Le phénomène Big Data et les mathématiques

Jean-Michel Loubes

Institut de mathématiques de Toulouse

Traditionnellement, un statisticien manipule des données certes en grand nombre, mais homogènes : elles présentent un même format, ou partagent une structure commune. Or, l'évolution rapide des systèmes d'information gérant des données de plus en plus volumineuses a causé de profonds changements de paradigme dans le travail du statisticien, devenant maintenant expert en données, ou *data scientist*. Caractériser précisément ce que cache l'appellation « Big Data » est une question sophistiquée : ce terme regroupe diverses réalités selon les différents champs d'application. La compétence d'un expert en données, rare encore sur le marché, se dispute aujourd'hui à prix d'or par les employeurs. Le rapport du McKinsey Global Institute (2011, disponible en ligne) a popularisé la caractérisation du Big Data par trois « V » : volume, variété, vélocité. De nombreux domaines des mathématiques ont un rôle à jouer à tous les niveaux : passons en revue quelles compétences sont mobilisées !

Apprentissage : des algorithmes qui doivent maintenant être revisités

Volume : le traitement de grandes masses de données impose une parallélisation des calculs pour obtenir des résultats en temps raisonnable. Cela est d'autant plus vrai lors du traitement en temps réel d'un flux (ou *streaming*) de données. Les acteurs majeurs que sont Google Inc., Amazon.com, Yahoo!, Twitter Inc., Netflix, Facebook, Microsoft Corporation, SAP, Oracle Corporation… développent dans des centres de données (ou *data centers*) des architectures spécifiques pour stocker, à moindre coût, de grandes masses de données (pages Web, listes de requêtes, informations sur les clients ou les articles à vendre, messages…) sous forme brute, sans unité de format.

L'infrastructure dominante, Hadoop (Apache Software Foundation, 2012), pour distribuer et archiver les données, est issue directement des besoins de Google (archiver des pages, compter des occurrences de mots…) et de son système de fichiers. Ce choix impose de repenser complètement la façon dont sont écrits les algorithmes : la parallélisation devient le maître mot ! Ceux qui ont entendu parler du modèle de programmation MapReduce (Google, 2004), qui domine actuellement, en savent quelque chose…

Tous les secteurs industriels sont touchés. Prenons ainsi la problématique, très générale, de la *régression logistique* : cette technique prédictive vise à construire un modèle permettant de prédire ou expliquer les valeurs prises par une variable cible qualitative (comme l'appréciation d'un client sur un produit) à partir d'un ensemble de variables explicatives qualitatives ou quantitatives (grandeur physique, âge, salaire, chiffre d'affaires…). La médecine, le secteur bancaire, l'économétrie, les assurances, le marketing, les sciences sociales, entre autres, y sont confrontés quotidiennement. Le problème de la régression logistique peut désormais être traité avec des outils stochastiques récents (comme les « classifieurs bayésiens naïfs » ou les « algorithmes du gradient stochastique »), ou des méthodes d'apprentissage directement interfacées avec Hadoop. La plus célèbre d'entre elles est le fameux algorithme des *k*-moyennes (ou *k-means algorithm*), utilisé pour réaliser un partitionnement d'une énorme collection de documents. De la même manière, de nombreux algorithmes et techniques classiques en mathématiques doivent être revisités, approfondis, améliorés.

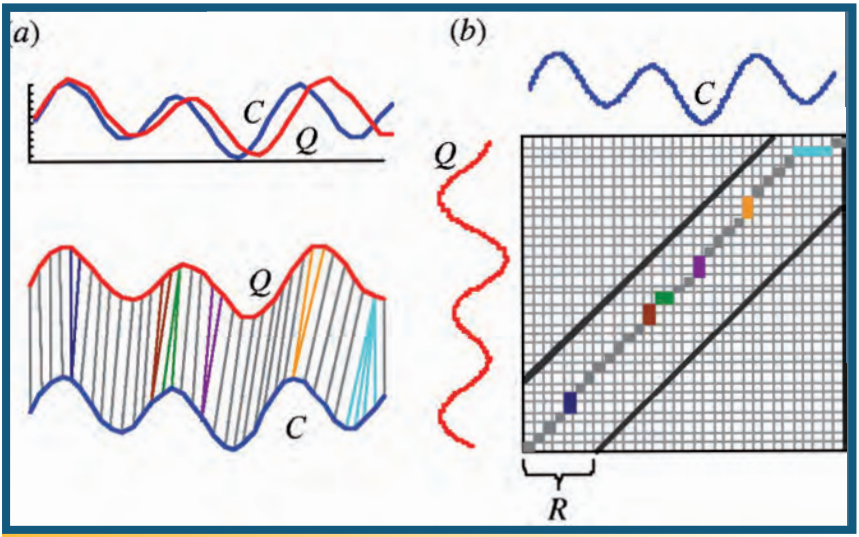
On a besoin d'experts
en statistiques, en algorithmique,
en programmation, en apprentissage et en optimisation !

Classification : vers de nouvelles idées utilisant la géométrie

Variété : le grand nombre de données étudiées entraîne de fait une très grande hétérogénéité. Sous le terme de variété se cache donc de multiples problématiques. Tout d'abord, les données, notamment dans un environnement industriel, peuvent se présenter sous des formes aussi hétérogènes que des textes, des séquences, des fonctions, des courbes, des spectres, des images, des graphes, des tableaux, des sons, des chaînes de caractères… voire des combinaisons de l'ensemble. Au-delà des difficultés techniques inhérentes au traitement d'une telle diversité se pose le problème de l'utilisation de toute cette masse d'informations hétéroclites. Chaque structure de données soulève des questions originales

spécifiques, et comparer ces différentes structures entre elles devient également compliqué à la fois d'un point de vue pratique mais également théorique. En effet, calculer ne serait-ce qu'une simple moyenne (ou des distances entre les objets concernés) requiert une analyse particulière pour intégrer leur géométrie.

On a besoin d'experts en géométrie !



Une utilisation en anthropologie pour quantifier les différences et les similarités entre deux crânes de gorilles :
(*a*) deux séries temporelles similaires mais déphasées ;
(*b*) pour aligner les deux séries, une matrice de recalage est construite pour rechercher le chemin de recalage optimal (en carrés colorés).

© Li Wei, Eamonn Keogh, Xiaopeng Xi et Sang-Hee Lee, 2007

D'autre part, les données sont observées avec une grande variabilité, qui masque bien souvent la nature du phénomène étudié. Il importe dès lors de révéler la structure contenue dans ces observations afin d'en extraire l'information qu'elles contiennent. Ainsi, pour calculer la distance entre deux courbes, celle usuelle (la « distance dans les espaces L^2 ») est beaucoup trop sensible à de légères déformations ou décalages des courbes. Un recalage préalable (ou *time warping*) est nécessaire.

Une autre approche prometteuse (le *scattering*) en analyse d'images consiste à représenter ces dernières dans une base d'ondelettes possédant

des propriétés d'invariance pour certains groupes de transformations (comme Stéphane Mallat et Laurent Sifre l'ont proposé en 2012). Cette représentation permet alors d'identifier ou de regrouper des formes ou des textures d'images particulières.

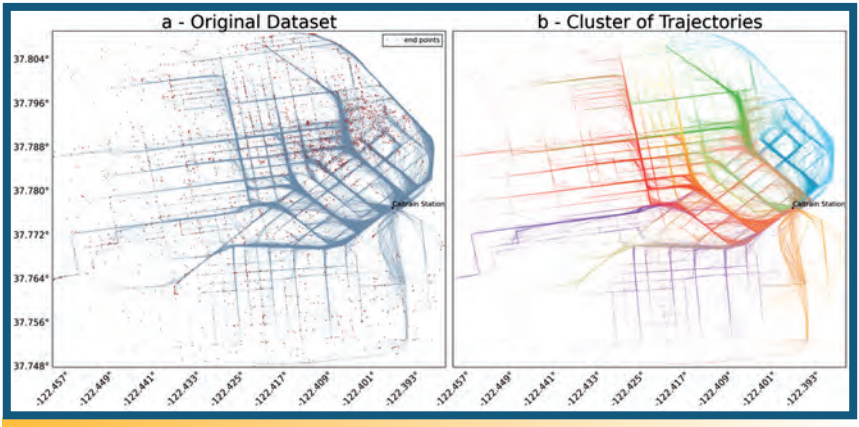
On a besoin d'experts en analyse !

Le dernier exemple concerne initialement l'analyse d'images, conjointement avec des principes de parcimonie. Emmanuel Candès, Justin Romberg et Terence Tao ont montré en 2004 que l'acquisition comprimée (ou *compressed sensing*) permet d'acquérir beaucoup plus rapidement, avec moins de mesures, une très bonne précision des images médicales à partir du moment où celles-ci sont structurées. La reconstruction de l'image fait ensuite appel à la résolution parcimonieuse d'un système linéaire. L'acquisition comprimée a bien d'autres applications, notamment pour l'analyse des très grandes matrices creuses produites en génétique, en génomique, en fouille de textes, en analyse d'incidents dans l'industrie…

On a besoin d'experts en théorie du signal, en analyse d'image et en algèbre linéaire !

Prise de décision : aller au-delà d'une vision séquentielle

Vélocité : la vélocité des données est également motrice de nouvelles recherches sur les algorithmes des méthodes de décision, qui deviennent nécessairement adaptatives ou séquentielles. Comment traiter au mieux



Estimation de flux de populations à partir de données GPS à San Francisco.
© Brendan Guillouet et Jean-Michel Loubes

le flux de données au fil de l'eau ? Le cadre classique de l'apprentissage en ligne ne s'y prête guère : il suppose un échantillon fixé, pour lequel le statisticien a le temps de faire tous les calculs nécessaires afin d'obtenir une règle de décision qui ne changerait pas de sitôt. Cette période est révolue !

Les approches intrinsèquement séquentielles, permettant de s'adapter progressivement aux nouvelles données qui arrivent, connaissent ainsi un vif regain d'intérêt. Pour l'étape d'optimisation conduisant aux règles de décision, c'est notamment le cas de la descente de gradient stochastique (comme l'ont montré Francis Bach et Éric Moulines en 2013). En outre, il convient de prendre en compte le vieillissement des données.

À titre d'exemple, on peut penser aux systèmes de recommandation, comme ceux qui interviennent dans la gestion du contenu des pages Web : le succès fulgurant de Criteo (www.criteo.com) souligne l'importance stratégique de cette problématique. Par ailleurs, l'organisation et le choix des prix de réserves dans les systèmes d'enchères au second prix à grande vitesse (étudiés par Nicolo Cesa-Bianchi, Ofer Dekel et Ohad Shamir en 2013) nécessitent également la mise en production ultra rapide et décentralisée d'algorithmes issus de modèles probabilistes avancés qui rappellent les problématiques du *trading* à haute fréquence.

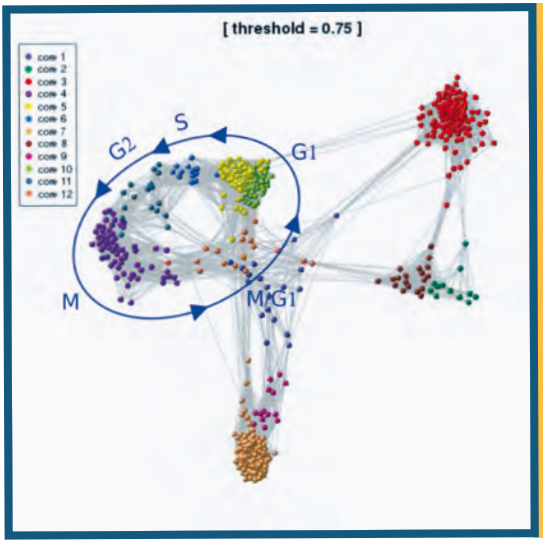
On a besoin d'experts en aide à la décision, en théorie des jeux et en analyse stochastique !

Ce dernier sujet, le *trading*, illustre en particulier le fait que, dans un cadre séquentiel, les décisions prises à un instant peuvent influencer les observations futures, qui du coup ne peuvent pas être traitées comme un échantillon. Ainsi, le système de recommandation ne peut se contenter de proposer les contenus dont l'appétence estimée est la plus grande car, s'il le fait, il se trouvera vite piégé, ne servant plus qu'un nombre très restreint de contenus pas assez diversifiés – les autres n'étant simplement pas assez proposés pour que l'intérêt que leur porte l'utilisateur puisse être perçu.

Toutes les branches des mathématiques sont mobilisées

À l'intérieur même des mathématiques, les contributions tendent à se diversifier. L'apprentissage profond (ou *machine learning*) s'est ainsi beaucoup appuyé sur la modélisation stochastique et statistique des données afin de construire des algorithmes fournissant des règles de

décision pertinentes et afin d'obtenir des garanties théoriques d'efficacité ou d'optimalité de ces méthodes en se fondant principalement sur des outils probabilistes (parfois d'inspiration géométrique). Les gigantesques masses de données et de signaux suscitent un immense intérêt pour l'étude des matrices aléatoires de grande dimension, et en particulier de leur spectre asymptotique, alors que les premiers résultats ont été initiés dans ce domaine par des questions de… physique théorique. Le passage à l'échelle du Big Data rend indispensables et centraux de nouveaux travaux dans le domaine de l'optimisation convexe et, plus généralement, de l'analyse. La modélisation des systèmes complexes, à commencer par les grands réseaux, suscitent des rapprochements avec certains travaux de géométrie, d'algèbre ou de théorie des graphes. Comprendre la géométrie d'un espace euclidien de plusieurs dizaines de dimensions est ainsi devenu un enjeu majeur dans des domaines tels que l'assurance !



Résultat d'un partitionnement de données : estimation d'un graphe d'interaction de gènes pour des mécanismes liés au diabète.

© Anne-Claire Brunet et Jean-Michel Loubes

Le rôle du mathématicien est important pour apporter un changement de perspective, souvent contre-intuitif mais efficace, en élevant le niveau d'abstraction ou en intégrant une approche stochastique. Il y a forcément un domaine d'activité qui saura vous séduire par la façon dont il mobilisera les mathématiques !

J.-M. L.