Institut de la finance structurée et des instruments dérivés de Montréal
Montreal Institute of Structured Products and Derivatives

NOTE TECHNIQUE

NT 12-01

Robustesse des méthodes de *backtesting*

Janvier 2012

Cette note technique a été rédigée sous la direction d'Élise St-Aubin Fournier.

Points importants

- La majorité des articles trouvés dans la littérature analyse les méthodes de *backtesting* appliquées au modèle de VaR.

- Plusieurs méthodes de *backtesting* sont analysées, et en général, la puissance des tests pour détecter un modèle de VaR invalide est assez faible.

- Les procédures de *backtesting* sur un horizon de 1 an (250 jours) s'avèrent peu efficaces, puisqu'on est en présence d'un problème d'événements rares.

- Les méthodes de *backtesting* qui évaluent seulement un quantile de la distribution ont une puissance plutôt faible. On peut augmenter la probabilité d'identifier un modèle non valide en analysant plusieurs niveaux de VaR ou en combinant plusieurs *backtests*.

- En général, les méthodes de *backtesting* ont tendance à ne pas rejeter assez souvent des modèles qui devraient l'être.

- Une grande majorité des entreprises étudiées utilisent des données contaminées dans leurs procédures de *backtesting*. Les méthodes de *backtesting* sont très sensibles aux données utilisées, ce qui peut donc biaiser les résultats de façon importante.

- On note des biais potentiels dans l'utilisation du *backtesting* : aléas moral et incitation à publier des méthodes de *backtesting* avec de bons résultats.

Les méthodes de *backtesting* sont utilisées pour évaluer un modèle, une stratégie ou une théorie, en l'appliquant à des données historiques. Ces méthodes sont de plus en plus populaires, puisqu'elles sont utilisées pour tester la validité des modèles de Valeur-à-Risque[1] (VaR) pour l'Accord de Bâle. Il aurait été intéressant d'avoir une vue globale du *backtesting* et de la fiabilité des résultats. Cependant, la grande majorité des articles trouvés dans la littérature analyse le *backtesting* appliqué aux calculs de VaR. Nous allons donc particulièrement nous concentrer sur ce point.

Propriétés à respecter

Une violation (ou *hit*) de la VaR se produit lorsque les profits ou pertes du portefeuille sont inférieurs à la VaR prévue *ex-ante*. On note $I_t(\alpha)$ la fonction indicatrice associée à une violation de la VaR à la date t pour un taux de couverture α% :

$$I_t(\alpha) = \begin{cases} 1 \; si \; x_t \leq VaR_{t|t-1}(\alpha) \\ 0 \; si \; x_t > VaR_{t|t-1}(\alpha) \end{cases}$$

Il existe un consensus général dans tous les articles étudiés sur les propriétés à tester par *backtesting* pour valider (ou invalider) un modèle de VaR. Christoffersen (1998) montre qu'en fait le problème de déterminer la validité d'un modèle de VaR peut se réduire à déterminer si la séquence des violations satisfait ces deux propriétés :

Propriété de couverture non conditionnelle : la probabilité que se réalise *ex-post* une perte plus grande que la VaR reportée *ex-ante* doit être égale au taux de couverture α% :

$$E[I_t(\alpha)] = P[I_t(\alpha) = 1] = \alpha$$

[1]Définition de la VaR : $P\left[Perte_t < VaR_{t|t-1}(\alpha)\right] = \alpha, pour\ un\ taux\ de\ couverture\ \alpha\%$

Propriété d'indépendance : les violations de la VaR à des dates différentes doivent être indépendamment distribuées. Donc, la variable $I_t(\alpha)$ est indépendante de $I_{t+k}(\alpha)$, pour toutes valeurs de *k* différentes de zéro.

On regroupe souvent ces deux propriétés en un seul critère : la séquence de violations $I_t(\alpha)$ doit être indépendante et identiquement distribuée selon une loi de Bernoulli de paramètre α. Lorsque l'une des deux propriétés n'est pas respectée, on peut affirmer que le modèle n'est pas valide. Dans le cas où ces deux hypothèses sont confirmées, on parle alors de couverture conditionnelle.

Différentes méthodes de *backtesting*

Il existe un très grand nombre de tests de *backtesting*. Nous allons ici énumérer les plus populaires, selon les articles étudiés, et évaluer leur fiabilité et leur robustesse (Campbell (2005), Haas (2001) et Christoffersen (1998)). Certaines de ces méthodes testent la propriété de couverture non conditionnelle, d'autres la propriété d'indépendance, et certains testent les deux propriétés jointes.

Proportion de violations (*Proportion of Failures*)

Cette méthode s'attarde particulièrement à la première propriété, soit la couverture non conditionnelle. On compare la proportion de *hits* dans l'échantillon ($\hat{\alpha}$) à la proportion imposée (α) :

$$\hat{\alpha} = \frac{1}{T}\sum_{t=1}^{T} I_t(\alpha)$$

Ensuite, plusieurs tests et critères statistiques sont disponibles pour évaluer la validité de ce résultat. Pour plus de détails, voir Campbell (2005) ou Haas (2001). Ce dernier propose une deuxième méthode amenée par Kupiec (2005), soit le temps jusqu'à la première violation (*Time until First Failure*). Cependant, la puissance de ce test est assez faible.

Ce type de *backtests* peut difficilement détecter un modèle de VaR qui sous-évalue systématiquement son risque, particulièrement pour un échantillon de petite taille (1 an par exemple, ce qui semble assez utilisé en pratique).

Test de Markov (Christoffersen, 1998)

Cette méthode teste le critère d'indépendance. On examine si la probabilité d'une violation de la VaR dépend de l'occurrence ou non d'une violation de VaR au jour précédent.

$$P(I_t(\alpha) = 1) = \hat{\alpha} = P(I_t(\alpha) = 1 | I_{t-1}(\alpha) = 1) = P(I_t(\alpha) = 1 | I_{t-1}(\alpha) = 0)$$

Ce test permet de détecter une dépendance entre deux *hits* consécutifs. Cependant, il ne permet pas de détecter tout autre type de dépendance. On peut aussi modifier cette procédure pour tester aussi la propriété de couverture non conditionnelle. En plus de demander que la probabilité d'avoir une violation soit la même peu importe s'il y a eu violation la veille, on teste si cette probabilité est égale à α.

Cependant, Campbell (2005) et Piontek (2010) arrivent à la conclusion que les tests joints, contrairement à ce qu'on peut penser, sont moins efficaces puisque la satisfaction d'un des deux critères rend l'analyse du deuxième plus difficile.

Approche basée sur la durée (Christoffersen et Pelletier, 2001)

Les auteurs utilisent l'intuition que si les violations de la VaR sont indépendantes les unes des autres, la durée entre deux violations devrait être indépendante du temps écoulé entre les deux derniers *hits*. Les tests prouvent que cette procédure de *backtesting* possède plus de puissance. Cependant, elle démontre de moins bons résultats lorsqu'on choisit une période pour le test d'un an, comme c'est souvent le cas en pratique. Quelques années plus tard, un article apportant une amélioration à cette technique a été publié (Candelon, Colletaz, Hurlin et Tokpavi, 2008). Cette nouvelle méthode utilise la fonction génératrice des moments (GMM). Elle base en fait ces tests statistiques sur les moments définis par cette fonction. Les auteurs ont testé leur méthode et ont obtenu de bons résultats, et ont surperformé l'approche traditionnelle pour des périodes de *backtesting* réalistes (1 an).

Campbell (2005) relève que les tests sur la durée semblent avoir plus de puissance que le test de Markov. Cependant, pour toutes les méthodes de *backtesting* qui testent la propriété d'indépendance, le régulateur doit spécifier l'hypothèse alternative. Il faut donc déterminer d'avance quels types d'anomalies sont recherchés dans les données. Ce type de test ne peut pas détecter d'autres types d'anomalies que ceux imposés par le régulateur. Cela représente une grande limite du modèle. Hurlin et Tokpavi (2008) confirment cette hypothèse en disant que les tests d'indépendance ne prennent pas en compte les dépendances d'ordre supérieur à un.

Tests basés sur plusieurs niveaux de VaR

Les deux mêmes critères devraient aussi être respectés pour plusieurs niveaux de VaR. De plus, les violations de VaR à tous les niveaux devraient être indépendantes les unes des autres. Cette procédure consiste donc à tester non seulement la VaR 1 %, mais aussi 5 %, 10 %, etc. Selon plusieurs auteurs, ces tests ont légèrement plus de puissance que les tests à un seul niveau de VaR.

Fonction de perte

L'information contenue dans la séquence de *hits* est plutôt limitée. Il serait intéressant d'analyser aussi l'amplitude des pertes lorsque celles-ci excèdent la VaR (aussi appelée *Expected Shortfall*). À partir de la séquence de violations, il est facile de déterminer la moyenne des pertes lorsque celles-ci excèdent la VaR. Cependant, afin de déterminer si cette moyenne est raisonnable ou non, il faut faire une hypothèse à propos de la fonction de perte réelle, ce qui complique le tout. Une conclusion de test stipulant que la moyenne des pertes excédentaires est « trop élevée » par rapport à nos hypothèses de fonction de perte signifierait donc que soit le modèle de VaR est mal choisi, ou que l'hypothèse de fonction de perte définie préalablement n'est pas valide. Ce type de *backtest* est donc davantage utile pour comparer des modèles entre eux que pour valider la précision d'un seul modèle.

Selon Campbell (2005), Hurlin et Tokpavi (2007) et Haas (2001), pour tous les *backtests* énoncés précédemment, la puissance n'est pas suffisante pour prédire correctement des erreurs dans les modèles de VaR. De plus, en pratique, on utilise surtout un horizon de 250 jours (1 an) pour tester une VaR 1 % quotidienne. On est donc en présence d'un problème d'événements rares, puisqu'on s'attendrait en moyenne à 2,5 violations de la VaR quotidienne sur 250 jours, ce qui est très peu pour assurer la fiabilité du test. Ces procédures ont donc une puissance très faible sur les petits échantillons. Selon les études, un échantillon d'au moins 1000 données serait nécessaire pour augmenter cette puissance. Cependant, un échantillon de la sorte réduirait la fréquence possible d'évaluation de la VaR.

En fait, selon Hurlin et Tokpavi (2007), les méthodes de *backtesting* ont tendance à ne pas rejeter la validité d'un grand nombre de modèle d'évaluation du risque portant sur les mêmes actifs, même si ces mesures ont des résultats sensiblement différents.

Haas (2001) affirme qu'un seul test de *backtesting* ne sera jamais suffisant. Afin d'avoir suffisamment de puissance pour juger la qualité d'un modèle de VaR, il serait préférable de combiner différentes méthodes. Hurlin et Tokpavi (2007) affirment aussi, suite à plusieurs tests, que les procédures actuelles de *backtesting* sont très peu discriminantes et qu'elles ont une tendance très forte à conclure que « tout va pour le mieux dans le meilleur des mondes ».

Deux types d'erreurs peuvent survenir dans le résultat d'un *backtesting* : rejeter un modèle correct (type I) ou accepter un modèle incorrect (type II). Les procédures de *backtesting* sont faites pour contrôler l'erreur de type I. Dans son article, Piontek (2010) a effectué plusieurs tests et arrive à la conclusion que l'erreur de type II peut être particulièrement élevée. Cependant, l'utilisation d'un modèle incorrect (validée par une procédure de *backtesting*) peut avoir un impact important sur le risque de l'entreprise.

Données contaminées

Dans leur article, Frésard, Pérignon et Wilhelmsson (2011) abordent le problème de la qualité des données utilisées pour faire du *backtesting*. Les données de la distribution des gains et des pertes d'un portefeuille sont considérées comme « propres » lorsqu'on tient compte seulement des changements dans la valeur du portefeuille provenant des positions à la date précédente. Si les données incluent autre chose (revenus supplémentaires dus à un changement dans la quantité d'actifs détenue ou les frais et commissions, par exemple), on considère ces données comme contaminées. En effet, le risque relié à ces montants n'est pas inclus dans le calcul de la VaR.

Les auteurs collectent des informations sur les 200 plus grandes banques commerciales dans le monde, et arrivent à la conclusion que moins de 6 % de ces banques utilisent des données non-contaminées pour *backtester* leur VaR. Ils montrent ensuite que l'utilisation de ces données contaminées a un impact important sur les résultats du *backtesting*, puisque les banques enregistrent ainsi beaucoup moins de violations de la VaR que les autres (la distribution des gains et des pertes est lissée par l'utilisation de mauvaises données). Finalement, ils montrent que les tests de *backtesting* tendent à rejeter beaucoup moins souvent les modèles de VaR lorsque des données contaminées sont utilisées. Les méthodes de *backtesting* utilisées actuellement sont donc très sensibles à la contamination des données, et cela peut biaiser les résultats de façon importante.

Biais potentiels

Il est important de noter qu'il existe plusieurs biais potentiels dans l'utilisation du *backtesting*. Tout d'abord, les nouvelles méthodes publiées ont forcément de bons résultats au niveau de la puissance des tests. En effet, les articles portant sur de nouvelles méthodes de *backtesting* ayant une puissance plus faible ne seraient probablement pas publiés. Cependant, lorsqu'ils sont testés par des auteurs différents, la puissance des tests est beaucoup moins élevée. Un questionnement sur la robustesse des résultats et la possibilité de les répliquer peut donc se poser.

Ensuite, si on prend en considération que c'est la même personne qui produit la VaR et qui la teste, on note un aléa moral important. Afin de réduire ce biais, certains suggèrent une séparation entre le département de *trading* et celui de gestion de risque. Cependant, plusieurs auteurs remettent en doute cette idée, puisqu'il s'agit encore de la même entreprise. Les gestionnaires en place peuvent tenter d'influencer les résultats du *backtesting*, même si les deux étapes (calcul et validation de la VaR) se font dans des départements différents.

Finalement, les procédures de *backtesting* se basent sur les résultats du modèle étudié en le testant sur des données historiques. Néanmoins, on sait déjà que de bons résultats dans le passé ne sont pas garants de bons résultats dans le futur. Cela pourrait donc être une autre source de biais.

Comme mentionné précédemment, il aurait été intéressant d'avoir une vue plus globale sur le sujet. Cependant, la majorité des articles sur le *backtesting* portent particulièrement sur son utilisation par les régulateurs de l'Accord de Bâle, et donc les méthodes de VaR. On peut conclure que si on émet des réserves à propos du *backtesting* pour une méthode en particulier, il est possible que ces réserves soient aussi présentes pour d'autres modèles à tester.

Bibliographie

CAMPBELL, Sean D. (2005). A Review of Backtesting and Backtesting Procedures, Washington: Federal Reserve Board.

CANDELON, Bertrand, COLLETAZ, Gilbert, HURLIN, Christophe et TOKPAVI, Sessi (2011). Backtesting Value-at-Risk: A GMM Duration-Based Test, *Journal of Financial Econometrics*, pp. 1-30.

CHRISTOFFERSEN, Peter (1998). Evaluating Interval Forecasts, *International Economic Review*, vol. 39, pp. 841-862.

CHRISTOFFERSEN, Peter et PELLETIER, Denis (2001). Backtesting Value-at-Risk : A Duration-Based Approach, *Journal of Empirical Finance*, vol. 2, pp. 81-108.

ESCANCIANO, J. Carlos et OLMO, Jose (2011). Robust Backtesting Tests for Value-at-Risk Models, *Journal of Financial Econometrics*, vol. 9, pp. 132-161.

EWERHART, Christian (2001). Market Risks, Internal Models, and Optimal Regulation: Does Backtesting Induce Banks to Report Their True Risks?, University of Zurich.

FRÉSARD, Laurent, PÉRIGNON, Christophe et WILHELMSSON, Anders (2011). The Pernicious Effects of Contaminated Data in Risk Management, *Journal of Banking and Finance*, vol. 35, pp. 2569-2583.

HAAS, Marcus (2001). New Methods in Backtesting, Financial Engineering Research Center.

HURLIN, Christophe et TOKPAVI, Sessi (2008). Une évaluation des procédures de Backtesting « Tout va pour le mieux dans le meilleur des mondes », *Finance,* vol 29, pp. 53-80.

KERKHOFF, Jeroen et MELENBERG, Bertrand (2003). Backtesting for Risk-Based Regulatory Capital, Tilburg University.

LUCAS, André (2001). Testing Backtesting: an Evaluation of the Basle Guidelines for Backtesting Internal Risk Management Models of Banks, *Journal of Money, Credit and Banking*, vol. 33, pp. 826-846.

PÉRIGNON, Christophe, DENG, Zi Yin et WANG, Zhi Jun (2008). Do Banks Overstate their Value-at-Risk?, *Journal of Banking and Finance*, vol. 32, pp. 783-794.

PIONTEK, Krzysztof (2010). The Analysis of Power for Some Chosen VaR Backtesting Procedures: Simulation Approach, *Advances in Data Analysis, Data Handling and Business Intelligence*, pp. 481-490.

