Imputation multiple par régression non-paramétrique

Vincent Audigier, Feriel Bouhadjera, Ndèye Niang

04 novembre, 2022

1 Mots clés

Données manquantes, imputation multiple séquentielle, régression non-paramétrique

2 Responsables du stage

- Vincent Audigier (CEDRIC-MSDMA, CNAM)
- Feriel Bouhadjera (CEDRIC-MSDMA, CNAM)
- Ndèye Niang (CEDRIC-MSDMA, CNAM)

3 Motivations

Les données manquantes constituent un problème fréquent dans la pratique de la statistique. La raison étant que les méthodes d'analyse ne sont généralement pas adaptées pour être mises en oeuvre sur des données incomplètes. L'imputation multiple constitue un des moyens efficaces pour gérer cette difficulté (Little and Rubin (2002), Schafer (1997), Audigier (2022)). Son principe consiste à remplacer chaque valeur manquante par plusieurs valeurs plausibles, conduisant à l'obtention de M tableaux complétés. Une fois ces tableaux obtenus, l'analyse souhaitée peut être menée sur chacun des tableaux (e.g. ajustement d'un modèle de régression) conduisant à une collection de M estimations des paramètres d'un modèle. Ces estimations peuvent ensuite être agrégées selon les règles dites de Rubin.

Ainsi, l'imputation multiple met en jeu deux modèles : un modèle utilisé pour compléter les données, appelé modèle d'imputation et un modèle utilisé pour analyser les données, appelé modèle d'analyse. Ces modèles ne sont généralement pas identiques, ne serait-ce que parce que le modèle d'analyse porte souvent sur la distribution d'une variable conditionnellement aux autres (comme dans le cadre d'un modèle de régression), alors que le modèle d'imputation porte nécessairement sur l'ensemble des variables incomplètes.

Si le modèle d'analyse peut être défini a priori, car l'utilisateur sait quelle analyse il souhaite appliquer sur ses données (incomplètes), celui du modèle d'imputation est généralement moins évident. De nombreux auteurs recommandent de choisir un modèle d'imputation faisant un nombre d'hypothèses le plus faible possible de façon à éviter l'introduction de biais dans l'analyse qui sera effectuée ultérieurement (Van Buuren (2012), Schafer (2003)).

Les modèles d'imputation sont généralement répartis en deux familles : les approches par modèle joint, consistant à définir une distribution multivariée sur l'ensemble des variables,

et les approches conditionnelles, consistant à spécifier chacune des distributions des variables (incomplètes) conditionnellement aux autres. Dans un contexte de variables quantitatives, la méthode de référence parmi les modèles joints est l'imputation selon le modèle gaussien multivarié, tandis que l'approche conditionnelle consiste à imputer par régression linéaire chacune des variables. De ce fait, ces approches ont recours à des hypothèses de distribution sur les données et en ce sens, peuvent introduire des biais lors de la phase d'imputation si ces hypothèses n'étaient pas vérifiées.

Ainsi, l'objectif de ce stage est d'investiguer une nouvelle classe de modèles d'imputation basés sur des techniques non-paramétriques, c'est-à-dire n'effectuant pas d'hypothèses sur la distribution des données. Plusieurs travaux en ce sens ont déjà été réalisés. Par exemple, Doove, Buuren, and Dusseldorp (2014) ont proposé des méthodes d'imputation basées sur les forêts aléatoires, tandis que l'on retrouve dans Geraci and McLain (2018) une méthode basée sur la régression quantile, ou encore une approche par plus proche voisins dans Kowarik and Templ (2016).

4 Démarche et mise en oeuvre

Un premier travail de ce stage consistera à proposer une nouvelle méthode d'imputation multiple selon une approche conditionnelle par des techniques de régression non-paramétriques (Tsybakov (2010)). Cette nouvelle méthode sera comparée à d'autres méthodes d'imputation (paramétriques ou non) dans le cadre où le modèle d'analyse consiste en un modèle de régression paramétrique. Dans un tel contexte, on s'attend à des performances comparables aux méthodes d'imputation séquentielles paramétriques par régression dans le cadre gaussien, mais potentiellement meilleures dans les autres cas (liaisons non-linéaires entre variables explicatives notamment).

Un second travail consistera ensuite à aborder le cas où le modèle d'analyse est non-paramétrique. Dans un premier temps, on évaluera la nouvelle méthode d'imputation dans un contexte d'imputation simple, ce qui signifie que le tableau de données incomplet ne sera imputé qu'une seule fois. Une fois le tableau imputé, on appliquera un modèle de régression non-paramétrique. Cette première étape permettra d'apprécier les propriétés de l'estimateur de la fonction de régression selon la technique d'imputation effectuée. Dans un deuxième temps, on effectuera cette comparaison lors d'une imputation multiple. Cette tâche est plus délicate car elle nécessite de définir une stratégie d'agrégation des M estimations du modèle d'analyse. Or, cela n'est pas classique dans un cadre non-paramétrique car les règles de Rubin ne s'appliquent pas. Ce travail consistera donc d'abord à définir une stratégie d'agrégation avant d'évaluer les performances de l'estimateur de la fonction de régression dans ce contexte d'imputation multiple.

Enfin, cette étude sera complétée par une mise en oeuvre pratique sur un jeu de données réelles océanographiques issues de plusieurs campagnes de mesure effectuées sur environ 9000 stations de mesure réparties sur l'océan global et portant sur les concentrations de 10 pigments phytoplanctoniques: Chlorophylle-a Totale, Divynil Chlorophylle-a, Chlorophylle-b, Divynil-Chlorophylle-b, Hexfucoxanthine, Butfucoxanthine, Fucoxanthine, Peridinine, Alloxanthine, Zeaxanthine.

5 Profil

Etudiant en Master 2 ou ingénieur en dernière année dans le domaine des mathématiques, de la biostatistique, de la statistique, ou de la science des données. Un bon niveau en analyse des données, en programmation R ainsi que des capacités à rédiger en Français et en Anglais sont attendus. Un intérêt particulier sera porté aux candidats souhaitant poursuivre en thèse.

Les dossiers de candidatures devront être composés d'un cv détaillé et d'une lettre de motivation mettant en évidence les raisons de la candidature. Ces éléments devront être transmis par mail aux trois adresses suivantes : vincent.audigier@lecnam.net ; feriel.bouhadjera@lecnam.net ; ndeye.niang_keita@cnam.fr

Références

Audigier, Vincent. 2022. "Gestion des données manquantes par imputation multiple." In *Données manquantes*, edited by Anne Gégout-Petit, Myriam Maumy, Gilbert Saporta, and Christine Thomas-Agnan. Editions TECHNIP. https://hal-cnam.archives-ouvertes.fr/hal-03693373.

Doove, Lisa L., Stef van Buuren, and Elise Dusseldorp. 2014. "Recursive Partitioning for Missing Data Imputation in the Presence of Interaction Effects." *Comput. Stat. Data Anal.* 72: 92–104.

Geraci, Marco, and Alexander C. McLain. 2018. "Multiple Imputation for Bounded Variables." *Psychometrika* 83: 919–40.

Kowarik, Alexander, and Matthias Templ. 2016. "Imputation with the r Package VIM." *Journal of Statistical Software* 74 (7): 1–16. https://doi.org/10.18637/jss.v074.i07.

Little, R., and D. Rubin. 2002. *Statistical Analysis with Missing Data*. New-York: Wiley series in probability; statistics.

Schafer, J. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall/CRC.

Schafer, J. 2003. "Multiple imputation in multivariate problems when the imputation and analysis models differ." *Statistica Neerlandica* 57 (1): 19–35. https://doi.org/10.1111/1467-9574.00218.

Tsybakov, A. B. 2010. *Introduction to Nonparametric Estimation*. 2nd ed. Springer Series in Statistics. Springer New York.

Van Buuren, S. 2012. *Flexible Imputation of Missing Data (Chapman & Hall/CRC Interdisciplinary Statistics)*. 1st ed. Hardcover; Chapman; Hall/CRC.