Essais en finance dans un environnement de Big data

Title: Essays in finance in a big data environment

Mots clés : Big data, machine learning, hypothèse d'efficience des marchés, sélection de variables, prévision, modélisation

Keywords: Big data, machine learning, efficiency market hypothesis, variable selection, forecasting

Description des objectifs, positionnement du projet et méthodologie

L'ère numérique a créé des quantités de données qui continuent de croître de façon exponentielle. L'*International Data Corporation* estime que le monde génère plus de données tous les deux jours que toute l'humanité depuis la nuit des temps jusqu'en 2003. Le Big data révolutionne donc le secteur financier et a le potentiel de façonner de manière significative la recherche future en finance.
Le projet de cette thèse est d'apporter de nouvelles analyses de théories financières en exploitant le big data tout en adaptant de nouvelles techniques statistiques et économétriques.

L'hypothèse de l'efficience des marchés (*Efficiency Market Hypothesis*, EMH) est un concept central de la théorie financière moderne. D'après l'EMH de Fama (1970) un marché dans lequel les prix reflètent pleinement et toujours l'information disponible est appelé efficient. En d'autres termes, à l'équilibre, les cours reflètent d'une manière fidèle toute l'information disponible. Cette définition de l'efficience des marchés implique que (i) les cours incluent instantanément les conséquences des événements passés et reflètent précisément les anticipations exprimées sur les événements futurs ; et (ii) aucun investisseur ne peut réaliser des rentabilités anormales en basant sa stratégie sur l'information disponible puisque celle-ci est à tout moment reflétée dans les cours boursiers. Il existe trois formes d'efficience pour définir le concept "d'information disponible" : faible, semi forte ou forte. L'efficience semi-forte suppose que les cours reflètent immédiatement toute l'information à caractère publique. Cette forme d'efficience peut être évaluée à partir des régressions prédictives dont l'objectif est de tester si les changements de prix (les rentabilités) peuvent être prédits à partir de variables financières ou macroéconomiques.
Un problème majeur central de ces modèles est une distorsion sévère de la taille du test en présence de prédicteurs hautement persistants couplés à une endogénéité de la régression. En effet, certaines variables peuvent présenter une mémoire courte (par exemple, les bons du Trésor), tandis que d'autres sont très persistantes (par exemple, la plupart des prédicteurs financiers et macroéconomiques). Comment gérer ces deux problèmes dans un cadre riche en données ? Par ailleurs, est-ce que d'autres prédicteurs, comme les indicateurs techniques ou les indicateurs de sentiments des investisseurs ne pourraient-ils pas apporter une information supplémentaire aux variables « classiques » ?

Le **Modèle d'Evaluation Des Actifs Financiers** (MEDAF) (*Capital Asset Pricing Model*, CPAM), développé par Sharpe (1964), s'attache à décrire les conditions d'équilibre et la formation des prix sur les marchés d'actifs financiers. La relation du MEDAF montre qu'il existe, à l'équilibre des marchés, une relation croissante entre l'espérance de rentabilités (ou rendements attendus) des titres et leur risque systématique, c'est-à-dire la variance des rentabilités du titre imputable à la variance des rentabilités du portefeuille de marché. Le marché rémunère donc uniquement le risque systématique des titres (non diversifiable), à savoir l'excès de rentabilités ou la prime de risque (compensation pour l'investisseur de son risque de placement pour "battre" le marché). Par la suite, Fama et French (1992) ont proposé un modèle à 3 facteurs qui est une amélioration du MEDAF en intégrant deux anomalies constatées sur les marchés bousiers au MEDAF pour expliquer la prime de risque d'un titre, et donc son effet sur les rendements attendus, par (1) la prime de risque du marché, (2) la prime de risque de taille (capitalisation des titres) et (3) la prime de risque de valeur. Un grand nombre d'études se sont intéressées à la recherche de facteurs expliquant la prime de risque des actions et ont abouti à des centaines de candidats potentiels.
L'une des tâches fondamentales auxquelles est confronté aujourd'hui le domaine de la tarification des actifs (*asset pricing*) est de discipliner la prolifération des facteurs. En particulier, une question qui reste ouverte est la suivante : quels seraient les facteurs les plus pertinents parmi tous les candidats potentiels ? Comment juger si un nouveau facteur ajoute un pouvoir explicatif à la tarification des actifs, par rapport aux centaines de facteurs que la littérature a jusqu'à présent produits ?

La volatilité des actifs financiers est d'une grande importance pour de nombreuses applications en finance. Des estimations et des prévisions fiables sont essentielles pour la gestion des risques et l'allocation des actifs. Contrairement aux séries de rentabilités, la volatilité financière est prévisible et a reçu une grande attention dans la communauté de recherche en économétrie financière. De nombreuses études utilisent le modèle autorégressif hétérogène de la variance réalisée (HAR-RV, *heterogeneous autoregressive model for realized variance*) mis au point par Corsi (2009) ainsi que ses différentes extensions pour prédire la volatilité des marchés boursiers. L'apport de variables explicatives de la volatilité apparaît comme une extension naturelle de ces modèles mais ce pose alors le problème de la sélection de ces variables, d'autant plus si le nombre de prédicteurs potentiels est important, afin d'obtenir à la fois un modèle parcimonieux et économique interprétable.

Les progrès des **techniques d'apprentissage automatique** (*machine learning*), grâce à une abondance sans précédent de sources de données dans de nombreuses disciplines, offrent des possibilités d'analyse de données économiques et financières. À l'ère du big data, les méthodes de rétrécissement (*shinkrage methods*) deviennent de plus en plus populaires dans l'inférence et la prédiction économétriques grâce à leurs propriétés de sélection de variables et de régularisation. En particulier, les méthodes de régressions pénalisées (ou régularisées) semblent pertinentes, comme le Lasso (Tibshirani, 1996), Elastic-Net (Zou et Hastie, 2005) et leurs extension,. De même, des méthodes de machine learning « non linéaire » pourraient être des alternatives intéressantes, comme, par exemple, les forêts aléatoires, les réseaux de neurones profonds ou encore les arbres boostés.

Bibliographie

Audrino F., Knaus S.D., (2016). Lassoing the HAR model: a model selection perspective on realized volatility dynamics. Econometric Reviews 35, 1485–1521.
Corsi F. (2009). A Simple Approximate Long-Memory Model of Realized Volatility. Journal of Financial Econometrics 7(2), 174–196.
Fama E. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance 25, 383–417.
Fama E.F., K.R. French (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56.
Feng G., Giglio S., Xiu D. (2020). Taming the Factor Zoo: A Test of New Factors. The Journal of Finance 75 (3), 1327-1370.
Goldstein I., Spatt C.S., Ye M. (2021). Big Data in Finance. Working paper No 28615, NBER.
Gu S., Kelly B., Xiu D. (2020). Empirical Asset Pricing via Machine Learning. The Review of Financial Studies 33, 2223–2273.
Neely C.J., Rapach D.E., Tu J., Zhou G. (2014). Forecasting the Equity Risk Premium: The Role of Technical Indicators. Management Science 60 (7), 1772–1791.
Sharpe W.F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. The Journal of Finance 19, 425–42.
Tibshirani R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B 58, 267-288.
Zou H., Hastie T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67, 301-320.