Sujet de stage: gestion des données manquantes avec données hétérogènes - modélisation des polytraumatisés graves

January 15, 2018

1 Domaine de recherche

Mots clés: données manquantes, décomposition en valeurs singulières, imputation multiple, modèles à variables latentes, complétion de matrices, données hétérogènes.

2 Laboratoire et équipe d'accueil

Le stage se déroulera au laboratoire de mathématiques appliquées de l'école Polytechnique CMAP (http://www.cmap.polytechnique.fr/spip.php?rubrique141). Le CMAP est un environnement dynamique à renommée internationale avec de nombreux étudiants, doctorants et enseignants-chercheurs. L'étudiant sera intégré à l'équipe statistique du CMAP et à l'initiative data-sciences. https://portail.polytechnique.edu/datascience/fr

Collaboration: avec Jean-Pierre Nadal, DR CNRS et directeur d'études au Centre d'Analyse et de Mathématique Sociales (CAMS) à l'EHESS, Laboratoire de Physique Statistique (CNRS-ENS-UPMC-Univ. Paris Diderot) et le groupe Traumabase avec Pr Catherine Paugam-Burtz et Dr Tobias Gauss (Pr des Universités-Praticien Hospitalier, Anesthésie et Réanimation Chirurgicale Polyvalente - Hôpital Beaujon, APHP Hôpitaux Universitaires Paris Nord Val de Seine).

3 Sujet

La problématique des données manquantes est incontournable dans la pratique statistique et pour autant la plupart des méthodes d'analyses ne peuvent pas être mises en œuvre directement à partir de données incomplètes. Cette thématique est en pleine expansion car le problème des données manquantes est exacerbé avec la muliplicité des données collectées qui proviennent souvent de différentes sources d'information. Il est alors crucial de disposer de méthodologies efficaces pour effectuer des analyses en

présence de données incomplètes et surtout de savoir quelle confiance on peut accorder aux résultats obtenus à partir de données partielles.

L'imputation multiple [1] est une méthode de référence pour faire de l'inférence en présence de données manquantes. Elle opère en trois étapes. Tout d'abord, M valeurs plausibles sont générées pour chaque valeur manquante conduisant à M tableaux imputés (complétés). Puis, la quantité d'intérêt θ et sa variance sont estimées sur chaque tableau et les estimations sont agrégées pour obtenir une estimation ponctuelle et un estimateur de la variance qui incorpore la variabilité supplémentaire due aux données manquantes et assure ainsi des taux de couverture des intervalles de confiance au niveau nominal.

Cette approche s'est démocratisée car une fois les données complétées, il est possible d'appliquer toutes les méthodes statistiques que l'on souhaite. Toutefois, les méthodes d'imputation multiple présentent encore de nombreux manques qui sont des champs de recherche actifs. En particulier, il n'existe que très peu de solutions satisfaisantes [2] pour compléter des données avec des variables mixtes, ou en grande dimension.

L'objectif de ce stage est de développer une méthode d'imputation multiple basée sur des ACP généralisées [3]. Ce travail est motivé par les très bons résultats empiriques des méthodes d'imputation pour les variables catégorielles [4] basées sur des décompositions en valeurs singulières pondérées et par les développements récents de modèles à variables latentes pour variables catégorielles [5]. L'idée est d'étendre ce dernier modèle aux variables mixtes en vue de proposer une imputation.

La stage fera l'objet d'une collaboration avec l'APHP (Assistance Publique - Hôpitaux de Paris) sur l'analyse et la modélisation de la prise en charge de patients traumatisés graves. Quand un patient arrive aux urgences, il y a une succession de décisions importantes qui sont prises notamment sur la gravité de son état et qui conduisent à lui apporter une attention plus ou moins importante et immédiate. Bien entendu, des erreurs de diagnostic à ces étapes peuvent être dramatiques. En se basant sur une très grande base de données extrêmement incomplète (constituée par la réunion de registres de plusieurs centres en Ile de France) qui détaille le parcours des patients et leurs caractéristiques, l'objectif ultime est de développer un modèle d'aide à la décision afin d'orienter et de supporter les urgentistes qui doivent établir les actions à mener pour le patient dans un très court délai. Les avancées méthodologiques réalisées pour la gestion des données manquantes sont indispensables pour pouvoir répondre de manière opérationnelle aux problématiques soulevées par la base Traumabase. Nous nous focaliserons en particulier sur l'élaboration de règles de décision par régressions logistiques avec données manquantes. Ces avancées seront donc réalisées en parallèle et des allers-retours fréquents entre théorie et applications sont envisagés. Cette collaboration a déjà été entammée lors de stages de Master 2.

4 Profil recherché

- étudiant de **M2 Statistiques, Probabilités, et/ou Apprentissage, data-sciences**

- fortes compétences mathématiques
- bonne connaissance des algorithmes EM et des méthodes de complétion de matrices
- connaissance du logiciel R est un plus
- intérêt pour les applications en santé
- la poursuite en thèse est envisagée

5 Contact

Le stage sera encadré par Julie Josse et Jean-Pierre Nadal.

julie.josse@polytechnique.edu
jpnadal@ehess.fr

Références

[1] Little, R.J.A & Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* John Wiley & Sons.
[2] Murray, J. & Reiter, J. (2016). Multiple imputation of categorical and continuous via bayesian mixture models. *Journal of American Statistical Association.*
[3] Allen, G.I., Grosenick, G. & Taylor, J. (2014). A Generalized Least-Square Matrix Decomposition. *Journal of the American Statistical Association.*
[4] Audigier, V., Husson, F. & Josse, J. (2015). MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing.*
[5] Fithian, W. and Josse, J. (2017) Multiple Correspondence Analysis & the Multilogit Bilinear Model. *Journal of Multivariate Analysis.* with multiple correspondence analysis. *Statistics and Computing.*