Sélection de variables pour les modèles d'imputation en grande dimension

Audigier Vincent, Matthieu Resche-Rigon

24 octobre, 2022

1 Mots clés

Données manquantes, sélection de variables, imputation multiple séquentielle

2 Responsables du stage

- Matthieu Resche-Rigon (Université de Paris)
- Vincent Audigier (CEDRIC-MSDMA, CNAM)

3 Motivations

La problématique des données manquantes est incontournable en épidémiologie clinique surtout à l'heure de l'exploitation de grandes bases observationnelles. Les techniques d'imputation multiples constituent une des stratégies efficaces pour y faire face. En particulier, les méthodes d'imputation séquentielle (imputations multiples par équations chainées, Van Buuren (2012)) sont très populaires car elles présentent l'avantage d'être très flexibles et peuvent ainsi être appliquées sur différents types de variables. Cependant, ces méthodes nécessitent de spécifier les modèles d'imputation de chacune des variables incomplètes. Or, dès que le nombre de variables excède quelques dizaines, la spécification de ces modèles devient vite complexe, et pourtant nécessaire si l'on souhaite rester parcimonieux ou si le nombre d'observations reste relativement faible.

On pourrait envisager de résoudre ce problème en appliquant des techniques de sélection de variables. La difficulté ici étant que les données sont incomplètes et que ces méthodes ne s'appliquent alors pas directement. Par ailleurs, cette sélection de variables est susceptible d'introduire des biais dans l'analyse des données imputées (problème mieux connu sous le nom de "non-congénialité"). Dès lors, comment mettre en oeuvre une procédure de sélection de variables telle que le lasso, le *stepwise* ou autre, tout en assurant la congénialité qui impose la compatibilité avec le modèle d'analyse et chaque modèle d'imputation ? Récemment, Bar-Hen and Audigier (2022) ont proposé une méthode de sélection de variables pouvant être appliquée en présence de données manquantes sur des jeux de données possédant plusieurs centaines de variables. L'objet de ce stage est de proposer une méthode de sélection automatique des modèles d'imputation basée sur cette méthode de sélection et de la comparer aux méthodes existantes. La procédure d'inférence obtenue sera alors évaluée par simulation.

4 Démarche et mise en oeuvre

Dans un premier temps, il s'agira d'évaluer la procédure en partant de données simulées selon une loi normale multivariée en la comparant à l'état de l'art le plus récent.

- Faire une étude bibliographique sur la gestion de la grande dimension en imputation multiple (séquentielle en particulier). On pourra notamment consulter les travaux récents sur l'imputation par régression régularisée (Zhao and Long (2017), Zahid and Heumann (2019))
- Proposer un plan de simulation permettant d'évaluer la méthode proposée en la comparant à l'état de l'art
- Mettre en oeuvre ce plan à l'aide du logiciel R
- Analyser les résultats

Dans un second temps, on cherchera à développer une méthode de sélection similaire à celle proposée dans Bar-Hen and Audigier (2022) dans le contexte de variables qualitatives. Cette méthode sera alors évaluée dans un contexte d'imputation multiple tel qu'effectuée précédemment dans le contexte de variables quantitatives.

Enfin, la méthodologie sera appliquée sur un jeu de données réelles. On s'appuiera sur le jeu de données de l'étude de cohorte FROG-ICU. FROG-ICU est une étude de cohorte prospective, observationnelle et multicentrique de survivants de soins intensifs qui ont été suivis pendant 1 an après leur sortie. Cette étude concerne 21 unités de soins intensifs médicales, chirurgicales ou mixtes en France et en Belgique. Tous les patients consécutifs admis en soins intensifs avec un besoin de ventilation mécanique invasive et/ou de soutien par des médicaments vasoactifs pendant plus de 24 heures après l'admission en USI et sortis de l'USI ont été inclus. Le principal critère d'évaluation était la mortalité, toutes causes confondues, un an après la sortie de l'unité de soins intensifs. Les paramètres cliniques et biologiques à la sortie de l'unité de soins intensifs ont été mesurés, notamment les biomarqueurs cardiovasculaires circulants suivants : peptide natriurétique N-terminal pro-B type, troponine I ultrasensible, adrénomédulline bioactive et ST2 soluble. L'objectif sera de produire un modèle pronostique final du décès à 1 an en tenant compte de manière optimale des données manquantes.

5 Profil

Etudiant de Master 2 ou ingénieur en dernière années 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 attendues.

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 deux adresses suivantes : vincent.audigier@lecnam.net and matthieu.resche-rigon@u-paris.fr

References

Bar-Hen, Avner, and Vincent Audigier. 2022. "An Ensemble Learning Method for Variable Selection: Application to High-Dimensional Data and Missing Values." *Journal of Statistical Computation and Simulation* 0 (0): 1–23. https://doi.org/10.1080/00949655.2022.2070621.

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

Zahid, Faisal M, and Christian Heumann. 2019. "Multiple Imputation with Sequential Penalized Regression." *Statistical Methods in Medical Research* 28 (5): 1311–27. https://doi.org/10.1177/0962280218755574.

Zhao, Yize, and Qi Long. 2017. "Variable Selection in the Presence of Missing Data: Imputation-Based Methods." *Wiley Interdisciplinary Reviews: Computational Statistics* 9 (5): e1402–n/a. https://doi.org/10.1002/wics.1402.