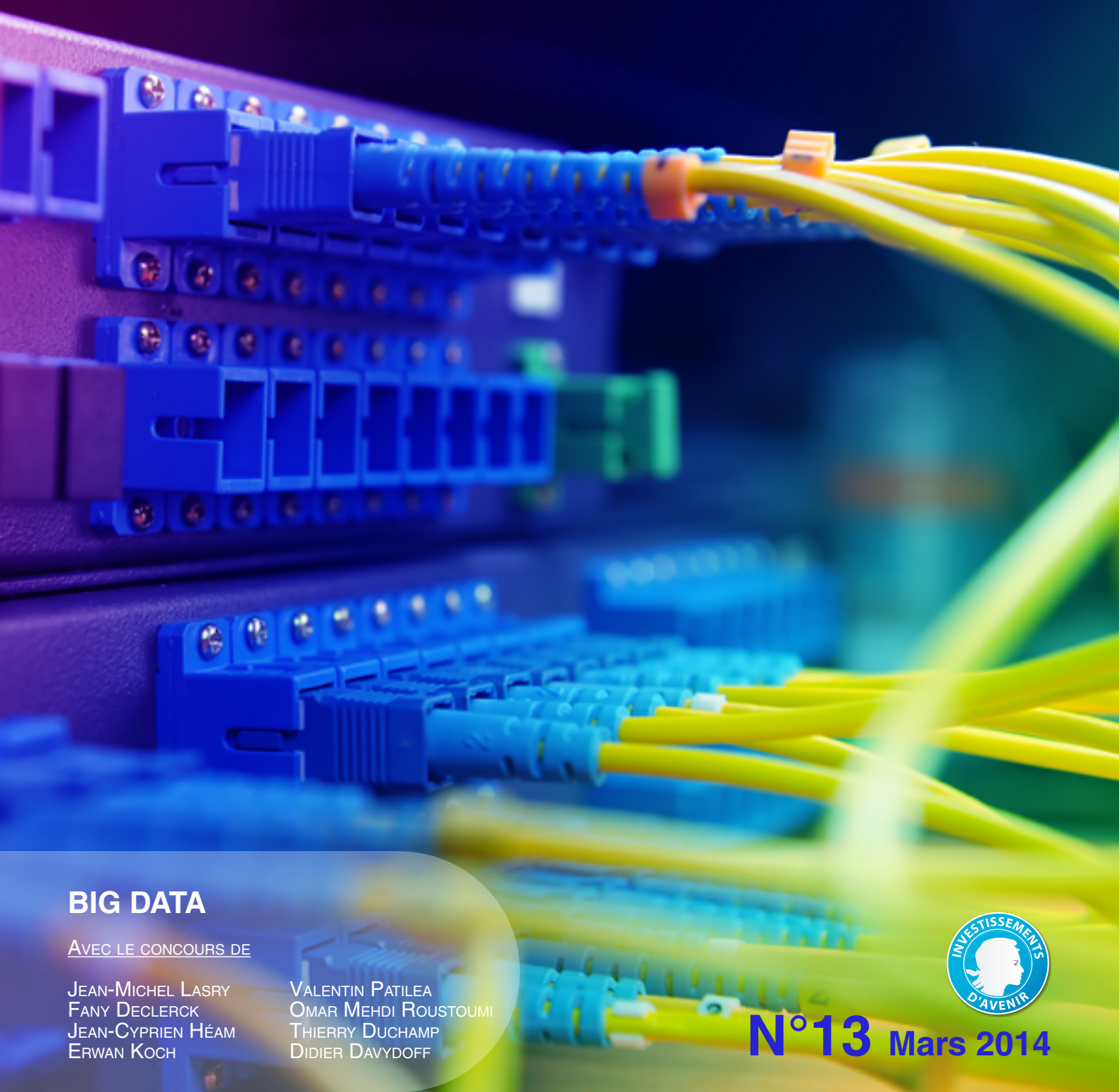
LES CAHIERS

Louis Bachelier



BIG DATA

Avec le concours de

Jean-Michel Lasry
Fany Declerck
Jean-Cyprien Héam
Erwan Koch
Valentin Patilea
Omar Mehdi Roustoumi
Thierry Duchamp
Didier Davydoff

N°13 Mars 2014

INSTITUT
Louis Bachelier

www.institutlouisbachelier.org

PROMOUVOIR, PARTAGER ET DIFFUSER LA RECHERCHE EN FINANCE

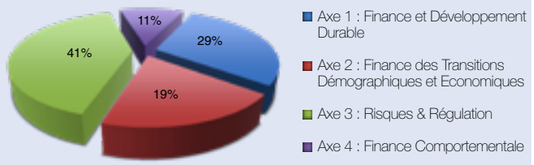
Créé en septembre 2008, l'Institut Louis Bachelier (ILB) est un centre de recherche en réseau de dimension internationale qui a pour objet de promouvoir, partager et diffuser la recherche et l'enseignement français en finance.

CRÉATION D'ÉQUIPES SCIENTIFIQUES D'EXCELLENCE

L'Institut Louis Bachelier constitue un dispositif unique réunissant, autour de partenariats industriels, les meilleures équipes de recherche en économie et mathématiques ; en atteste la labellisation LABEX (Laboratoire d'Excellence) obtenue par l'Institut Louis Bachelier dans le cadre de son projet Finance et Croissance Durable.

- **Création de programmes de recherche en lien direct avec l'industrie financière** : 30 chaires et initiatives de recherche ont été créées sous l'égide de l'Institut Europlace de Finance (EIF) et de la Fondation du Risque (FDR) depuis 2007, regroupant plus de 200 chercheurs.

RÉPARTITION DES 30 CHAIRES ET INITIATIVES DE RECHERCHE SELON LES QUATRE AXES STRATÉGIQUES DU LABEX FINANCE ET CROISSANCE DURABLE



- **Gestion et montage de projets R&D innovants en collaboration avec le Pôle Finance Innovation.**
- **Contribution et soutien à l'émergence de nouvelles formations aux niveaux licence, master et doctorat en phase avec les besoins de la Place de Paris.**
- **Coopération avec des universités et centres de recherche français, européens, américains et asiatiques.**

VALORISATION DE LA RECHERCHE

L'Institut Louis Bachelier assure la diffusion la plus large et la plus efficace des résultats issus de ses programmes de recherche auprès notamment des autorités de régulation françaises et européennes.

- **Revue trimestrielle "Les Cahiers Louis Bachelier"** qui présente des travaux de recherche issus de ses chaires et initiatives de recherche dans un langage accessible à un large public.
- **Publication de discussion papers** visant à éclairer les pouvoirs publics ainsi que les professionnels de la finance sur des sujets d'actualité.
- **Portail de la "Recherche en Finance" en partenariat avec l'AGEFI.**
- **Réseau communautaire de la recherche en finance :** www.louisbachelier.org

ESPACE DE RÉFLEXION ET DE DÉBATS À L'ÉCHELLE EUROPÉENNE

L'Institut Louis Bachelier est un véritable lieu de rencontre et de mise en relation destiné à favoriser les interactions entre le monde de la recherche et les acteurs économiques.

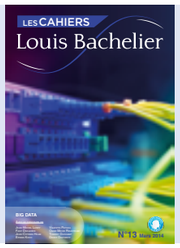
- **Financial Risks International Forum :** cette manifestation annuelle a pour objectif de présenter les meilleurs travaux de recherche internationaux et d'échanger, par le biais de débats et de tables rondes, sur les préoccupations des acteurs économiques.
- **Les Semestres Thématiques :** organisés sous forme de conférences, de séminaires et de cours, les semestres thématiques visent à favoriser les échanges entre académiques et professionnels sur une problématique commune.
- **La journée des chaires :** organisée annuellement, cette manifestation a pour but de confronter les travaux réalisés dans le cadre des chaires et initiatives de recherche de l'Institut Louis Bachelier.
- **Les Matinales Scientifiques :** ont pour objet de faire le point sur les derniers développements de la recherche en finance à travers les projets de recherche soutenus par l'Institut Europlace de Finance.

SOMMAIRE

FONDATIONS DE RECHERCHE

FONDATION DU RISQUE
Louis Bachelier

INSTITUT EUROPLACE DE FINANCE
Louis Bachelier



PUBLICATION DE L'INSTITUT LOUIS BACHELIER
Palais Brongniart
28 place de la Bourse - 75002 PARIS
Tél. 01 73 01 93 40
www.institutlouisbachelier.org
www.louisbachelier.org

CHEFS DE PROJET
Cyril Armange
Loïc Herpin

CONTACT
cyril.armange@institutlouisbachelier.org
loic.herpin@institutlouisbachelier.org

DIRECTEUR DE LA PUBLICATION
Jean-Michel Beacco

RÉDACTRICE EN CHEF
Isaure du Fretay

AVEC LA PARTICIPATION DE
Coralie Bach

PARTENAIRES
• MPG Partners
• IODS

CONCEPTION GRAPHIQUE, COUVERTURE ET RÉALISATION
Gaël Nicolet
La Cote Bleue
10-12 place Vendôme - 75001 Paris
Tél. 01 44 76 85 85
www.lacotebleue.fr

IMPRIMEUR
Kava
42, rue Danton - 94270 Le Kremlin-Bicêtre
Tél. 06 14 32 96 87

ÉDITO

Mégadonnées, analytique 2.0, zettaoctets, infomagique, infom@gic…
Ces terminologies indiquent-elles une nouvelle révolution scientifique ?

Depuis plusieurs décennies, on observe un développement rapide et continu de l'outil informatique, des capacités de stockage de données et des temps de réponse et de calcul. Ces capacités technologiques accrues sont cependant intégrées par à-coup par les entreprises du fait de coûts d'adaptation significatifs en matériel, formation de personnel et gouvernance.

La précédente rupture de ce type pour la finance et l'assurance a eu lieu au début des années 90 avec d'une part la création des marchés de cotation électroniques et d'autre part le suivi en temps réel des comptes courants, des crédits permanents et de la gestion de stocks en temps réel. D'analyses s'appuyant sur quelques milliers d'observations, on est passé à des analyses fondées sur quelques dizaines de millions d'observations, soit une multiplication par 10 000 des tailles des bases utilisées. Ce saut, pas uniquement quantitatif, a permis aussi de disposer d'informations nouvelles. Celles-ci ont conduit à d'autres formes de marchés avec par exemple le trading haute-fréquence, l'introduction de régulations adaptées, etc.

La nouvelle rupture concernant les données est du meme type que la précédente avec un effet d'échelle similaire. En effet les questions actuellement posées sont assez proches de celles du début des années 90 : Comment ne pas se laisser submerger par les données ? Doit-on employer des méthodes automatiques d'analyse de ces données ou avoir des approches réfléchies de ces données massives ? L'intérêt est-il dans le nombre de données ou dans l'existence de nouveaux types de données et de questions à considérer ? Toutes les entreprises doivent-elles s'adapter à ce nouvel environnement ou cette adaptation est elle trop couteuse par rapport au gain espéré ? Comment protéger les aspects privés dans l'utilisation des nouvelles données ? Quelles parts respectives de la gouvernance laissées aux dirigeants, aux services informatiques, aux spécialistes marketing, risque...concernant ces nouveaux développements ?

Les méthodes automatiques des années 90, légèrement améliorées, connues sous le nom générique de "data mining" sont de nouveau proposées pour l'analyse de données massives. Cependant un "forage" au hasard dans les bases de données se révèle couteux pour une faible productivité. Avant de faire un forage, il vaut mieux spécifier ce que l'on cherche et cerner la zone ou forer. Que cherche t'on et que peut-on espérer trouver ?

Il faut distinguer deux apports potentiels de ces bases :

- On peut utiliser ces données pour améliorer les réponses à des questions classiques. Un exemple type est l'utilisation de données de géolocalisation pour améliorer la prévision des risques d'accidents automobiles et proposer de nouveaux types de contrats d'assurance automobile. De meme des données du web peuvent servir à mieux connaître les choix de consommation des personnes et à mieux cibler les campagnes commerciales. Enfin, des compteurs intelligents facilitent le suivi des consommations d'électricité en temps réel et la mise en place de processus de production efficace.

- Il existe d'autres données, qui vont permettre de résoudre des questions qui ne pouvaient etre considérées plus tôt. Ainsi les données sur les bilans détaillés des banques et de leurs contreparties, sur les compositions de portefeuilles des gestionnaires de fonds vont permettre de mieux comprendre les effets d'interaction , leur importance dans l'analyse des risques systémiques. De façon similaire, en combinant les données de sites web, on peut espérer comprendre comment des annonces diffusées sur différents médias interagissent pour influencer les consommateurs ?

Pour répondre à de telles questions, il faut mettre en place de nouveaux modèles et introduire des méthodes statistiques adaptées. Ces méthodes, introduites au cours des 15 dernières années, doivent etre utilisées de façon adéquates en fonction du problème considéré. Elles portent des noms tels que : Lasso, régression "sparse", apprentissage statistique, segmentation, granularité, panel non linéaire avec effets individuels et temporels, compression… (des références sur ces méthodes sont fournies plus bas).

Insistons finalement sur trois aspects :

1. Ces données massives sont souvent de faible qualité. Les traitements préliminaires pour les rendre plus fiables peuvent s'avérer très couteux, limitant l'intérêt de les utiliser.

2. Les méthodologies utilisées doivent avoir un niveau de complexité numérique bien maitrisé. En particulier le nombre d'opérations nécessaires pour traiter *n* données ne doit pas augmenter trop vite avec *n*. De ce point de vue une gestion de portefeuille de type momentum fondée sur un grand nombre d'actifs sera moins couteuse du point de vue numérique qu'une gestion moyenne-variance par exemple.

3. La disponibilité de données en temps réel n'implique pas nécessairement des réponses en temps réel, ces dernières devant tenir compte des personnes auxquelles elles sont destinées. Ainsi la connaissance des risques automobile en continu n'empechera pas les primes d'assurance d'etre ajustées par exemple mensuellement.

Ce cahier de l'Institut Louis Bachelier fournit des exemples de questions et méthodes liées aux mégadonnées : analyse de la liquidité à partir de données haute fréquence, compréhension des interconnections entre banques à partir de données de bilan, utilisations potentielles des régressions "sparse", etc.

Christian Gouriéroux

Lectures sur les mégadonnées

- Beath,C., Becerra-Fernandez, I., Ross, S., et T., Short (2012) : "Finding Value in the Information Explosion", MIT Sloan Management Review.
- Mayer-Schonberger,V., et K., Cukier (2013) : "Big Data: A Revolution that Will Transform How We Live, Work and Think", John Murray.
- Nichols,W.(2012) : "Advertising Analytics 2.0", Harvard Business Review.

Lectures sur les nouvelles approches statistiques

- Gagliardini, P., et C., Gourieroux (2014) : "Granularity Theory", à paraître Cambridge University Press.
- Gagliardini, P., Gouriéroux, C., et M., Rubin (2013) : "Positional Portfolio Management", CREST DP.
- Hastie, T., Tibshirani, R., et J., Friedman (2009) : "The Elements of Statistical Learning, Data Mining, Inference and Prediction", 2nd ed., Springer.
- Novicki,K., et T., Snijders (2001) : "Estimation and Prediction for Stochastic Blockstructures", J.Amer.Statist. Assoc.,96,1077-1087.
- Tibshirani,R. (1996) : "Regression Shrinkage and Selection via Lasso", JRSS B, 58, 267-288.

SERVICE ABONNEMENT

Si vous souhaitez vous abonner aux Cahiers Louis Bachelier , merci de contacter l'équipe de l'Institut Louis Bachelier par courrier électronique à l'adresse suivante : **contact@institutlouisbachelier.org**

Veillez à bien préciser l'objet de votre message ainsi que vos coordonnées complètes; vous recevrez ainsi chaque numéro par courrier postal à l'adresse communiquée.

NB : Le service abonnement vous est proposé gratuitement cependant chaque édition des Cahiers Louis Bachelier est limitée !

Big data : quels enjeux pour la recherche et les industriels ?

BIOGRAPHIE

Jean-Michel Lasry

Jean-Michel Lasry est conseiller scientifique au Crédit Agricole - CIB (anciennement CALYON) et Président du Comité de pilotage de la chaire "Finance et Développement Durable". Il était auparavant membre du Comité Exécutif des marchés de capitaux ainsi que responsable de la recherche quantitative de CALYON pendant 4 ans. Il a également été Directeur général adjoint de la banque CPR à Paris pendant quatre ans. De 1996 à 1999, Jean-Michel Lasry a été responsable de la recherche quantitative et capital manager des activités de marchés de Paribas. De 1994 à 1996, il a été directeur général de la Caisse Autonome de Refinancement (CDC). De 1990 à 1993, il a été membre du Comité exécutif de la direction des affaires bancaires et financières de la CDC, ainsi que membre du Conseil d'Administration de CDC Gestion. Jean-Michel Lasry a été professeur à l'Université Paris-Dauphine et à l'École Polytechnique pendant 17 ans. Il a publié plus de 100 articles scientifiques dans des revues d'économie et de mathématiques et a dirigé 25 thèses de doctorat.

En novembre 2013, l'université Paris Dauphine et Havas se sont associés pour donner naissance à la Chaire "Economie des nouvelles données" au sein de l'Institut Louis Bachelier. Ce programme de recherche regroupe sponsors industriels et experts scientifiques afin de répondre aux défis économiques et scientifiques que représente le big data. Pourquoi l'arrivée de ces données marque-t-elle une rupture ? Les chercheurs disposent-ils de techniques adaptées à l'analyse de ces informations ? Quelles sont les opportunités offertes par le big data ? Jean-Michel Lasry, à l'initiative de la Chaire, revient sur ces différentes problématiques.

Jean-Michel Lasry, le sujet du big data est devenu très médiatique ces dernières années. En quoi ce phénomène est-il nouveau ?

Le big data marque une rupture sur plusieurs points. D'abord en termes quantitatifs. Les volumes de données disponibles et les flux de création de nouvelles données sont supérieurs de plusieurs ordres de grandeur à ceux que nous observions à la fin des années 1990. Ces informations sont ensuite largement accessibles en temps réel, ce qui contraste avec le passé, même récent pour une grande partie des données. Ensuite, elles sont de natures très diverses. Elles regroupent des choses aussi différentes que la géolocalisation massive liée à la démocratisation des objets connectés, l'enregistrement détaillé horodaté et généralisé des consommations individuelles, via les tickets de supermarché par exemple, ou encore le suivi en continu des constantes biologiques permis notamment par les capteurs de rythme cardiaque. Toutes ces mesures sont souvent liées à l'apparition de nouveaux outils. En parallèle, les capacités de stockage et de calcul ont fortement augmenté, tout en devenant plus accessibles en termes de coûts. Pour résumer, le big data représente une révolution par l'ampleur des données disponibles et par la démocratisation d'outils de mesure, de stockage et d'analyse.

Cette masse de données disparates provient de sources multiples. Les professionnels ne sont plus les seuls producteurs d'information…

Effectivement. Il ne s'agit plus uniquement de résultats d'études au process bien établi. Désormais, les données surgissent d'une multitude d'internautes, via leur participation à des blogs, à des réseaux sociaux… Leur moindre clic est enregistré par le site visité mais aussi par les cookies intégrés dans les navigateurs. Les données naissent aussi de toutes sortes d'objets connectés : depuis les smartphones, à certaines raquettes de tennis qui enregistrent les mouvements durant une partie, en passant par les stations météo personnelles connectées.

Enfin, les actes de la vie économique sont systématiquement enregistrés qu'il s'agisse de flux financiers, d'échanges commerciaux ou de simples actions de prospection.

> *"Le big data représente une révolution par l'ampleur des données disponibles et par la démocratisation des outils de mesure, de stockage et d'analyse."*

Cette arrivée massive de données n'a-t-elle rien de comparable avec des expériences passées ?

Disons que nous avons connu des phénomènes précurseurs. Les données recueillies par les biologistes en génétique, par les spécialistes de traitement linguistique, puis par les spécialistes de traitement d'images, ont fourni un avant-goût des problématiques rencontrées aujourd'hui. Parmi les exemples les plus emblématiques, on peut citer l'analyse du génome, l'imagerie médicale ou encore la détection de spam sur le web.

Pour répondre à toutes ces questions, les spécialistes de l'apprentissage, des statistiques et des sciences informatiques ont dû définir des techniques nouvelles, bien éloignées de celles traditionnellement utilisées. Toute une discipline s'est ainsi constituée au cours des deux dernières décennies afin de proposer des méthodes quantitatives que l'on peut regrouper sous la bannière "l'apprentissage statistique" ou le "machine learning"[1].

Sur quels principes ces méthodes reposent-elles?

Il s'agit de définir des algorithmes complexes afin de rechercher systématiquement des structures permettant d'extraire de l'information. Ce travail est mené dans un contexte où la grande dimension apparaît a priori comme décourageante. Une idée sous-jacente à ces méthodes quantitatives est la notion de parcimonie. Elle postule que les objets d'intérêts admettent une représentation parcimonieuse, c'est-à-dire qu'ils peuvent être représentés à l'aide d'un nombre limité de variables. Trouver de manière effective ces variables, via des algorithmes efficaces, devient l'enjeu de la statistique en grande dimension, discipline en pleine expansion.

Au-delà des questions mathématiques et informatiques, quels sont les principaux enjeux du big data pour la recherche ?

L'arrivée de cette multitude de données ouvre la voie à de nouvelles recherches, et ce dans de nombreux domaines. Il peut s'agir d'études sociologiques sur les réseaux sociaux par exemple, ou encore d'études microéconomiques sur la consommation via l'analyse des tickets de caisse… Les champs des possibles sont très vastes et concernent toutes les disciplines.

Les données massives n'ont de valeur que si nous leur donnons du sens. Il s'agit à la fois d'un travail technique, algorithmique, mais aussi d'une démarche de modélisation en fonction du contexte, des usages existants ou à inventer.

Les entreprises sont également de plus en plus nombreuses à se pencher sur ce sujet. Comment ces nouvelles données impactent-elles leur activité ?

Ce phénomène bouleverse les relations entre producteurs et consommateurs. Les marques peuvent désormais connaître leurs clients sans passer par un intermédiaire, grâce aux réseaux sociaux notamment. Elles ont ainsi la capacité d'établir une relation individualisée avec chacun d'entre eux, et de renforcer leur image par des moyens beaucoup plus complexes que la publicité classique. Grâce à une application Facebook, Warner a par exemple désormais une image beaucoup plus fine des goûts cinématographiques de ses spectateurs, et peut construire une relation plus étroite avec eux. Autrement dit les technologies actuelles offrent la possibilité de construire des CRM d'un type complètement nouveau.

C'est dans ce contexte qu'a été créée en novembre 2013, au sein de l'Institut Louis Bachelier, la Chaire "Economie des nouvelles données" Havas-Dauphine. Quels sont ses objectifs?

La Chaire se veut pluridisciplinaire et transversale et a pour but de faciliter l'accès des chercheurs en économie et en gestion au travail sur le big data. Les chercheurs en économie et en gestion de Dauphine, et plus généralement de PSL, pourront ainsi établir plus facilement des coopérations scientifiques avec des chercheurs spécialisés en apprentissage statistique et en machine learning. Soutenue par plusieurs sponsors industriels, la Chaire mènera des études à la fois sur des thèmes théoriques, comme la mise au point de nouvelles méthodes algorithmiques et statistiques, et sur des domaines plus pratiques de recherche appliquée liés à l'activité de ses partenaires. Dans ce dernier cas, une partie de ces recherches tourneront probablement autour du thème de la relation client. Nous souhaitons, grâce à cette initiative, susciter des échanges entre les professionnels et les chercheurs pour conjuguer les savoir-faire et les expertises. Il s'agit d'apporter rapidement des réponses aux mutations en cours. Par la qualité et la diversité des compétences dont elle dispose, la France peut avoir l'ambition d'être un des grands pôles d'excellence dans le monde sur les problématiques big data.

1. *Discipline scientifique concernée par le développement, l'analyse et l'implémentation de méthodes automatisables qui permettent à une machine (au sens large) d'évoluer grâce à un processus d'apprentissage. Wikipedia*

À retenir

- Le big data marque une rupture tant par la quantité que par la nature des données disponibles.
- Les professionnels, comme les instituts de sondage par exemple, ne sont plus les seuls à produire des données. L'activité des internautes ainsi que l'utilisation d'objets connectés créent une masse d'information très hétérogène.
- L'arrivée de ces informations ouvre la voie à de nouvelles recherches, et ce dans la plupart des disciplines scientifiques.
- Pour les entreprises, le big data constitue une chance de repenser la relation client.

Faut-il imposer de la transparence au marché obligataire ?

Permettre aux investisseurs d'accéder plus rapidement à des informations financières, voici l'un des enjeux du Big Data. Faut-il encore que les acteurs se montrent coopératifs en transmettant leurs données. Or, le marché obligataire fonctionne aujourd'hui de façon assez opaque, préférant les échanges de gré à gré à l'utilisation de plate-forme publique. Cette organisation est-elle efficiente ? Garantit-elle une bonne liquidité ainsi qu'une juste formation des prix ? Bruno Biais et Fany Declerck se sont penchés sur ces questions.

D'après l'article de Bruno Biais et Fany Declerck "Liquidity, Competition & Price Discovery in the European Corporate Bond Market" et un entretien avec Fany Declerck.

Trouver le cours d'une action est aujourd'hui facile. Les informations concernant l'évolution des échanges et les dernières transactions réalisées sont à portée de clic. L'exercice s'avère plus compliqué pour les obligations. Et pourtant en Europe, le marché obligataire brasse deux fois plus d'argent que le marché boursier. Bien que considérables, ces échanges de dette d'entreprise s'effectuent généralement dans la plus grande opacité. Les bases de données en ligne, comme celle de Bloomberg par exemple, se sont développées ces dernières années et permettent d'augmenter la quantité d'information disponible. Pour autant, elles restent relativement peu utilisées. La majorité des transactions continue de s'effectuer de façon traditionnelle, via des échanges téléphoniques entre acheteurs, vendeurs et dealers, confinant les données à un cercle restreint d'initiés. Une telle organisation permet-elle une liquidité suffisante ? Assure-t-elle une bonne transmission des informations ainsi qu'une juste formation des prix ? Autant de questions auxquelles l'étude de Bruno Biais et Fany Declerck apporte des éléments de réponse.

Calibrer l'émission obligataire pour répondre aux attentes du marché

Sur la base d'un échantillon de transactions effectuées entre 2003 et 2005, les auteurs ont étudié le système de négociation d'obligations d'entreprises. Ils se sont plus particulièrement penchés sur le marché secondaire (revente de titre). Il apparaît que le niveau d'échange d'un titre varie selon plusieurs éléments. Tout d'abord, la taille de l'émission : plus elle est importante, plus le trading sera actif. Ensuite, la maturité. Les obligations d'une durée de cinq ou dix ans sont les plus échangées. De même, la notation joue directement sur la demande des acheteurs. Ainsi, les titres bénéficiant de la meilleure note (AAA) séduisent les investisseurs par leur faible degré de risque. Mais les titres moins bien notés (BBB), donc risqués, sont également très échangés.

Une situation qui s'explique par le niveau d'information fourni par l'entreprise, autre élément clé dans le trading obligataire. Les obligations risquées font l'objet d'une communication plus fréquente, or les investisseurs réagissent à l'information.

Un marché européen liquide…

La transparence du marché influe également sur l'écart entre le prix du vendeur et le prix de l'acheteur. Aux Etats-Unis, une régulation a ainsi été mise en place afin d'obliger les dealers à communiquer l'heure, le prix et la quantité des titres vendus en temps réel. Elle a ainsi permis de réduire de 5 à 10% l'écart entre le prix du vendeur et le prix de l'acheteur. Aucune équivalence n'a été instaurée en Europe pour le moment.
Or, étonnement, l'étude montre que le marché européen s'avère plus liquide que le marché américain. Il enregistre globalement plus de transactions pour des frais moins élevés, et au sein de l'Europe, le marché euro est lui-même plus liquide que le marché livre sterling. Les chercheurs interprètent ces résultats comme une conséquence positive de l'intégration économique européenne. En effet, l'Union Européenne, et plus encore la zone euro, favorisent la multiplication des acteurs générant un niveau de concurrence plus fort. Le marché est éclaté, composé à la fois des grandes banques internationales et des établissements nationaux. Les Etats-Unis, pour leur part, fonctionnent de façon moins compétitive avec un nombre restreint de gros acteurs.

> *« Il faut cinq jours pour que toutes les informations post transactions soient répercutées sur le marché. »*

…malgré un manque d'information

La situation européenne n'est pas idéale pour autant. Fany Declerck et Bruno Biais mettent en avant une défaillance dans le processus de découverte des prix. Ils constatent que le jour de la transaction, l'information contenue dans les transactions n'est pas intégrée dans les prix proposés par les dealers. Au total, il faut au moins cinq jours pour que l'ensemble des informations soit répercuté. L'ajustement du prix est donc retardé. Résultat les investisseurs achètent au mauvais prix.
Malgré une bonne liquidité, le marché obligataire européen souffre donc d'un manque d'information. Les dealers profitent de cette opacité pour maintenir des coûts élevés et retarder l'ajustement des prix. Une première tentative d'amélioration a été menée en juillet 2011 avec la création de NYSE BondMatch. Cette bourse électronique, dédiée aux obligations d'entreprises européennes, rend accessible les données liées aux échanges de titre. Une avancée « théorique » qui, dans la pratique, ne rencontre pas le succès espéré. Avec un niveau complet de transparence, pré et post-transaction donc, la plate-forme reste pour l'heure sous utilisée.
Une réglementation poussant à plus de transparence, à l'image de celle instaurée Outre-Atlantique, pourrait donc être nécessaire. Une transmission d'information plus fiable et plus rapide devrait permettre aux entreprises émettrices de mieux valoriser leurs obligations. Elle faciliterait aussi probablement la revente des titres sur le marché secondaire, augmentant ainsi la liquidité et donc l'attractivité des obligations d'entreprise. Toutefois, il faudrait préalablement vérifier l'impact d'une telle régulation en comparant un échantillon de transactions soumises à une exigence de communication au reste du marché. Cette expérience ne peut se réaliser sans le soutien des autorités européennes. Elles seules ont la capacité de contraindre les dealers à communiquer leurs données en temps réel.

À retenir

- Le marché obligataire européen est plus liquide que le marché américain. Il est pourtant moins transparent.
- Cette liquidité s'explique par l'intégration économique européenne. Celle-ci a ouvert le marché à un important nombre d'acteurs et a généré une forte concurrence.
- Cependant, le marché souffre d'une mauvaise transmission d'information. Les données liées à une transaction (prix, quantité, heure) mettent plus de cinq jours à être reportées sur le marché. La correction des prix est donc retardée.

Pour aller plus loin...

- *The Microstructure of the Bond Market in the 20th Century, working paper. Bruno Biais, Richard C. Green. Carnegie Mellon University Research Showcase 2007*
- *Transparency and Liquidity: A Controlled Experiment on Corporate Bonds, Michael A. Goldstein, E. Hotchkiss and E. Sirri. 2007, Review of Financial Studies, 235-273.*
- *Dealer Networks, Norman Schuerhoff and Li Dan, working paper*

Retrouvez l'article de Fany Declerck sur www.louisbachelier.org

Méthodologie

Bruno Biais et Fany Declerck se sont appuyés sur les bases IIC et ICMA afin d'analyser un échantillon de transactions effectuées entre 2003 et 2005. L'échantillon est composé de 300 obligations libellées en euro et 300 obligations libellées en livre sterling. Les titres ont des notations allant de AAA à BBB, et ont été émis par des entreprises de différents secteurs (matières premières, biens de consommation et services, industrie, santé, etc). L'échantillon est ainsi comparable à celui utilisé par l'étude américaine (US Trace). Dans le détail, l'étude porte sur 1 844 826 transactions pour lesquelles les chercheurs ont analysé l'heure de la transaction, le prix, la quantité, les caractéristiques et le "dealer code".

Recommandations

- Le marché obligataire européen fonctionne aujourd'hui de façon opaque. Une plus grande transparence pourrait permettre d'améliorer son efficacité et d'ajuster les prix des obligations plus rapidement.
- Pour valider cette hypothèse, il faudrait comparer deux échantillons de transactions : un premier où les informations sont rendues publiques, et un second sans obligation de communication.
- Cette expérience doit être motivée par la Commission Européenne afin d'obliger les dealers à fournir les informations demandées.

BIOGRAPHIE

Fany Declerck

Fany DECLERCK est professeur de finance à la Toulouse School of Economics. Après un master en économétrie et une thèse en finance, elle a obtenu une bourse Marie Curie afin de passer 3 mois au Centre for Studies in Economics and Finance (Université de Salerne). En mai 2013 elle était chercheur invité à Berkeley et en mai 2014 chercheur invité à la Banque de France. Son expérience académique est complétée par une expérience professionnelle dans le privée : avant de rejoindre Toulouse elle était chercheur affilié au sein d'Euronext. Ses recherches portent principalement sur la microstructure des marchés financiers. Son travail empirique repose sur de lourdes bases de données actions et obligations. Elle a publié dans le Journal of Banking and Finance ainsi que dans le Journal of Financial Markets.

La recherche d'une diversification explique-t-elle l'interconnexion bancaire ?

Les banques sont liées financièrement les unes aux autres via les transactions interbancaires. Elles gèrent ainsi leurs besoins en liquidité, mais pas seulement… L'obligation faite aux banques, de fournir leur bilan détaillé, permet d'accéder à de nouvelles données et de tester d'autres hypothèses. L'interconnexion serait-elle aussi un moyen pour les banques de diversifier leur positionnement ? Représente-t-elle uniquement un facteur de risque ou contribue-t-elle au bon fonctionnement du marché bancaire ?

D'après un entretien[1] avec Jean-Cyprien Héam et Erwan Koch et l'article "Diversification and Endogenous Financial Networks", co-écrit par Jean-Cyprien Héam et Erwan Koch, 2014.

Compétition et coopération sont-elles compatibles ? Il semblerait. Les établissements financiers en sont la preuve. Si les banques rivalisent entre elles pour accroître leurs parts de marché, elles nouent également des partenariats, via notamment les échanges interbancaires. Dans ce contexte paradoxal, la faillite d'une banque est à la fois une bonne et une mauvaise nouvelle pour ses homologues. La première justification donnée à cette interconnexion est la liquidité. Les transactions interbancaires permettent à chaque établissement de gérer ses risques à court terme et de faire face à ses créances. La littérature est abondante sur ce point. Cependant, limiter les relations entre banques à cette problématique serait réducteur. D'autres facteurs doivent être considérés.
Banques, comme compagnies assurances, peuvent décider de se rapprocher pour créer un produit commun, transférer un risque (dans le cadre de la réassurance par exemple) ou diversifier leur positionnement. C'est ce dernier point que Jean-Cyprien Héam et Erwan Koch ont souhaité approfondir : La recherche d'une diversification est-elle une explication valable au fonctionnement en réseau des institutions financières ? Faut-il s'inquiéter de cette interconnexion et mieux la contrôler ? Ou contribue-t-elle au bon fonctionnement du marché ? Pour mener leur étude, les chercheurs se sont appuyés sur de nouvelles données mises à disposition par le régulateur. En effet, les banques sont tenues de communiquer chaque trimestre leur bilan détaillé. Une obligation qui deviendra bientôt hebdomadaire pour les établissements les plus importants. L'arrivée de ces informations ouvre la voie à de nouvelles recherches, à l'image de celle présentée dans cet article.

Bénéficier du positionnement de ses concurrents

L'interconnexion des banques vient d'abord de l'organisation du marché. Ainsi, tous les établissements ne suivent pas le même business model : enseignes mutualistes et groupes commerciaux relèvent par exemple de deux logiques distinctes. De même, pour des raisons historiques, certains se sont fortement développés sur un segment spécifique (une zone géographique, un type de clientèle…). Face à ce constat, chaque banque cherche à définir la meilleure stratégie pour optimiser ses investissements. Sachant que l'acquisition d'un nouveau client coûte cher, elle préfère souvent conclure un partenariat avec un concurrent déjà installé plutôt que de mener ce travail de conquête elle-même. Ses choix sont ensuite guidés par un arbitrage entre risque et rentabilité. Plusieurs paramètres vont jouer sur le niveau de diversification et donc d'interconnexion. Les auteurs ont notamment étudié la rentabilité des prêts octroyés par les différentes banques, la corrélation entre les rendements et le poids de la contrainte réglementaire en capital. Il en ressort que plus l'établissement est sensible au risque, plus il recherche la diversification. Celle-ci est un moyen de limiter les écarts de rentabilité et donc de réduire le risque. Au contraire, une institution neutre face au risque sera uniquement guidée par la recherche de profit.

> *"L'interconnexion répond en partie à une démarche d'optimisation des banques"*

Veiller à un bon niveau de régulation

La réglementation a également un fort impact sur le degré d'interconnexion. Les règles prudentielles imposent aux banques de conserver un certain niveau de fonds propres pour chaque investissement réalisé. Les actifs interbancaires n'échappent pas à cette règle. Plus cette contrainte sera forte, plus les établissements réduiront leurs achats d'actions ou d'obligations d'autres banques.

Dans leur recherche, Jean-Cyprien Héam et Erwan Koch insistent sur ce point. Il s'agit d'instaurer le bon degré de régulation qui limite le risque systémique tout en permettant un niveau optimal de prêt à l'économie réelle. Les chercheurs montrent un effet qu'une interconnexion trop forte pourrait générer une contagion. A l'inverse, peu d'interconnexion pénaliserait la stratégie de diversification des établissements et, par conséquent, le fonctionnement du marché bancaire.

Entre diversification et contagion

Les arbitrages des banques s'effectuent donc en fonction de ces divers éléments et de leur connaissance de l'activité de leurs concurrents. Chaque établissement investit ensuite auprès de ses partenaires en pensant que ces liens auront un impact positif sur son activité. Il optimise son bilan en fonction de la situation des autres banques du réseau. Mais traditionnellement, la réglementation appréhende l'interconnexion uniquement sous le prisme du risque. Selon elle, plus les banques sont liées entre elles, plus le risque de contagion est élevé. Le besoin de diversification n'est pas pris en compte.
Ce dernier apparaît pourtant comme une explication vraisemblable à l'interconnexion financière. Un motif valable parmi d'autres. Plusieurs éléments doivent être considérés afin de cerner l'activité interbancaire, des études complémentaires sont ainsi nécessaires. Il est en tous cas intéressant de connaître l'impact des différents motifs de l'interconnexion afin d'évaluer la sensibilité du système bancaire à certains chocs. La compréhension de ces mécanismes doit également guider la définition de la réglementation la plus appropriée.

1. *Les opinions exprimées sont celles des auteurs et ne reflètent pas nécessairement celles des institutions auxquelles ils appartiennent.*

Méthodologie

Jean-Cyprien Héam et Erwan Koch ont construit un modèle où le réseau interbancaire résulte du souhait de diversification des banques. Ce choix dépend d'un ensemble de paramètres dont les auteurs cherchent à identifier l'importance. Ces paramètres sont la rentabilité des prêts octroyés, la corrélation entre les rendements, le poids de la contrainte réglementaire en capital… Dans un premier temps, il s'agit de voir comment un établissement gère ses opérations interbancaires en fonction de sa connaissance des bilans des autres banques. Puis, de comprendre comment l'ensemble du réseau se construit à partir de ce principe d'optimisation individuelle.

À retenir

- L'interconnexion des banques est souvent perçue comme une réponse aux besoins de liquidité des établissements financiers.
- Il existe des interconnexions de long terme ne reposant pas sur des aspects de liquidité. Par exemple, une banque peut chercher à se diversifier. En nouant un partenariat avec un concurrent spécialisé sur un segment, elle s'ouvre à ce segment spécifique.
- Au niveau de chaque banque, l'interconnexion est vue comme un élément positif. Toutefois, sur un plan global, les interconnexions peuvent générer un risque de contagion.

Pour aller plus loin...

- *Financial Stability Board Data Gaps Initiative, 2014 "Senior Supervisors Group Report on Counterparty Data", www.financialstabilityboard.org*
- *Acemoglu, D., Ozdaglar, A., and A. Tahbaz-Salehi, 2013: "Systemic Risk and Stability in Financial Networks", NBER Working Paper 18727.*
- *Elliott, M., Golub, B., and M. Jackson, 2014: "Financial Networks and Contagion", mimeo.*

Retrouvez l'article de Jean-Cyprien Héam et Erwan Koch sur www.louisbachelier.org

Recommandations

- Evaluer les différents motifs de l'interconnexion bancaire est indispensable pour mesurer la sensibilité du système à divers chocs.
- Comprendre ce phénomène permet de guider la réglementation sur l'arbitrage entre diversification et contagion. Ce travail sert également à proposer un cadre d'analyse des nouvelles données collectées par le régulateur.
- D'autres modèles analysant d'autres motifs d'interconnexion doivent être développés pour dresser une cartographie la plus large possible des mécanismes de formation des réseaux financiers.

BIOGRAPHIE

Jean-Cyprien Héam

Jean-Cyprien Héam est économiste au sein de la Direction des Etudes de l'Autorité de Contrôle Prudentiel et de Résolution et doctorant au CREST (Paris). Ses thémes de travail concernent le risque systémique via des analyses en réseau ainsi que le risque de liquidité. Il est diplômé de l'ENSAE et de l'Ecole Centrale de Lyon.

Erwan Koch

Erwan Koch termine sa thèse à l'ISFA et au laboratoire de finance/assurance du CREST. Ses travaux portent sur les risques spatiaux et en réseaux, avec des applications aux extrêmes climatiques ainsi qu'à la contagion en finance. Ingénieur de l'Ecole Centrale de Paris, il est également diplômé d'un master en modélisation mathématique et climatologie de la même Ecole ainsi que du master d'actuariat de l'Université Paris-Dauphine.

Statistique et informatique : l'indispensable coopération

La multiplication des données complexifie le travail de modélisation et d'analyse. Comment trouver l'information pertinente au milieu de ce flot hétéroclite? La puissance des machines suffit-elle à retirer la richesse des données ? Quel est l'apport de la statistique sur ces problématiques ?

D'après un entretien avec Valentin Patilea, responsable du site rennais du Centre de Recherche en Economie et Statistique (CREST).

BIOGRAPHIE

Valentin Patilea

Valentin Patilea est professeur de statistique à l'Ecole Nationale de la Statistique et de l'Analyse de l'Information (Ensai). Après un master en mathématiques à Bucarest et un master en économie mathématique et économétrie à Toulouse, il a obtenu son doctorat en statistique à Louvain-la-Neuve. Il dirige actuellement le site rennais du Centre de Recherche en Economie et Statistique (CREST). Valentin Patilea a publié de nombreux articles scientifiques dans les meilleurs journaux en statistique et économétrie. Il est régulièrement invité pour donner de séminaires et visiter des universités et centres de recherche de renommé, ainsi que pour donner de conférences dans des colloques prestigieux. Valentin Patilea est co-porteur du programe de recherche Nouveux Défis pour les Nouvelles Données.

Face à l'invasion "Big Data", les professionnels se mettent en quête de la méthodologie "magique" capable d'isoler l'information essentielle pour répondre aux questions économiques ou financières qui les intéressent. Car, en elle-même, cette multitude de données ne présente qu'un faible intérêt. Au sein de ce flot d'informations, seul un petit nombre s'avère pertinent. Une base de données n'a donc de valeur et d'utilité que si elle est mise à jour et nettoyée régulièrement. Or, plus les éléments sont nombreux, plus ce travail de sélection et d'analyse s'avère complexe. Alors comment réussir cette démarche ? Comment extraire de la masse les structures, les liens, les causalités ? Selon Valentin Patilea, la solution se trouve dans la combinaison de la statistique et de l'informatique : deux clés qui permettraient de révéler toute la valeur des données.

Adapter les techniques statistiques traditionnelles

Avec la multiplication des données, les analystes se retrouvent face à de nouveaux défis. Valentin Patilea prend l'exemple d'une variable économique ou financière, discrète ou continue, qu'il souhaite étudier à l'aide d'une grande quantité d'information, parfois collectée de manière automatique. C'est le cas, typiquement, des informations récupérées à la volée sur le web et les réseaux sociaux. L'approche habituelle repose sur les modèles statistiques de régression, qui permet de modéliser la relation entre la variable étudiée et les variables explicatives qui résument l'information disponible. Toutefois, les approches classiques, comme les régressions linéaires ou logistiques, peuvent s'avérer inutilisables, tant du point de vue méthodologique que du calcul numérique. Pour cause, un trop grand nombre de variables explicatives, parfois même supérieur au nombre d'individus observés. Il est alors nécessaire d'adapter la modélisation classique à la réalité des données massives.

Réduire la complexité à l'aide du principe de parcimonie

Le problème de modélisation admet parfois une représentation parcimonieuse, c'est-à-dire qu'il suffit d'un petit nombre de variables explicatives parmi celles disponibles pour bien expliquer la variable d'intérêt. Dans ce cas une stratégie se dessine : sélectionner de façon automatique, à partir des données, les variables réellement pertinentes. Le principe de parcimonie est donc compatible avec l'idée que seule une petite partie de l'information contenue dans les données massives est vraiment utile.

Une adaptation simple des techniques usuelles, basée sur l'idée de pénalisation, permet d'apporter une réponse efficace pour les problèmes parcimonieux. Par exemple, pour adapter le critère des moindres carrés, on pourrait ajouter une pénalité proportionnelle au nombre de coefficients non nuls parmi les coefficients de régression afin de contraindre l'algorithme à préférer les représentations parcimonieuses. Cependant la forme d'une telle pénalité n'est pas adaptée à un calcul effectif d'une solution. Plusieurs variantes de cette méthode existent. La plus populaire, le LASSO (Least Absolute Shrinkage Selection Operator) permet d'obtenir une méthode théoriquement performante : avec une grande probabilité, seules les variables explicatives pertinentes sont sélectionnées. De plus cette méthode peut être implémentée de façon simple et efficiente.

> *« La statistique en grande dimension adapte les techniques traditionnelles à l'accroissement des données »*

Le thème de la parcimonie est aussi prometteur pour modéliser des séries chronologiques de grande dimension. Une technique de type LASSO et ses variantes permettent de relever les autocorrélations significatives et ainsi mettre en lumière des interactions temporelles entre les composantes du vecteur observé au fil du temps. Ceci peut servir par exemple pour anticiper des risques de contagion entre les institutions bancaires. Ces techniques statistiques par pénalisation s'appliquent aussi dans le cas de ruptures structurelles où les autocorrélations changent à certaines dates et restent stables entre ces dates. Autrement dit, le concept de parcimonie n'est pas restreint à celui de paramètres nuls, il s'applique aussi à des paramètres constants par périodes de temps.

Résumer le contenu d'une donnée complexe

De nombreuses applications en finance et assurance produisent des données qui peuvent être considérées comme appartenant à une entité d'observation de nature continue, appelée aussi donnée fonctionnelle ou courbe. C'est le cas, par exemple, des courbes de volatilité ou des enregistrements GPS utilisés parfois en assurance. Les progrès technologiques permettent des grilles d'observation de plus en plus fines, permettant ainsi de capter pratiquement toute l'information sur l'entité. Une fois observée, la courbe peut être approchée avec une grande précision par une combinaison linéaire d'un certain nombre, souvent très faible, de courbes élémentaires bien choisies. En ne retenant que les courbes élémentaires et les coefficients de la combinaison pour chaque entité d'observation, la méthode permet, d'une part, de compresser les données et, d' autre part, d'utiliser des modélisations standards.

Une grande partie des techniques statistiques utilisables avec les données massives ont été conçues il y a plusieurs années. Elles ont simplement été adaptées pour répondre aux problématiques de l'accroissement des données. Pour le chercheur, le phénomène "Big Data" actuel ne représente donc pas une rupture scientifique au niveau de la modélisation statistique. Toutefois, l'arrivée massive de ces données renforce la légitimité de cette science. Si l'informatique apporte la puissance de calcul, la statistique fournit les outils d'analyse. D'où l'importance qu'informatique et statistique, parfois opposées, travaillent ensemble. Cependant, comme l'information accroit toujours beaucoup plus vite que la puissance et la capacité des ordinateurs, il est essentiel, avant d'initier une recherche, de définir un protocole d'étude afin de savoir quelle est la question économique ou financière à laquelle on s'intéresse, et quelles variables sont susceptibles d'y répondre.

À retenir

- La statistique répond depuis longtemps aux problématiques d'analyse des données.
- Les techniques se sont simplement adaptées afin de faire face à l'accroissement des informations.
- La puissance de calcul ne peut pas remplacer l'analyse statistique. Les deux sont complémentaires.

Pour aller plus loin...

- *Bühlmann, P., et S.A. van de Geer (2011), Statistics for High-Dimensional Data. Springer, New-York.*
- *Ramsay, J.O., et B.W. Silverman (2005), Functional data analysis, 2nd ed. Springer, New-York.*
- *Rigollet, P., et A.B. Tsybakov (2011), "Sparse estimation by exponential weighting", Statistical Science, vol. 27, 558-575.*
- *Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso", Journal of the Royal Statistical Society, Series B, vol. 58, 267-288.*

Retrouvez l'article de Valentin Patilea sur www.louisbachelier.org

Recommandations

- Chercher de l'information à l'aveugle est inefficace. Il faut préalablement préciser le protocole d'utilisation afin de définir quelles données doivent être conservées.
- Une base de données doit être mise à jour et nettoyée régulièrement.
- Avant de lancer une exploration de données, il est nécessaire de comparer le coût de cette opération et sa rentabilité.

7th Financial Risks **INTERNATIONAL FORUM**

INSTITUT Louis Bachelier

BIG DATA IN FINANCE AND INSURANCE

Paris, March 20 & 21, 2014

CCI Paris Ile-de-France
27, avenue de Friedland – 75008 PARIS

Design by Paul Morgan : www.paulmorgan.fr

Program and online registration
http://risk2014.institutlouisbachelier.org/

Associate partner:

Venue:

With the support of :

LE REGARD DE NOS PARTENAIRES

Le “Big Data” au service de l’Industrie Bancaire

Par Omar Mehdi Roustoumi et Thierry Duchamp

Certaines entreprises utilisent le Big Data depuis plusieurs années, bien avant que le terme Big Data ne devienne à la mode. Citons par exemple Google, Facebook, Twitter ou encore Salesforce.com, qui font partie des précurseurs les plus connus de l’utilisation de cette technologie. S’il semble aujourd’hui que celle-ci a fait ses preuves auprès de grandes multinationales, essentiellement web, qu’en est-il de ses applications et sa valeur ajoutée pour les métiers de la Banque ? Le Big Data peut-il répondre à leurs nombreuses spécificités, de la banque de détail à la banque d’investissement, tout en les aidant à mieux respecter leurs contraintes prudentielles ?

La problématique Big Data telle qu’exprimée par Gartner[1] peut se résumer à l’association de trois propriétés :

- capacité de stockage,
- capacité de calculs,
- faible coût.

Elle repose sur l’utilisation simultanée de plusieurs serveurs dits banalisés ou “grand public”. Ces machines, produites en grandes quantités, sont beaucoup moins chères que leur version haut de gamme, plus puissantes. Etant également moins fiables, le logiciel doit être conçu pour résister aux pannes. Dans la pratique, une solution Big Data permet donc de stocker une grande quantité de données (jusqu’à plusieurs peta-octets), de réaliser un grand nombre de calculs sur ces données, d’ajouter dynamiquement des machines pour augmenter les capacités tout en résistant aux pannes matérielles.

Cette technologie repose sur trois piliers :

- des publications scientifiques régulières, en particulier par Google depuis 2003,
- une validation pratique de cette technologie par son implémentation, par Google encore, pour ses besoins propres,
- l’utilisation de sa version Open Source par de nombreux acteurs, initialement du Web (eBay, Facebook), et maintenant par le système d’information d’entreprise (Saleforces.com).

Les besoins Big Data en banque

Les problématiques de stockage et de traitement de données d’une banque de détail sont très différentes de celles d’une banque d’investissement. La première aspire à mieux répondre aux besoins de ses clients et en attirer de nouveaux. En termes de traitement de données, il s’agit de pouvoir analyser les comportements bancaires de la clientèle afin de mieux comprendre et anticiper ses besoins. En somme, Il s’agit d’analyser les comportements socio-économiques dans le but d’améliorer les stratégies marketing de la banque ainsi que sa relation-client. La banque d’investissement, quant à elle, aspire à accroitre ses résultats en prenant les bonnes décisions d’achat et de vente des différents produits cotés sur les marchés ou échangés de gré-à-gré. Cette prise de décision passe par la maitrise de son exposition aux risques financiers (marché, contrepartie, taux, liquidité, change…).Autrement dit, il s’agit de pouvoir analyser en temps réel les données de marché dont elle dispose afin de maximiser sa rentabilité et minimiser son exposition aux risques.
Depuis déjà plusieurs décennies, les technologies numériques ne cessent de révolutionner l’industrie bancaire. Les perspectives de gain, en finance de marché en particulier, tiennent désormais à la faculté d’analyser un spectre très élargi d’informations financières dans des temps record.

L’analyse ponctuelle comme première utilisation du Big Data en finance

La première utilisation du Big Data est une analyse périodique de données déjà disponibles mais non exploitées. Il s’agit ici d’ajouter un système à l’existant afin d’y dupliquer les données pour analyse. Cette copie est effectuée de façon interne, c’est-à-dire sans confier les données à un tiers, respectant ainsi les obligations de confidentialité. L’intérêt du Big Data est, dans ce cas et pour des coûts plus faibles, de permettre l’utilisation de données auparavant non valorisées. En banque de détail, l’utilisation typique est l’analyse multi canal des clients, pour identifier les offres les plus adaptés et ainsi mieux structurer l’approche commerciale.

Deuxième utilisation du Big Data : la conservation de données

Mais le Big Data n’est pas cantonné à une fonction d’analyse. Il peut tout à fait être le premier et unique détenteur des données : c’est par exemple l’utilisation faite par Facebook, qui depuis 2011 stocke et traite plus d’1,5 million de messages à la seconde en pic et 6 milliards de messages par jour. Une banque peut également conserver l’ensemble de ses données y compris toutes les versions, tout en y ajoutant des informations telles que des horodatages (« timestamps »). L’audit de chaque entrée sera complété par l’enregistrement de tous les accès et actions dans le système. Utilisé de cette façon, le Big Data répond à l’objectif de stocker davantage de données et de les garder en ligne, c’est-à-dire utilisables par les opérationnels, tout en offrant une traçabilité complète.

Les perspectives ouvertes par l’analyse dynamique

Le temps réel est nécessaire pour résoudre les problématiques “front” des établissements financiers. En effet, en environnement front, un très grand nombre de données varient à chaque instant. Or une grande partie de ces données n’est pas utilisée, faute de capacité de stockage et/ou de capacité de traitement ou d’analyse. Une implémentation concrète du Big Data, permet de conserver les données, y compris toutes les modifications dans le temps (versions), tout en permettant l’évolution du format. Des données sont donc ajoutées en permanence avec une grande souplesse. L’adjonction d’un moteur de recherche permet de prospecter efficacement ces données en temps réel, tout comme Google permet de chercher dans tout le web mondial et sait instantanément présenter les 10 résultats importants du moment.
Cette utilisation automatisable par programmation, ouvre de nouveaux horizons : détection de fraude ou optimisation de stratégies de trading. Dans ce dernier cas, l’analyse est à la fois dynamique, pour la prise de décision, et statique pour le back-testing. Mis à disposition d’un opérateur middle office, le Big Data est un moyen d’une puissance incomparable pour la détection d’anomalies car il est alors possible d’accéder à l’ensemble des données de la banque sans aucune limite d’historique. Les capacités utilisées étant alors la recherche « libre », mais aussi l’ensemble des fonctions d’audit.

Pour conclure

Le concept du Big Data ne recouvre pas seulement la problématique de la taille, mais aussi celle du coût et du temps de traitement des données.
Son utilisation permet de répondre à tous les besoins de l’industrie bancaire, avec des temps de traitement réduits (quelques minutes au lieu de quelques heures), à moindre coût (des serveurs banalisés) sur un socle adaptable (des serveurs peuvent être ajoutés). Ce sont ces trois derniers points, temps, coûts et élasticité, qui différencient le Big Data des technologies classiques.
Ainsi, grâce au Big Data, les calculs d’indicateurs de risque et des autres ratios réglementaires seront traités de manière plus efficace ; de plus, l’analyse en temps réel offre de nouvelles opportunités en termes d’arbitrages et d’aide à la décision. À ce titre la réglementation EMIR offrira de nouvelles perspectives au Big Data qui ne manquera pas de produire de précieux enseignements sur le marché OTC, dont la transparence est prévue dès le 12 février 2014.
D’innombrables autres exemples démontrent les perspectives tangibles des applications de cette technologie en banque et dans les autres industries. Une nouvelle ère vient de s’ouvrir où les défis du Big Data seront constamment renouvelés, grâce à l’augmentation exponentielle des données et des capacités de leur stockage et de leur traitement. En 2014, le Big Data est sans doute l’avenir dont les data scientists aspirent modestement aujourd’hui... à écrire la préhistoire.

1. *Originellement puis revue en 2012 : capacité de stockage (Volume) et de traitement (Vélocité) illimités pour tout type de documents (Variété).*

Méthodologie

La mise en œuvre des Big Data dans le secteur bancaire se décline sous plusieurs formes, en respectant certains principes de base alliant performance, rapidité, flexibilité, robustesse, sans limite de volume. Les possibilités offertes par le Big Data permettent de stocker et d’analyser tout type de données, structurées ou non structurées, en statique et en dynamique. Les données à traiter sont soumises à des algorithmes issus de l’ingénierie financière, orientant les capacités de stockage et de traitement des Big Data selon l’exigence des métiers de la banque en matière d’information, d’analyse, d’efficacité et de célérité de la prise de décision. Parce qu’une décision de trading, d’investissement ou de financement ne se fait pas sans une bonne maîtrise de ses risques, les enjeux des établissements financiers vis-à-vis des Big Data touchent un large éventail de métiers, du front au back office et leur permettent de mieux répondre aux exigences prudentielles.

À RETENIR

↗ Le Big Data ne doit pas seulement être vu comme une technologie de remplacement des technologies classiques. Il permet de traiter les données plus rapidement et à moindre coût

↗ Les trois propriétés importantes d’une solution Big Data d’entreprise : coûts plus faibles, meilleure continuité de service, élasticité de la solution.

+ POUR ALLER PLUS LOIN...

- Highly Available Transactions: Virtues and Limitations: Peter Bailis, Aaron Davidson, Alan Feket, Ali Ghodsi, Joseph M. Hellerstein, Ion Stoica, UC Berkeley and University of Sydney (2013).
- Consumer Credit Risk Models via Machine-Learning Algorithms: Amir Khandani and Adlar Kim, Journal of Banking & Finance34 (2010).

Recommandations

- Imaginer des avantages compétitifs et des modèles économiques que permet le Big Data, grâce à des nouvelles possibilités de stockage de volumes et de traitement de données.
- Il faut aussi prendre en compte les obligations réglementaires auxquelles le Big Data permet de se conformer : audit, détection de fraudes (“Rogue trading”), consolidation globale du risque.

LE REGARD DE NOS PARTENAIRES

Les besoins spécifiques de données pour la recherche empirique ?

Par Didier Davydoff

La majorité des recherches en finance publiées dans les revues scientifiques sont des travaux empiriques. La facilité d'accès aux données et leur qualité sont ainsi un facteur de production décisif pour la communauté académique. Dans ce domaine, les Américains sont partis avec un temps d'avance, avec en particulier la base CRSP de cours boursiers produite par l'Université de Chicago et la base Compustat de données fondamentales sur les entreprises cotées. Un temps que les Européens s'attachent à rattraper.

L'Europe a deux différences avec les Etats-Unis. Tout d'abord, les marchés financiers européens restent encore fragmentés. La World Federation of Exchanges recense 16 Bourses membres de l'association en Europe, malgré les regroupements d'entreprises de marché comme le London Stock Exchange, Euronext, OMX ou Deutsche Börse. Aux Etats-Unis, il n'y en a toujours que deux (NYSE Euronext et NASDAQ OMX). Deuxième difficulté spécifique à l'Europe : même si quelques pôles de recherche disposant de ressources imposantes ont émergé dans chaque pays, il reste que le budget moyen à la disposition des laboratoires et des unités d'enseignements européens est en moyenne sensiblement plus faible qu'aux Etats-Unis. La création de IODS (INSEAD OEE Data Services) en 2011 s'est inscrite dans ce contexte.

La plupart des données utiles en finance sont produites par et pour les acteurs du marché, pas pour la recherche académique. Souvent, les données sont accessibles à travers des stations de travail ergonomiques présentant des écrans de visualisation interactifs. Mais au-delà de cette visualisation, le travail de recherche nécessite en général de sélectionner les données pertinentes en utilisant comme critère de sélection toutes les variables, et non pas seulement celles couramment utilisées par les praticiens.

Il faut aussi pouvoir charger des données en masse pour ensuite effectuer les traitements spécifiques à la recherche. C'est pourquoi, chaque fois que possible, il est demandé aux fournisseurs de données de livrer des fichiers à plat, qui seront stockés sur des serveurs accessibles par des moteurs de requête utilisables selon tout critère de sélection. Il faut par ailleurs que les bases de données soient reconnues pour leur qualité. On se posera par exemple les questions suivantes :

- Comment sont traitées les données manquantes ? S'il n'existe pas de cours de bourse sur une valeur un jour déterminé, c'est peut-être qu'il y a eu une suspension de cotation et il ne faut alors surtout pas reprendre le cours de la veille pour combler le cours manquant. A l'inverse, certaines informations ayant une source différente du flux principal doivent être recherchées avant d'être déclarées comme manquantes. S. Ince and R. Burt Porter (2006) ont ainsi montré que 7 % des observations sur les dividendes d'actions américaines dans la base Thomson Datastream Database (TDS) divergeaient de la base académique de référence CRSP.
- La base est-elle exempte du biais du survivant ? Les fonds d'investissement qui disparaissent des bases de données ou les titres qui sont retirés de la cote sont en moyenne moins performants que la moyenne avant de disparaître. L'étude déjà citée a par exemple montré que pour cette raison, la base (TDS) surestimait de 2.40 % la performance moyenne des actions américaines.
- Les classifications sont-elles exactes ? Il est par exemple essentiel de pouvoir identifier sans ambiguité la ligne principale d'actions d'une société donnée, et de ne pas la confondre avec les lignes secondaires.
- Quels contrôles sont effectués sur l'exactitude des données ? Les informations sont-elles simplement renseignées par les acteurs, avec tous les risques d'erreur involontaires ou non que cela comporte, ou bien sont-elles vérifiées systématiquement ?

Pour assurer cette qualité, la sélection des fournisseurs de données est essentielle. IODS a ainsi choisi comme partenaire sur les cours boursiers EUROFIDAI, un institut de recherche créée par le CNRS en 2003. Pour les données de gouvernance et les fusions-acquisitions, les bases proposées ont été construites par des chercheurs, de l'université de Paris Dauphine dans le premier cas, de SKEMA dans le second. Pour les données fondamentales sur les entreprises françaises, c'est ALTARES qui a été sélectionné : en effet ce fournisseur ne se contente pas de charger les informations disponibles auprès des greffes des tribunaux de commerce. Un contact individuel est pris au moins une fois par an avec toutes les entreprises ayant un chiffre d'affaires supérieur à 10 M€, ce qui permet de vérifier les informations, de les affiner et plus généralement de collecter des données au-delà du minimum légal, avec par exemple la composition des comités exécutifs et l'identité des responsables des principales fonctions dans l'entreprise.

L'exigence de qualité ne doit pas empêcher de diversifier les types de données utilisées. Les avancées de la recherche sont souvent née de l'utilisation de nouvelles données, qui auparavant soit n'existaient pas, soit n'étaient pas visibles. Par exemple, à la fin des années 80, les exploitations des données des marchés électroniques – la Bourse de Paris ayant été précurseur dans ce domaine – ont donné lieu aux premières publications de ce qui allait devenir dans les années suivantes un courant abondant de recherche sur la microstructure des marchés. Aujourd'hui, la masse de données grandissante des plateformes électroniques de trading obligataire est peut-être une nouvelle frontière.

C'est du lien entre les bases de données que peuvent naître aussi les innovations de la recherche. Il faut alors soit disposer d'identifiants communs à ces bases, mais ce n'est pas toujours possible s'agissant de fournisseurs de données indépendants voire concurrents ; soit construire, des tables de passage permettant par exemple de passer d'une base de données fondamentales sur les entreprises aux bases boursières.

Méthodologie

Les séries temporelles macrofinancières sur l'épargne peuvent être classées selon plusieurs dimensions :

- La nature économique du support d'épargne. La nomenclature d'opérations de la comptabilité nationale est utile car elle constitue une partition de l'ensemble des opérations financières possibles. Mais sa granularité n'est parfois pas assez fine pour l'analyse. Les informations plus spécialisées comme les statistiques monétaires ou les statistiques des associations professionnelles sont en général cohérentes avec la nomenclature de comptabilité nationale.
- La dimension géographique (le pays ou le groupe de pays)
- Le type de données : encours, flux, diffusion dans la population, performance financière
- La saisonnalité : séries brutes ou corrigées des variations saisonnières
- La devise : monnaie nationale ou convertie en euros ou en dollars

Les métadonnées doivent être documentées, afin de préciser par exemple les méthodes de désaisonnalisation, les méthodes de conversion statistiques et les ruptures statistiques.

À RETENIR

↗ La fragmentation du marché financier européen doit être prise en compte pour la construction de bases de données reconnues en finance.

↗ Les bases de données en finance sont en majorité produites par et pour les acteurs du marché. Elles doivent être sélectionnées, vérifiées, complétées et interconnectées pour répondre aux besoins de la recherche empirique.

↗ La non-correction des erreurs de données peut conduire à des résultats de la recherche empirique complètement faux.

+ POUR ALLER PLUS LOIN...

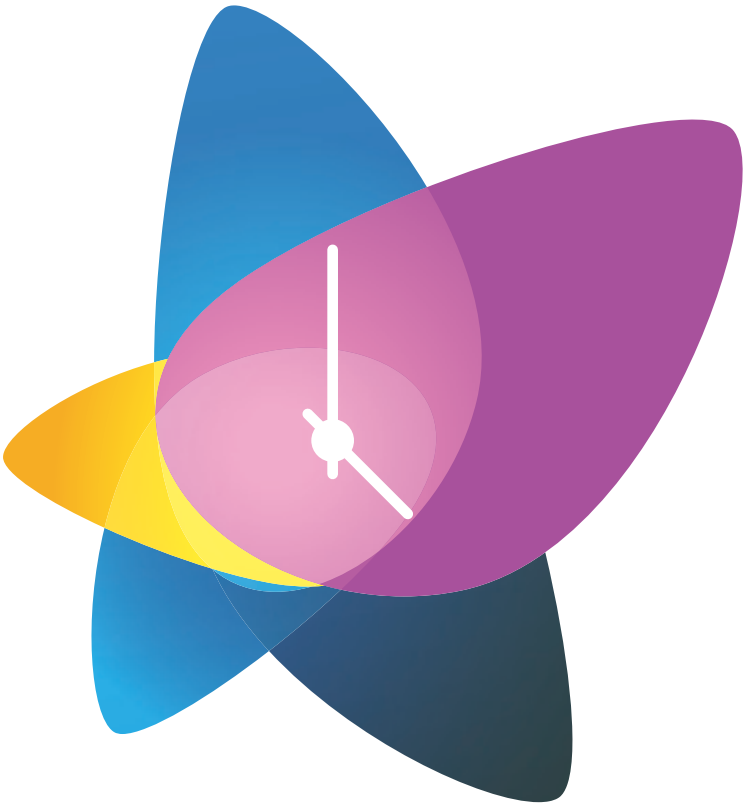
- Ozgur S. Ince and R. Burt Porter (2006) Individual Equity Return Data from Thomson Datastream: Handle with Care ! Journal of Financial Research, Volume 29, Issue 4, pages 463–479
- Laurent Frésard, Christophe Pérignon, Anders Wilhelmsson The Pernicious Effects of Contaminated Data in Risk Management (2011) Journal of Banking & Finance, Volume 35, Pages: 2569-2583
- Roman Brückner, Patrick Lehmann, Martin H. Schmidt, Richard Stehle (2013) Fama/French Factors for Germany: Which Set Is Best? Working paper of the School of Business and Economics at Humboldt University in Berlin

Recommandations

- Lorsqu'une recherche a nécessité la construction d'une base de données spécifique, il est souhaitable d'en mutualiser l'accès à l'ensemble de la communauté académique afin que les résultats de la recherche soient vérifiables, et pour en assurer une mise à jour qui permettra des développements de recherche ultérieurs.
- Les acteurs et les fournisseurs de données du marché devraient s'assurer que les données sont accessibles aux chercheurs.

Journée des Chaires Louis Bachelier

SAVE THE DATE



29 Avril 2014

Palais Brongniart à Paris

la journée DES CHAIRES

INSTITUT Louis Bachelier

FONDATION DU RISQUE Louis Bachelier

Renseignements et inscription sur
www.louisbachelier.org

INSTITUT EUROPLACE DE FINANCE Louis Bachelier