,5	NO	5	0,0000	-0,8690		1,7372*	2,8501***	1,6691*
2,5			(1,0000)	(0,3849)		(0,0825)	(0,0044)	(0,0138)
2,5	TF	3	-0,7281	-1,2932	-1,7372*		1,8964*	1,3441
2,5			(0,4666)	(0,1960)	(0,0825)		(0,0580)	(0,1790)
2,5	SEP3	1	-2,3213**	-1,8964*	-2,8501***	-1,8964*		-1,4171
2,5			(0,0203)	(0,0580)	(0,0044)	(0,0580)		(0,1566)
2,5	ST4	2	-1,7372*	-1,6691*	-2,4636**	-1,3441	1,4171	
2,5			(0,0825)	(0,0952)	(0,0138)	(0,1790)	(0,1566)	
5	SimHist	4		1,0797				-1,3895
5				(0,2804)	(0,0888)	(0,0324)	(0,0284)	(0,1648)
5	RMetrics	1	-1,0797		-0,3334	-2,1407**	-0,1125	-1,6509*
5			(0,2804)		(0,7388)	(0,0324)	(0,9104)	(0,0989)
5	NO	3	-1,7026*	0,3334		-	0,8172	-3,5038***
5			(0,0888)	(0,7388)		(0,0000)	(0,4139)	(0,0005)
5	TF	6	2,1400**	2,1407**	4,1818***		4,4315***	1,2140
5			(0,0324)	(0,0324)	(0,0000)		(0,0000)	(0,2248)
5	SEP3	2	-2,1929**	0,1125	-0,8172	-4,4315***		-3,7917***
5			(0,0284)	(0,9104)	(0,4139)	(0,0000)		(0,0002)
5	ST4	5	1,3895	1,6509*	3,5038***	-1,2140	3,7917***	
5			(0,1648)	(0,0989)	(0,0005)	(0,2248)	(0,0002)	

***p<0,01 **p<0,05 *p<0,1

Les p-values sont entre parenthèses au-dessous des statistiques; les valeurs de α sont en %,

On reconnait dans (4.8) l'hypothèse nulle exacte du backtest *uc*. Maintenant, l'hypothèse nulle du test DM appliqué à la fonction d'identification de la *tick loss* est :

$$E\left[|(I_t(M_i) - \alpha)| - \left|\left(I_t\left(M_j\right) - \alpha\right)\right|\right] = 0. \qquad (4.9)$$

Le Tableau 5 donne les matrices 6×6 relatives au test DM effectué avec la fonction d'identification de S_{tick}. Le seuil critique est toujours fixé à 5%. À noter déjà que la colonne Identif Rank contient une suggestion de classement des six modèles. Ce classement est déterminé en calculant 6 moins le nombre de statistiques négatives de la ligne du modèle concerné. Le chiffre 6 est le nombre de modèles concurrents. Vu qu'il ne tient pas compte de la significativité des statistiques, il est présenté à titre suggestif afin d'aider à diriger l'analyse pour naviguer plus vite parmi les chiffres des matrices du test DM.

Nous avons enfin des résultats qui semblent avoir un peu plus de sens. En effet, la ligne ST4 est toute négative avec en prime les cinq statistiques toutes significatives à 5 % ou mieux (p-values = 0,0247, 0,0101, 0,0001, 0,011 et 0,0191 toutes < 5 %:). Ceci positionnerait ST4 en tant que meilleur modèle de VaR sans conteste pour $\alpha = 1$ %. Par contre, nous avons encore un problème. Dans la ligne SimHist, on peut voir que SimHist est en deuxième position, étant significativement inférieur à ST4, significativement supérieur à NO, non distinguable par rapport à SEP3 et à TF. De plus, SimHist ne serait pas distinguable non plus de RMetrics (stat = -1,4028 < 0 et p-value = 0,1608 > 5 %). Plus encore, RMetrics revient dans la discussion pour $\alpha = 2{,}5$ % où SEP3 serait le meilleur modèle, mais que RMetrics ne serait pas distinguable de ce premier modèle au seuil critique de 5 % (stat = -1,8964 < 0 et p-value = 0,058 > 5 %). On serait tenté de mettre le seuil critique plutôt à 10%, cela règlerait le problème pour $\alpha = 2{,}5$ %, mais pas celui relatif à $\alpha = 1$ %. En parlant de revoir le seuil critique, cela fait penser à la puissance du test DM.

Sans plus tarder, on passe au Tableau 6 présentant les matrices DM relatives à la fonction d'identification associée à la *log loss*. Voici d'abord l'hypothèse nulle correspondant à ce test, sachant que $\partial G(u_t)/\,\partial u_t = \partial\{-log(u_t)\}/\,\partial u_t = -1/u_t$:

$$E\left[\left|(I_t(M_i) - \alpha)\frac{-1}{u_t(M_i)}\right| - \left|\left(I_t\left(M_j\right) - \alpha\right)\frac{-1}{u_t(M_j)}\right|\right] = 0. \qquad (4.10)$$

Tableau 6: Résultats du test DM utilisant la fonction d'identification de *log loss*

			Modèles concurrents					
			SimHist	RMetrics	NO	TF	SEP3	ST4
α	Modèles	Rang	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)	Stat (*p*-value)
1	SimHist	2		-2,9686***	-4,1829***	-2,1017**	-1,6251	2,1967**
1				(0,0030)	(0,0000)	(0,0357)	(0,1043)	(0,0281)
1	RMetrics	5			-0,6858	1,7658*	2,1351**	3,8004***
1			(0,0030)		(0,4929)	(0,0775)	(0,0328)	(0,0001)
1	NO	6	4,1829***	0,6858		4,4202***	4,3020***	4,8450***
1			(0,0000)	(0,4929)		(0,0000)	(0,0000)	(0,0000)
1	TF	4	2,1017**	-1,7658*	-4,4202***		1,0492	3,1268***
1			(0,0357)	(0,0775)	(0,0000)		(0,2942)	(0,0018)
1	SEP3	3	1,6251	-2,1351**	-4,3020***	-1,0492		2,7303***
1			(0,1043)	(0,0328)	(0,0000)	(0,2942)		(0,0064)
1	ST4	1	-2,1967**	-3,8004***	-4,8450***	-3,1268***	-2,7303***	
1			(0,0281)	(0,0001)	(0,0000)	(0,0018)	(0,0064)	
2,5	SimHist	3		-2,1971**	-2,4168**	-1,3908	1,0301	0,3048
2,5				(0,0281)	(0,0157)	(0,1644)	(0,3030)	(0,7605)
2,5	RMetrics	6	2,1971**		0,7889	1,4411	2,5455**	2,3153**
2,5			(0,0281)		(0,4302)	(0,1497)	(0,0110)	(0,0207)
2,5	NO	5	2,4168**	-0,7889		2,8989***	4,5929***	4,1779***
2,5			(0,0157)	(0,4302)		(0,0038)	(0,0000)	(0,0000)
2,5	TF	4	1,3908	-1,4411	-2,8989***		3,4651***	2,9138***
2,5			(0,1644)	(0,1497)	(0,0038)		(0,0005)	(0,0036)
2,5	SEP3	1	-1,0301	-2,5455**	-4,5929***	-3,4651***		-1,7851*
2,5			(0,3030)	(0,0110)	(0,0000)	(0,0005)		(0,0743)
2,5	ST4	2	-0,3048	-2,3153**	-4,1779***	-2,9138***	1,7851*	
2,5			(0,7605)	(0,0207)	(0,0000)	(0,0036)	(0,0743)	
5	SimHist	4		0,8588	1,8372*	-4,0782***	2,6812***	-1,9793**
5				(0,3905)	(0,0663)	(0,0000)	(0,0074)	(0,0479)
5	RMetrics	2	-0,8588		-0,1736	-2,5797***	0,0773	-1,4898
5			(0,3905)		(0,8622)	(0,0099)	(0,9384)	(0,1364)
5	NO	3	-1,8372*	0,1736		-5,8017***	0,8219	-4,6251***
5			(0,0663)	(0,8622)		(0,0000)	(0,4112)	(0,0000)
5	TF	6	4,0782***	2,5797***	5,8017***		6,1282***	3,1291***
5			(0,0000)	(0,0099)	(0,0000)		(0,0000)	(0,0018)
5	SEP3	1	-2,6812***	-0,0773	-0,8219	-6,1282***		-5,4911***
5			(0,0074)	(0,9384)	(0,4112)	(0,0000)		(0,0000)
5	ST4	5	1,9793**	1,4898	4,6251***	-3,1291***	5,4911***	
5			(0,0479)	(0,1364)	(0,0000)	(0,0018)	(0,0000)	

***p<0,01 **p<0,05 *p<0,1

Les p-values sont entre parenthèses au-dessous des statistiques; les valeurs de α sont en %,

On examine d'abord les nouvelles valeurs des statistiques pour voir qu'à $\alpha = 1$ %, ST4 est toujours le meilleur modèle sans conteste avec toutes les statistiques de sa ligne très significativement négatives. Ensuite, dans la ligne de SimHist, classé deuxième, ce modèle est maintenant très significativement supérieur à RMetrics. La ligne de RMetrics montre qu'il est, en effet, très significativement dépassé par SimHist, ST4 et SEP3 et qu'il ne serait comparable à 5 % qu'au modèle NO (stat = -0,6858 < 0 et *p*-value = 0,4929 > 5 %) et TF (stat = 1,7658 > 0 et *p*-value = 0,0775 > 5 %). Une bonne nouvelle également du côté de $\alpha = 2{,}5\%$, le problème RMetrics ne se pose plus, non plus, sa ligne ne présente que des statistiques positives dont trois significatives à 5 % et deux non significatives montrant que RMetrics ne pourrait être désormais comparable qu'à NO (stat =0,7889 > 0 et *p*-value = 0,4302 > 5 %) et TF (stat =1,4411 > 0 et *p*-value = 0,1497 > 5 %).

Nous discutons, enfin, les résultats concernant $\alpha = 5$ %. Le modèle le plus performant est SEP3 dont la ligne est toute négative. Cette même ligne indique que SEP3 est comparable à NO et à RMetrics. Ceci est parfaitement cohérent avec les conclusions des backtests standards à $\alpha = 5$ %. On sait qu'au degré de confiance 95 % de la queue gauche, le modèle normal est souvent très performant pour modéliser la VaR. RMetrics montre une bonne performance. Au passage, dans toutes les discussions des résultats des backtests on peut remarquer que RMetrics n'est jamais bien loin du modèle normal concernant la VaR.

D'une façon globale, au Tableau 6 associé à la fonction d'identification *log loss*, on remarque qu'il y a beaucoup plus de statistiques significatives et très significatives. Les résultats sont nettement plus tranchants par rapport au tableau précédent, le Tableau 5 relatif à la fonction d'identification associée à *tick loss*. Ce grand changement vient du fait que la *log loss* est 0-homogène alors que la *tick loss* ne l'est pas (*tick loss* est plutôt 1-homogène). La puissance du test DM dépend du degré d'homogénéité de la fonction sous-jacente. La puissance de DM est maximale avec les fonctions 0-homogènes et décroît à mesure que le degré d'homogénéité augmente. Patton et Sheppard (2009) soulèvent ce fait concernant les fonctions de score mesurant les prédictions de la volatilité du marché. Nolde et Ziegel (2017) et Patton et al. (2019) recommandent le plus possible d'utiliser les 0-homogènes pour évaluer les modèles de VaR et CVaR. À l'issue de ce que nous venons de voir, nous conseillons vivement d'utiliser la fonction de score *log loss* et sa fonction d'identification vu

que le risque est réel de tomber sur un modèle ou plusieurs qui pourraient passer à travers tous les tests et les procédures en tant que bons modèles de VaR pour les données alors qu'en réalité ils ne le sont pas.

Jusqu'ici on a discuté les résultats du backtesting comparatif et vu les différents liens avec le backtest inconditionnel standard *uc.* Qu'en est-il du côté des backtests conditionnels standards et du lien pouvant exister avec le backtesting comparatif ? On pourrait penser au test de Giacomini and White (2006), mais on fera mieux de présenter un test conditionnel similaire au test de calibration conditionnel de Nolde et Ziegel (2017) et ce que Patton et al. (2019) appellent test de *Goodness-of-fit* ou encore test d'optimalité.

4.5. Test d'optimalité

Le point de départ est simplement la condition de premier ordre de la minimisation d'une fonction de score qui, comme vu dans la sous-section précédente, revient à écrire que l'espérance de la fonction d'identification associée est nulle à l'optimum. On va l'écrire, cette fois, conditionnellement aux informations connues jusqu'à la veille, $\mathcal{F}_{t-1}$, laquelle espérance conditionnelle devant être nulle pour tout $t = 1, \dots, T$:

$$E[V(u_t, y_t)|\,\mathcal{F}_{t-1}] = E\left[(I_t - \alpha)\frac{\partial G(u_t)}{\partial u_t}|\,\mathcal{F}_{t-1}\right] = 0 \ \forall t. \tag{4.11}$$

Intuitivement, cela permet de mesurer à quel point un modèle donné est ajusté pour une fonction de score donnée, et ce, non seulement en moyenne globale, mais à l'échelle quotidienne $\forall t = 1, \dots, T$. L'intérêt majeur est que cela capture l'impact des clustering des hautes pertes. Le lien avec les backtests conditionnels standards *cc et DQ* est immédiat puisque cela mesure à quel point les violations de VaR sont i.i.d. Récapitulons en disant que ce qui mesure le degré de clustering revient à mesurer le degré de *Goodness-of-Fit* d'un modèle du point de vue d'une fonction de score donnée. L'hypothèse nulle (4.11) réécrite pour la *tick loss* et pour la *log loss* donne respectivement :

$$E[(I_t - \alpha)|\,\mathcal{F}_{t-1}] = 0 \ \forall t \quad (tick\ loss) \tag{4.12}$$

$$E\left[(I_t - \alpha)\frac{1}{u_t}|\,\mathcal{F}_{t-1}\right] = 0 \ \forall t \quad (log\ loss), \tag{4.13}$$

la dérivée étant $\partial G(u_t)/\partial u_t = 1$ pour la *tick loss* et $\partial G(u_t)/\partial u_t = -1/u_t = 1/v_t$ pour *log loss*. Pour la première, la *tick loss*, on reconnait immédiatement dans (4.12) l'hypothèse nulle des backtests standards conditionnels *cc et DQ* (3.17). Si on choisit d'effectuer ce test d"optimalité (ou *Goodness-of-fit*) avec l'approche du backtest *cc* de Christoffersen (1998) ou celle *DQ* de Engle et Manganelli (2004) on aurait les mêmes résultats que ceux affichés et discutés concernant le Tableau 2, colonnes (8) et (10). À remarquer que nous aurons soulevé régulièrement des liens de similitude entre les constructions autour de la *tick loss* et les backtests standards, ce qui justifierait la popularité de cette fonction de score, à l'exception du problème de la non 0-homogénéïté de cette fonction qui pourrait poser un sérieux problème lors du test DM, comme identifié plus haut.

Pour revenir au cas général d'une fonction de score S, on crée la variable $\lambda_t = V(u_t, y_t) = (I_t - \alpha)\,\partial G(u_t)/\partial u_t$. Cette variable vérifiant $E[\lambda_t|\mathcal{F}_{t-1}] = 0$ pourrait être vue comme représentant les résidus généralisés de l'ajustement du modèle en question selon la fonction de score S, d'où l'appellation *Goodness-of-Fit* inspirée de Patton et al. (2019). En suivant l'approche par régression DQ de Engle et Manganelli (2004), on considère l'équation :

$$\lambda_t = \xi_0 + \xi_1\lambda_{t-1} + \xi_2 v_t + \epsilon_t$$

où $u_t = -v_t$, ϵ_t est le terme d'erreur de la régression, ξ_0, ξ et ξ_3 sont des coefficients à estimer. On fait implicitement l'hypothèse que l'ensemble d'information $\mathcal{F}_{t-1}$ est incorporé au lag de la veille λ_{t-1}. D'autre part, si H_0 est vraie alors les trois coefficients doivent être simultanément nuls : $\xi_0 = \xi_1 = \xi_2 = 0$. Au lieu d'utiliser la formule matricielle habituelle de DQ d'Engle et Manganelli (2004), on va plutôt conduire ce test simultané par Wald. Le grand avantage est de pouvoir estimer convenablement la matrice variance-covariance en entrée du test de Wald avec correction HAC (avec *prewhitening*). L'intérêt de cette façon de procéder est majeur, comme on va le voir bientôt. Le seuil critique est toujours fixé à 5%. Les résultats de *Goodness-of-fit* des modèles concurrents sont présentés au Tableau 7.

Concernant S_{tick} du Tableau 7, les colonnes (3) et (4) sont les mêmes que les colonnes (8) et (10) du Tableau 2; voir les résultats et les commentaires discutés dans la sous-section 3.1.3. Ce qui nous intéresse ici est que le modèle RMetrics est sévèrement rejeté pour les

niveaux de première et deuxième importances pour la réglementation, $\alpha = 1$ %, pas 2,5 %. Ceci est rassurant par rapport à ce qu'on a dit, le modèle RMetrics qui semble être le meilleur en enregistrant le meilleur score. Il le fait uniquement parce qu'il économise le capital de couverture plus tôt que les autres modèles dès qu'il y a une période de clustering de basses pertes. Ce faisant, ce modèle ne passe pas les tests conditionnels qui vérifient que la couverture est optimale tout au long des 2710 jours de l'échantillon.

Concernant les résultats du test d'optimalité selon la fonction de score S_{log}, les colonnes (5) et (6) du même Tableau 7 montrent que les conclusions sont très proches pour $\alpha = 1$ %. Ce qui diffère pour $\alpha = 2{,}5\%$ est que maintenant SimHist n'est pas rejeté (p-value = 0,1389 > 5 %). Le modèle RMetrics est rejeté pour $\alpha = 1$ % et 2,5%, comme pour S_{tick}. D'autre part, aucun modèle n'est rejeté pour $\alpha = 5$ % avec des p-values confortablement > 5 %, tout comme le sont celles du backtest *cc*. À noter que l'estimation HAC des matrices variance-covariance est faite en suivant Andrews et Monahan (1992). On présente également dans le tableau les résultats obtenus avec l'estimateur de Newey et West (1987), au lag = 5, pour être plus sûr de suffisamment tenir compte de l'hétéroscédasticité et des fortes autocorrélations des séries. La colonne (8) montre les p-values du test d'optimalité selon S_{log} obtenues par NW (à comparer avec les p-values de la colonne (6). On voit que les p-values sont pratiquement identiques. La même chose est constatée pour S_{tick} dont les p-values NW sont dans la colonne (7) par rapport à celles de la colonne (4). D'une façon globale, selon les deux fonctions S_{tick}et S_{log}, les modèles ST4, SEP3 et SimHist présentent un *Goodness-of-fit* convenable, aussi bien pour $\alpha = 1$ % que 2,5 %. Par ailleurs, aucun souci particulier concernant le niveau complémentaire $\alpha = 5$ % pour aucun des six modèles. En somme, le test d'optimalité que nous proposons est simple et peut être un complément précieux au test DM appliqué aux fonctions d'identification associées aux scores.

Tableau 7: Résultats du test d'optimalité - *Goodness-of-fit*

		—— vcov HAC (+ prewhite) ——				—– Newey-West —–		—– vcov OLS —–	
		—– tick loss —–		—– log loss —–		tick loss	log loss	tick loss	log loss
α	Modèles	Stat. (3)	p-value (4)	Stat. (5)	p-value (6)	p-value (7)	p-value (8)	p-value (9)	p-value (10)
5	SimHist	5,4109	0,1441	2,4775	0,4794	0,1582	0,5036	0,1294	0,2114
5	RMetrics	9,0565	0,0285	3,2528	0,3543	0,0292	0,3541	0,0713	0,1324
5	NO	6,6512	0,0839	2,0517	0,5618	0,0938	0,5857	0,1210	0,4207
5	TF	8,3292	0,0397	4,2134	0,2393	0,0484	0,2626	0,0619	0,0432
5	SEP3	5,7326	0,1254	2,1402	0,5438	0,1356	0,5665	0,1408	0,4093
5	ST4	5,3569	0,1475	2,1467	0,5425	0,1649	0,5688	0,1193	0,2873
2,5	SimHist	8,7815	0,0323	5,4957	0,1389	0,0342	0,1471	0,0004	0,0005
2,5	RMetrics	11,5222	0,0092	7,9991	0,0460	0,0093	0,0457	0,0038	0,0014
2,5	NO	8,5335	0,0362	6,0918	0,1072	0,0393	0,1162	0,0015	0,0010
2,5	TF	8,2295	0,0415	5,7030	0,1270	0,0445	0,1374	0,0051	0,0018
2,5	SEP3	5,8572	0,1188	4,2503	0,2357	0,1218	0,2466	0,0075	0,0028
2,5	ST4	5,9695	0,1131	4,2673	0,2340	0,1170	0,2442	0,0092	0,0027
1	SimHist	3,3670	0,3384	3,4461	0,3278	0,3330	0,3211	0,0230	0,0001
1	RMetrics	13,4720	0,0037	8,0576	0,0448	0,0037	0,0431	0,0012	0,0006
1	NO	13,2963	0,0040	10,4567	0,0151	0,0040	0,0153	0,0000	0,0000
1	TF	10,4507	0,0151	8,0917	0,0442	0,0157	0,0445	0,0005	0,0000
1	SEP3	7,0776	0,0695	6,0235	0,1105	0,0666	0,1056	0,0022	0,0000
1	ST4	3,2881	0,3493	2,6300	0,4523	0,3495	0,4479	0,0084	0,0000

Valeurs α en %

Un dernier point adresse une critique à l'endroit de Patton et al. (2019) concernant le test d'optimalité. Ces auteurs signalent qu'il pourrait arriver souvent que le test d'optimalité ne soit pas assez sélectif, qu'il pourrait rejeter trop ou pas assez parmi les modèles concurrents. Sachant qu'il y a une relation directe du test d'optimalité avec les backtests standards conditionnels *cc*, ce que disent ces auteurs reviendrait à constater que le backtest *cc* ne serait pas suffisamment sélectif et, dans ce cas, l'utilité des backtests de type conditionnel serait remise en question, ce qui est bien entendu loin d'être vrai. Dans les colonnes (9) et (10), on présente les p-values calculées en prenant telles quelles les matrices variance-covariance OLS (sans aucune correction). On peut voir que les p-values sont radicalement différentes par rapport à celles corrigées (colonnes (4) et (6) pour S_{tick} et S_{log} respectivement).

En fait, sans correction HAC de la matrice variance-covariance, le test d'optimalité aurait rejeté tous les modèles aussi bien pour $\alpha = 1$ % que 2,5 %. Nous pensons que leur constat

de rejets ou non rejets pourrait être lié à l'emploi de la procédure DQ telle quelle, sans correction pour robustesse. De plus, Patton et al. (2019) ne semblent pas mentionner dans leur papier une quelconque correction lors du test des coefficients de la régression DQ. Nous croyons finalement que les résultats quelque peu décevants de leur test d'optimalité viendraient de ce manque de correction, d'autant plus qu'ils précisent autour de leur équation (40) qu'ils tentent de réduire l'impact de la persistance de la série des VaR en éliminant purement ce terme de leur variable en question.

On aura vu comment les deux volets standards et comparatif du backtesting peuvent fonctionner ensemble dans deux objectifs. D'abord, dans le sens collaboratif, le volet standard sert de repère fixe faisant en sorte que le comparatif soit "averti" d'un problème donné, comme le cas de RMetrics. D'autre part, dans le sens évolutif, le comparatif avec sa richesse mathématique beaucoup plus puissante peut aider à développer de nouveaux outils plus efficaces pour un backtest final plus solide. C'est le cas, par exemple, de la nouvelle technique que nous proposons pour tester la couverture conditionnelle avec des régressions similaires à l'approche DQ, mais à l'aide du test de Wald dont la matrice variance-covariance estimée avec correction HAC, le tout utilisant la fonction d'identification associée à la *log loss* que nous recommandons vivement pour les travaux sur la VaR.

5. Conclusion

Nous pensons que le backtesting du risque de marché basé essentiellement sur la VaR 1 % est là pour rester. Il sera très probablement maintenu en parallèle après que l'un ou l'autre des backtests de la CVaR aura été techniquement fiabilisé et devenu un standard pour Bâle. Dans cette optique, la VaR garde toute son importance comme mesure du risque de marché.

Le test DM (Diebold et Mariano, 1995) est actuellement une pratique quasiment incontournable pour valider le backtesting comparatif des modèles de VaR. Notre contribution dans la présente recherche se situe à ce niveau, en primeur à notre connaissance, et est multiple. D'abord, nous montrons pourquoi le test DM ne devrait pas

être directement utilisé sur les fonctions de score. Ces dernières ne sont pas conçues pour mesurer une erreur entre une prédiction d'une entité aléatoire versus une réalisation ex-post, comme le suppose implicitement le test DM. D'autre part, l'objectif de la VaR n'est pas, non plus, d'égaler les réalisations ex-post, mais plutôt de couvrir les réalisations négatives du rendement. En définitive, bien que cette pratique soit très répandue dans la littérature, les fonctions de score ne peuvent pas et ne doivent pas servir pour construire les *loss differentials* du test DM. Nous documentons avec des cas concrets le fait que les fonctions de score peuvent sélectionner de faux meilleurs modèles et que le test DM ne les détecte pas, voire confirme leur supériorité par rapport aux autres modèles, et ce, à cause des raisons citées. Comme deuxième volet de la contribution, nous démontrons que les fonctions d'identification associées aux fonctions de score sont compatibles avec la théorie du test DM et documentons cela avec des résultats numériques concrets qui étayent notre démarche.

Par ailleurs, nous mettons aussi le doigt sur un problème lié à la puissance du test DM. En effet, Patton et Sheppard (2009) montrent que les fonctions de score n'assurent pas la même puissance du test DM et que cette puissance est maximale pour les fonctions 0-homogènes. Nolde et Ziegel (2017) et Patton et al. (2019) soutiennent qu'il est préférable d'employer le plus possible ce type de fonctions pour la VaR et la CVaR. En tant que troisième contribution, nous documentons un cas concret concernant la *tick loss* qui est non 0-homogène. Sa fonction d'identification occasionne une puissance insuffisante du test DM au point que les conclusions du test sont incohérentes (trop d'erreurs de type II). Dans ce sens, nous suggérons fortement d'utiliser avantageusement la fonction *log loss,* étant strictement consistante et 0-homogène. Ce serait un remplacement très avantageux de la *tick loss,* plutôt 1-homogène, malgré sa popularité dans la littérature où elle est souvent utilisée seule pour conclure quant aux modèles de VaR. D'une manière générale, les fonctions de score ne sont pas équivalentes ni interchangeables. Pour valider des modèles, il est crucial que l'outil de validation soit lui-même valide et approprié pour ce faire.

Le modèle RiskMetrics, dans sa parure la plus simple, joue un rôle central dans ce travail. C'est ce modèle que les fonctions de score désignent unanimement comme le meilleur modèle alors qu'il est le plus mauvais, étant sévèrement rejeté par le backtest standard *uc*.

Le diagnostic de ce phénomène montre qu'en périodes de hautes pertes, les couvertures quotidiennes de ce modèle sont plutôt très médiocres et comparables à celles du modèle normal. Par contre, ce modèle "profite" des périodes creuses de basses pertes, réagit plus vite et réduit plus le capital. Ce faux meilleur modèle obtient le meilleur score uniquement grâce à des économies de capital en basses pertes au point que les mauvais scores des jours de hits ne se voient pas dans le score global. Cela identifie clairement un défaut majeur du backtest comparatif. Le pire est, comme signalé plus haut, que le test DM ne détecte pas cette défaillance quand il est directement appliqué sur les fonctions de score, comme c'est la pratique actuelle dans la littérature.

En dernier résultat, nous proposons un moyen de rendre plus efficace le test de calibration conditionnelle de Nolde et Ziegel (2017) en y apportant une amélioration clé. Sans cette modification, ce test obtient un avis plutôt mitigé de la part de ses propres concepteurs, entre autres (voir, par exemple, Patton et al., 2019). Avec ce test amélioré, nous proposons à la fois un outil supplémentaire pour la détection des faux meilleurs modèles, un outil pour évaluer le *Goodness-of-fit* des modèles par rapport aux fonctions de scores utilisées, et, enfin, pour servir de backtest afin évaluer l'hypothèse de couverture conditionnelle d'une façon nouvelle et complémentaire aux backtests *cc* et *DQ.*

L'une des clés de voûte de ce travail est la conduite en parallèle du backtesting de la VaR dans ses deux volets standard et comparatif. Les mises au point régulières en comparant les résultats de l'un versus l'autre apportent une amélioration incontestable. De leur côté, les backtests standards servent de premiers repères en donnant un premier "avis" départageant les modèles concurrents. De l'autre côté, la richesse conceptuelle et la souplesse mathématique du backtesting comparatif peuvent ouvrir la voie à des possibilités de développer des outils nouveaux plus adaptés et plus efficaces. C'est le cas, par exemple, du test DM utilisant les fonctions d'identification pour valider les scores du côté backtesting comparatif qui peut être vu et utilisé pour classer les modèles selon leurs performances en termes de couverture inconditionnelle *uc*. Cela illuste la richesse et le grand avantage de conduire en parallèle les deux volets standard et comparatif qui devraient être systématiques et inséparables pour une conclusion du backtesting plus solide.

À titre d'avenue d'une future recherche et en termes de généralisation, toutes les précisions que nous apportons quant au backtesting comparatif de la VaR devraient également concerner celui de la CVaR. Vu les similitudes techniques et l'emploi commun du test DM pour la VaR et la CVaR, nous recommandons donc fortement d'utiliser ce test sur les fonctions d'identification associées plutôt que directement sur les scores CVaR et d'utiliser, pour une puissance optimale du test DM, la fonction de score, $FZ0$, démontrée 0-homogène pour la CVaR dans Patton et al. (2019, proposition 1, éq. (6)).

Annexes

A.1. Transformations linéaires des fonctions de score

On réécrit ici la forme générale (4.1)

$$S(v_t) = (I_t - \alpha) \times G(-v_t) - I_t \times G(y_t) + h(y_t). \qquad \text{(A.1)}$$

Les transformations linéaires $S^L(.) = a \times S(.) + b$, où $a > 0$ et $b \in \mathbb{R}$, ne changent pas les classements des modèles concurrents. En effet, $E[a \times S(.) + b] = a \times E[S(.)] + b$. Cela revient juste à amplifier simultanément les scores de tous les modèles par un facteur $a > 0$ et les translater tous par une quantité b, ce qui est neutre pour le classement des modèles étant une relation d'ordre.

Les transformations linéaires sont également neutres quant aux résultats du test DM (Diebold et Mariano, 1995) quand le *loss differential,* D_{ij}, est construit à partir des fonctions de score directement. La statistique de ce test est bâtie sur la différence normalisée des vecteurs de scores $D_{ij}^L = \{S^L(M_i) - S^L(M_j)\} = a \times \{((S(M_i) - S(M_j))\}$. La transalation b est éliminée dans la différence. Le facteur a est simplifié lors de la normalisation par l'écart type du *loss differential* D_{ij}^L.

La même conclusion s'applique pour le *loss differential,* D_{ij}, du test DM construit à partir des fonctions d'identification associée aux fonctions de score strictement consistantes. La fonction d'identification associée est la dérivée première de la fonction de score par rapport à la variable de la VaR. La constante b est éliminée dans la dérivée. Le facteur

multiplicatif a disparait avec la normalisation par l'écart type du *loss differential* correspondant.

Enfin, les transformations linéaires ne changent pas non plus les résultats du test d'optimalité. Les résidus généralisés, λ_t, s'obtiennent de la fonction d'identification associée qui est la dérivée première de la fonction S par rapport à v_t : $\lambda_t = \partial S(v_t)/\partial v_t$. On note λ_t^L les résidus correspondant à la transformée linéaire de S. Le calcul donne $\lambda_t^L = a\lambda_t$. En effet, $\partial\{a \times S(v_t) + b\}/\partial v_t = a \times \partial S(v_t)/\partial v_t) = a\lambda_t$. D'autre part, l'hypothèse nulle du test d'optimalité $H_0: E[\lambda_t|\mathcal{F}_{t-1}] = 0$ implique $E[\lambda_t^L|\mathcal{F}_{t-1}] = aE[\lambda_t|\mathcal{F}_{t-1}] = 0$, d'où la neutralité des transformations linéaires pour le test d'optimalité.

Il est aisé de généraliser les résultats précédents quand le terme additif de la transformation linéaire est fonction de y_t: $b = b(y_t)$. Vu que le score est l'espérance mathématique de S, $E[b(y_t)]$ est une constante pour tous les modèles. Tous les résultats où b est une constante sont donc valides, à condition que la fonction $b(y_t)$ soit intégrable sur le domaine de définition d'intérêt pour que l'intégrale du calcul d'espérance existe.

Un exemple d'utilisation des transformations linéaires est de changer d'échelle pour une meilleure lecture ou d'ajouter une constante pour rendre positifs les scores des modèles.

A.2. Autres formulations des fonctions de score

Les fonctions de score de la VaR strictement consistantes peuvent s'écrire avec la formulation générale suivante (voir Gneiting et Raftery, 2007, éq. (40))

$$S(v_t) = (I_t - \alpha) \times G(-v_t) - I_t \times G(y_t) + h(y_t). \qquad \text{(A.2)}$$

Deux variantes de la formulation (A.1) sont répandues dans les travaux récents. La première s'écrit (voir, par exemple, Fissler, Ziegel et Gneiting, 2016, éq. (1)) :

$$S(v_t) = (I_t - \alpha)\big(G(-v_t) - G(y_t)\big). \qquad \text{(A.3)}$$

La deuxième variante est (voir par exemple Nolde et Ziegel (2017, éq. (2.2)[11]) :

[11] L'équation (2.2) originale de Nolde et Ziegel (2017) est formulée pour la queue droite de distribution où $y_y, v_t > 0$ dans ce papier.

$$S(v_t) = (I_t - \alpha) \times G(-v_t) - I_t \times G(y_t). \qquad \text{(A.4)}$$

Il est aisé de voir que l'expression (A.3) peut être obtenue de la formulation générale (A.2) en posant $h(x) = \alpha G(x)$ qui est intégrable puisque G est continue (donc intégrable). Quant à la deuxième variante (A.4) on pose $h(x) = 0$ qui est intégrable, bien entendu. Les deux variantes (A.3) et (A.4) sont donc incluses dans la formulation générale (A.2). Plus encore, les trois formulations sont équivalentes et donnent le même classement des modèles concurrents, les mêmes résultats du test DM et les mêmes conclusions du test d'optimalité (voir preuve dans l'annexe A.1).

Il est à noter que les calculs des scores de la présente note sont faits selon la formulation (A.4).

A.3. Continuité et différentiabilité des fonctions de score de VaR

On réécrit la fonction de score de VaR de la forme (A.2) en utilisant la variable $u_t = -v_t$:

$$S(u_t) = (I_t - \alpha)G(u_t) - I_t G(y_t) + h(y_t) \qquad \text{(A.5)}$$

où I_t est la fonction indicatrice $I_t = 1_{y_t<u_t}$. On considére que la fonction G est continue, différentiable et strictement croissante et que la fonction h est intégrable pour que la fonction de score S soit strictement consistante. En dehors des points de discontinuité de I_t, cette fonction est continue et vaut 0 ou 1. Quand $u_t \neq y_t$, S est alors naturellement continue, étant le fruit de produits et de sommes de fonctions continues et de constantes. On focalise maintenant sur les points $u_t = y_t$. La fonction $I_t = 1_{y_t<u_t}$ a la particularité d'être continue à droite de y_t et vaut 0, et discontinue à gauche de y_t et vaut 1. La fonction de score S est, donc, continue à droite de y_t et sa limite à droite quand $u_t \to y_t^+$ s'écrit :

$$\lim_{u_t \to y_t^+} S(u_t) = (0 - \alpha)G(u_t) - 0 \times G(y_t) + h(y_t) = -\alpha G(y_t) + h(y_t) \qquad \text{(A.6)}$$

car $I_t = 0$ quand $u_t \to y_t^+$. D'autre part, quant à la limite à gauche quand $u_t \to y_t^-$, on a $I_t = 1$ et

$$\lim_{u_t \to y_t^-} S(u_t) = (1 - \alpha)G(y_t) - 1 \times G\big(y_y\big) + h(y_t) = -\alpha G(y_t) + h(y_t). \qquad \text{(A.7)}$$

Les deux expressions (A.6) et (A.7) étant identiques, les limites à gauche et à droite aux points $u_t = y_t$ sont alors égales. Par conséquent, malgré la présence de I_t dans son

expression (A.5), la fonction de score S peut être considérée continue par extension aux points $u_t = y_t$.

On passe à la différentiabilité de S. Au voisinage des points $u_t \neq y_t$, la fonction indicatrice I_t y étant une constante (qui vaut 0 ou 1), la fonction de score S est naturellement différentiable en tout point $u_t \neq y_t$, vu que G est différentiable sur tout son domaine de définition. Dans le cas des points $u_t \neq y_t, \forall t$ la dérivée de S existe et s'écrit pour un réel strictement positif $\epsilon > 0$:

$$\begin{aligned}
\lim_{\epsilon \to 0} \frac{S(u_t + \epsilon) - S(u_t)}{\epsilon} &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} [\ \{(1_{y_t < u_t + \epsilon} - \alpha) G(u_t + \epsilon) - 1_{y_t < u_t + \epsilon} G(y_t)\} \\
&\qquad - \{(1_{y_t < u_t} - \alpha) G(u_t) - I_t G(y_t)\}\ ] && \text{(A.8)} \\
&= \lim_{\epsilon \to 0} \frac{1}{\epsilon} [(1_{y_t < u_t} - \alpha)(G(u_t + \epsilon) - G(u_t))] && \text{(A.9)} \\
&= (I_t - \alpha) \times \lim_{\epsilon \to 0} \frac{G(u_t + \epsilon) - G(u_t)}{\epsilon} && \text{(A.10)} \\
&= (I_t - \alpha) \frac{\partial G}{\partial u_t}(u_t). && \text{(A.11)}
\end{aligned}$$

Le passage de (A.8) à (A.9) est dû au fait $1_{y_t < u_t + \epsilon} = 1_{y_t < u_t}$, car ϵ est suffisamment petit pour que $u_t + \epsilon$ reste loin des points de discontinuité de I_t quand $\epsilon \to 0$. Le terme de droite de la ligne (A.10) revient à l'expression de la définition de la dérivée de la fonction G évaluée au point u_t, comme montré dans (A.11).

On étudie maintenant la différentiabilité de la fonction de score S au voisinage des points $u_t = y_t$. Pour ce faire, on doit calculer les dérivées à droite et celles à gauche des points $u_t = y_t$ et de les comparer. L'intuition est de remarquer que la pente de la fonction I_t à droite et celle à gauche des points $u_t = y_t$ sont les mêmes et sont nulles pour les deux côtés. Soit $\epsilon > 0$ un réel qui tend vers 0^+(à droite de 0). La dérivée à droite de S à un point $u_t = y_t$ se calcule quand $\epsilon \to 0^+$ tel que :

$$\lim_{\epsilon\to 0^+}\frac{S(y_t+\epsilon)-S(y_t)}{\epsilon} = \lim_{\epsilon\to 0^+}\frac{1}{\epsilon}[\ \big(1_{y_t<y_t+\epsilon}-\alpha\big)G(y_t+\epsilon)-1_{y_t<y_t+\epsilon}G(y_t+\epsilon)$$
$$-\big(1_{y_t<y_t}-\alpha\big)G(y_t)+1_{y_t<y_t}G(y_t)\ ]$$
$$= \lim_{\epsilon\to 0^+}\frac{1}{\epsilon}[(1-\alpha)G(y_t+\epsilon)-G(y_t+\epsilon)+\alpha G(y_t)+0] \quad \text{(A.12)}$$
$$= \lim_{\epsilon\to 0^+}\frac{1}{\epsilon}\big[-\alpha\big(G(y_t+\epsilon)-G(y_t)\big)\big]$$
$$= -\alpha\times\lim_{\epsilon\to 0^+}\frac{G(y_t+\epsilon)-G(y_t)}{\epsilon}$$
$$= -\alpha\frac{\partial G}{\partial u_t}(y_t). \quad \text{(A.13)}$$

La ligne (A.12) découle de la précédente en appliquant $1_{y_t<y_t}=0$ et $1_{y_t<y_t+\epsilon}=1$ quand $\epsilon\to 0^+$. D'une façon similiaire, la dérivée à gauche d'un point $u_t=y_t$ se calcule cette fois avec $\epsilon<0$ quand $\epsilon\to 0^-$ tel que :

$$\lim_{\epsilon\to 0^-}\frac{S(y_t+\epsilon)-S(y_t)}{\epsilon} = \lim_{\epsilon\to 0^-}\frac{1}{\epsilon}[\ \big(1_{y_t<y_t+\epsilon}-\alpha\big)G(y_t+\epsilon)-1_{y_t<y_t+\epsilon}G(y_t+\epsilon)$$
$$-\big(1_{y_t<y_t}-\alpha\big)G(y_t)+1_{y_t<y_t}G(y_t)\ ]$$
$$= \lim_{\epsilon\to 0^-}\frac{1}{\epsilon}[(0-\alpha)G(y_t+\epsilon)-0+\alpha G(y_t)+0] \quad \text{(A.14)}$$
$$= -\alpha\times\lim_{\epsilon\to 0^-}\frac{G(y_t+\epsilon)-G(y_t)}{\epsilon}$$
$$= -\alpha\frac{\partial G}{\partial u_t}(y_t). \quad \text{(A.15)}$$

La ligne (A.14) découle de la précédente en appliquant $1_{y_t<y_t}=0$ et $1_{y_t<y_t+\epsilon}=0$ quand $\epsilon\to 0^-$. Les expressions (A.13) et (A.15) étant identiques, la dérivée à droite est alors égale à celle à gauche de tout point $u_t=y_t$. De plus, l'expression de ces dérivées revient à évaluer la formule (A.11) aux points $u_t=y_t$, vu que $I_t|_{u_t=y_t}=1_{y_t<y_t}=0$. Par conséquent, la fonction de score S peut être considérée différentiable en tout point et sa dérivée évaluée à un point u_t vaut :

$$\frac{\partial S}{\partial u_t}(u_t)=(I_t-\alpha)\frac{\partial G}{\partial u_t}(u_t) \quad \forall u_t, t=1\dots T. \quad \text{(A.16)}$$

L'expression (A.16) confirme l'intuition que le terme I_t se comporte comme une constante dans les calculs de dérivée de la fonction de score strictement consistante.

Références

Andrews, Donald Wilfrid Kao, and J Christopher Monahan. 1992. "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator." *Econometrica* 60 (4): 953–66.

Basel Committee On Banking Supervision (BCBS). 2016. "Minimum Capital Requirements for Market Risk, Publication No 352." *Bank for International Settlements (BIS)*, nos. Jan-2016: 1–92.

Basel Committee On Banking Supervision (BCBS). 2019. "Minimum Capital Requirements for Market Risk, Publication No 457." *Bank for International Settlements (BIS)*, nos. Jan-2019: 1–136.

Berkowitz, Jeremy, Peter Christoffersen, and Denis Pelletier. 2011. "Evaluating Value-at-Risk Models with Desk-Level Data." *Management Science* 57 (12): 2213–27.

Christoffersen, Peter F. 1998. "Evaluating interval forecasts." *International Economic Review* 39 (4): 841–62.

Diebold, Francis X, and Robert S Mariano. 1995. "Comparing Predictive Accuracy." *Journal of Business & Economic Statistics* 13 (3): 253–63.

Engle, Robert F, and Simone Manganelli. 2004. "CAViaR: Conditional autoregressive value at risk by regression quantiles." *Journal of Business and Economic Statistics* 22 (4): 367–81.

Fernández, Carmen, and Mark FJ Steel. 1998. "On Bayesian Modeling of Fat Tails and Skewness." *Journal of the American Statistical Association* 93 (441): 359–71.

Fissler, Tobias, and Johanna F Ziegel. 2016. "Higher Order Elicitability and Osband's Principle." *Annals of Statistics* 44 (4): 1680–1707.

Fissler, Tobias, Johanna F Ziegel, and Tilmann Gneiting. 2016. "Expected Shortfall Is Jointly Elicitable with Value at Risk-Implications for Backtesting." *Risk Magazine*.

Giacomini, Raffaella, and Halbert White. 2006. "Tests of Conditional Predictive Ability." *Econometrica* 74 (6): 1545–78.

Gneiting, Tilmann. 2011. "Making and Evaluating Point Forecasts." *Journal of the American Statistical Association* 106 (494): 746–62.

Gneiting, Tilmann, and Adrian E Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–78.

Harvey, David, Stephen Leybourne, and Paul Newbold. 1997. “Testing the Equality of Prediction Mean Squared Errors.” *International Journal of Forecasting* 13 (2): 281–91.

Kupiec, Paul H. 1995. “Techniques for verifying the accuracy of risk measurement models.” *Journal of Derivatives* 3 (2): 73–84.

Liu, Wei, Artur Semeyutin, Chi Keung Marco Lau, and Giray Gozgor. 2020. “Forecasting Value-at-Risk of Cryptocurrencies with Riskmetrics Type Models.” *Research in International Business and Finance* 54: 101259.

Lucas, André, and Xin Zhang. 2016. “Score-Driven Exponentially Weighted Moving Averages and Value-at-Risk Forecasting.” *International Journal of Forecasting* 32 (2): 293–302.

Mehta, Amit, Max Neukirchen, Sonja Pfetsch, and Thomas Poppensieker. 2012. “Managing Market Risk: Today and Tomorrow.” *McKinsey & Company McKinsey Working Papers on Risk* 32: 24.

Mincer, Jacob A, and Victor Zarnowitz. 1969. “The Evaluation of Economic Forecasts.” In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, 3–46. NBER.

Newey, Whitney K, Kenneth D West, and others. 1987. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica* 55 (3): 703–8.

Nolde, Natalia, and Johanna F Ziegel. 2017. “Elicitability and Backtesting: Perspectives for Banking Regulation.” *The Annals of Applied Statistics* 11 (4): 1833–74.

Osband, Kent. 1985. “Providing Incentives for Better Cost Forecasting.” PhD thesis, University of California, Berkeley.

Patton, Andrew J, and Kevin Sheppard. 2009. “Evaluating Volatility and Correlation Forecasts.” In *Handbook of Financial Time Series*, 801–38. Springer.

Patton, Andrew J, Johanna F Ziegel, and Rui Chen. 2019. “Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk).” *Journal of Econometrics* 211 (2): 388–413.

Rigby, Robert A, Mikis D Stasinopoulos, Gillian Z Heller, and Fernanda De Bastiani. 2019. *Distributions for Modeling Location, Scale, and Shape: Using Gamlss in R*. CRC press.

RiskMetrics, TM. 1996. “Technical Document, 4-Th Edition.” *New York, NY, JP Morgan Inc., December*.

Rockafellar, R Tyrrell, and Stanislav Uryasev. 2002. “Conditional Value-at-Risk for General Loss Distributions.” *Journal of Banking & Finance* 26 (7): 1443–71.

Saerens, Marco. 2000. “Building Cost Functions Minimizing to Some Summary Statistics.” *IEEE Transactions on Neural Networks* 11 (6): 1263–71.

Saissi Hassani, Samir, and Goerges Dionne. 2021. “Nouvelle Réglementation Internationale Du Risque de Marché : Rôles de La Var et de La Cvar Dans La Validation Des Modèles.” *Revue Assurances et Gestion Des Risques-Insurance and Risk Management* 87 (3-4): 169–207.

Stasinopoulos, Mikis D, Robert A Rigby, Gillian Z Heller, Vlasios Voudouris, and Fernanda De Bastiani. 2017. *Flexible Regression and Smoothing: Using Gamlss in R*. CRC Press.

Taylor, James W. 2019. “Forecasting Value at Risk and Expected Shortfall Using a Semiparametric Approach Based on the Asymmetric Laplace Distribution.” *Journal of Business and Economic Statistics* 37 (1): 121–33.

Thomson, William. 1979. “Eliciting Production Possibilities from a Well-Informed Manager.” *Journal of Economic Theory* 20 (3): 360–80.

Woodall, Louie. 2021. “Most Eu Banks Use Historical Simulation Approach to Var.” *Risk.Net* https://t.co/m286flrnYm?amp=1.