Méthodes clusterwise pour le traitement de données hétérogènes et incomplètes

Audigier Vincent, Niang Ndeye

17 Mai 2020

1 Préambule

- Université d'inscription : Conservatoire national des arts et métiers
- Ecole doctorale : SMI
- Laboratoire d'accueil : CEDRIC
- Encadrement :
  - Directeur de thèse : Ndèye Niang
  - Co-directeur de thèse : Vincent Audigier

2 Cadre général

Sciences sociales, sciences médicales, sciences du comportement, domaine bancaire, chimiométrie, ... bien rares sont les domaines où les jeux de données sont exempts de données manquantes. Ceci est d'autant plus incontournable à l'ère des données massives où l'augmentation du volume des données augmente mécaniquement la probabilité d'observer des données incomplètes.

Les données manquantes posent un vrai problème pour l'analyse car les modèles ne sont généralement pas prévus pour être ajustés sur ce type de données. Si dans certains contextes on peut parfois se limiter à une suppression des individus incomplets, ceci devient impossible dès lors que le nombre de variables est grand, ne serait-ce que parce qu'aucun individu n'est alors complet.

Parmi les solutions offertes pour gérer ce problème, les techniques d'imputation multiple sont sûrement les plus populaires (Rubin (1987)). Le principe est de compléter plusieurs fois le tableau incomplet selon un modèle (dit modèle d'imputation), puis d'ajuster sur chacun de ces tableaux imputés le modèle souhaité (dit modèle d'analyse) et enfin d'agréger les résultats selon des règles bien définies.

Définir un modèle d'imputation n'est pas trivial, celui-ci doit ajuster correctement l'ensemble des données alors que celles si peuvent être assez hétérogènes : variables de nature différentes, individus structurés en groupes, etc.

Cette hétérogénéité est fréquemment observée en pratique dans de nombreux secteurs. On peut par exemple observer des variables de nature différente dans les enquêtes où les personnes interrogées renseignent à la fois leur catégorie socio professionnelle (variable qualitative) ou leur nombre d'enfants (variable quantitative) ; en agroalimentaire, où on s'intéresse à la fois à différents procédés de fabrication des produits et à des mesures physico chimiques sur ces derniers ; en médecine où les patients sont décrits par différents indicateurs caractérisant les principes actifs administrés ainsi que la dose associée ; on pourrait aussi citer l' agronomie, la psychologie, etc. L'hétérogénéité liée à structure en groupes des individus est elle aussi rencontrée dans un grand nombre d'applications, par exemple en sciences sociales, sciences de l'éducation, environnement, études cliniques, ou plus généralement dans le domaine scientifique.

Le cadre général de cette thèse est d'investiguer une nouvelle classe de modèles pour imputer des données mixtes en grande dimension quand il existe une structure inconnue des individus en groupe.

3 Méthodologie et objectifs de la thèse

3.1 Méthodologie

Des approches d'imputation multiple dites séquentielles (van Buuren (2018)) ont été proposées pour gérer en partie cette hétérogénéité. Pour un tableau donné, elles consistent à imputer chacune des variables selon un modèle essentiellement lié à la nature de celles-ci (régression linéaire pour une variable quantitative, logistique pour une variable binaire, etc). Les variables sont imputées tour à tour fournissant ainsi un jeu de données complété sans avoir eu à définir explicitement la distribution de l'ensemble des variables. Cette opération est répétée plusieurs fois de façon à produire plusieurs tableaux imputés.

Bien que les modèles utilisés pour imputer chacune des variables permettent de gérer l'hétérogénéité liée à la nature diverse des variables, ceux-ci ne sont généralement pas adaptés à l'imputation d'individus structurés en groupe, ou pour la prise en compte d'un grand nombre de variables. Certains travaux ont été récemment proposés pour le cadre où la structure sur les individus est connue et où le nombre de variables est modeste (cf Audigier et al. (2018)) mais ceci n'est généralement pas le cas.

Pour parvenir au développement d'une méthode d'imputation séquentielle dédiée aux données hétérogènes, nous proposons de nous appuyer sur les méthodes de régression typologique ou clusterwise. En effet, en régression, lorsque les individus présentent une structure en groupes inconnue a priori, ces méthodes permettent d'apporter une réponse à travers la recherche simultanée d'une partition des données en classes et le modèle de régression local associé à ces classes. Ces méthodes ont déjà fait l'étude de nombreux travaux, mais n'ont jamais été envisagées à des fins d'imputation multiple. Les premiers travaux relatifs à ces méthodes sont attribués à Späth (1979) selon DeSarbo and Cron (1988). Mais on peut aussi citer ceux de Bock (1969) et Diday (1976) puis ceux de Charles (1977). DeSarbo and Cron (1988) proposent une méthode de régression linéaire typologique fondée sur un modèle de mélange de gaussiennes avec des estimateurs du maximum de vraisemblance et l'algorithme EM. Plus récemment Preda and Saporta (2005) l'ont utilisée dans le cadre de la régression PLS sur données fonctionnelles. Notons aussi les travaux plus récents de Hennig (2002) dans le contexte de la robustesse ou ceux de Bougeard, Niang Keita, et al. (2018), Bougeard, Cariou, et al. (2018), Bougeard, Abdi, et al. (2018) dans un contexte de données multiblocs.

Les méthodes clusterwise précédentes couvrent essentiellement le cas où les données sont de nature quantitative. Pour pouvoir imputer des jeux de données mixtes, il est nécessaire dans un premier temps de développer les méthodes clusterwise pour ce type de données. Par la suite, il sera possible de proposer une méthode d'imputation séquentielle pour données hétérogènes.

3.2 Objectifs

3.2.1 Méthodes clusterwise pour données mixtes

Dans un premier temps, il s'agira de développer des méthodes de régression clusterwise pour des variables explicatives quantitatives et qualitatives. Dans un second temps, on s'intéressera au problème de la classification supervisée, i.e. au cas d'une variable réponse qualitative.

Par ailleurs, les méthodes clusterwise peuvent être rapidement limitées en grande dimension. La difficulté ici étant que les modèles locaux sont rapidement sur-paramétrés quand le nombre de variables est grand. Ce problème est encore plus prégnant en imputation séquentielle car en raison des données manquantes les modèles d'imputation ne peuvent être ajustés que sur une partie des individus. Le problème de la grande dimension est communément rencontré dans les méthodes d'imputation séquentielle mais reste encore ouvert. Des stratégies ad-hoc consistant à sélectionner les variables explicatives des modèles d'imputation sont généralement utilisées. Ces stratégies sont très critiquables car la sélection est effectuée a priori faute de pouvoir gérer les données manquantes. Des travaux récents se sont portés sur l'utilisation de méthode de régression Lasso (Zhao and Long (2016)) ou ridge (Deng et al. (2016)). Il s'agira ici de proposer une approche clusterwise parcimonieuse afin de gérer la grande dimension (Suk and Hwang (2010), Bougeard, Cariou, et al. (2018)).

3.2.2 Imputation multiple séquentielle par méthodes clusterwise

La suite consistera à proposer une méthode d'imputation multiple basée sur les développements précédents. Toutefois, cette application n'est pas directe car les différentes imputations du tableau de données doivent être effectuées selon des paramètres différents afin de refléter leur variabilité. Pour y parvenir, il est classique d'utiliser des approches Bayésiennes, mais les approches bootstrap sont de plus en plus plébiscitées (Audigier, Husson, and Josse (2017), Doove, Buuren, and Dusseldorp (2014) pour citer quelques méthodes récentes) car elles présentent des avantages computationnels (Bartlett and Hughes (2019)) et offrent une solution non-paramétrique pour générer les paramètres du modèle d'imputation, plus simple à développer qu'une modélisation Bayésienne. Ainsi, une autre contribution consistera à s'appuyer sur un bootstrap non-paramétrique pour développer une approche d'imputation multiple clusterwise.

3.2.3 Application

Tous les apports méthodoloqiques précédemment présentés seront systématiquement évalués par simulation en confrontant les méthodes proposées à l'état de l'art le plus récent. Par ailleurs, ils se concluront par une mise en application sur des données réelles provenant de l'Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (ANSES).

4 Echancier

Le travail de thèse se déroulera selon les étapes suivantes

- Etape 1 (3 mois) : effectuer une étude bibliographique détaillée des méthodes de régression typologique dont le but est de dégager des axes de recherches permettant l'extension de la régression clusterwise aux données hétérogènes et de grande dimension.
- Etape 2 (9 mois) : proposer et évaluer des approches clusterwise dans le cas où les variables explicatives sont mixtes en considérant le cas d'une variable réponse quantitative ou qualitative. Publication
- Etape 3 (9 mois) : développer des méthodes clusterwise parcimonieuses pour les données mixtes et les évaluer. Publication
- Etape 4 (9 mois) : développement et l'évaluation de nouvelles méthodes d'imputation multiple séquentielle pour des données mixtes structurées en sous-groupes d'individus. Publication.

Les 6 derniers mois seront consacrés à la rédaction de la thèse.

5 Moyens consacrés

Les moyens tant matériels (informatiques) qu'humains (compétences propres) du laboratoire d'accueil et des autres partenaires (ANSES et CEDRIC/CNAM) seront mis à disposition du doctorant.

Références

Audigier, V., F. Husson, and J. Josse. 2017. "MIMCA: Multiple Imputation for Categorical Variables with Multiple Correspondence Analysis." *Statistics and Computing* 27 (2): 501–18. doi:10.1007/s11222-016-9635-4.

Audigier, Vincent, Ian R. White, Shahab Jolani, Thomas P. A. Debray, Matteo Quartagno, James Carpenter, Stef van Buuren, and Matthieu Resche-Rigon. 2018. "Multiple Imputation for Multilevel Data with

Continuous and Binary Variables." *Statist. Sci.* 33 (2). The Institute of Mathematical Statistics: 160–83. doi:10.1214/18-STS646.

Bartlett, Jonathan W., and Rachael A. Hughes. 2019. "Bootstrap Inference for Multiple Imputation Under Uncongeniality and Misspecification."

Bock, HH. 1969. "The Equivalence of Two Extremal Problems and Its Application to the Iterative Classification of Multivariate Data." *Mathematisches Forschungsinstitut*, 10.

Bougeard, Stéphanie, Hervé Abdi, Gilbert Saporta, and Ndèye Niang. 2018. "Clusterwise analysis for multiblock component methods." *Advances in Data Analysis and Classification* 12 (2). Springer Verlag: 285–313. doi:10.1007/s11634-017-0296-8.

Bougeard, Stéphanie, Véronique Cariou, Gilbert Saporta, and Ndèye Niang. 2018. "Prediction for regularized clusterwise multiblock regression." *Applied Stochastic Models in Business and Industry* 34 (6): 852–67. doi:10.1002/asmb.2335.

Bougeard, Stéphanie, Ndeye Niang Keita, Xavier BRY, and Thomas Verron. 2018. "Current multiblock methods: competition or complementarity A comparative study in a unified framework." *Chemometrics and Intelligent Laboratory Systems* 182: 131–48. doi:10.1016/j.chemolab.2018.09.003.

Charles, Christian. 1977. "Régression Typologique et Reconnaissance Des Formes." PhD thesis, Université Paris IX.

Deng, Yi, Changgee Chang, Moges Seyoum Ido, and Qi Long. 2016. "Multiple Imputation for General Missing Data Patterns in the Presence of High-Dimensional Data." *Scientific Reports* 6 (1). Nature Publishing Group: 1–10.

DeSarbo, Wayne S, and William L Cron. 1988. "A Maximum Likelihood Methodology for Clusterwise Linear Regression." *Journal of Classification* 5 (2). Springer: 249–82.

Diday, E. 1976. "Classification et Sélection de Paramètres Sous Contraintes." *Rapport de Recherche IRIA-LABORIA*, no. 188.

Doove, L., S. van Buuren, and E. Dusseldorp. 2014. "Recursive Partitioning for Missing Data Imputation in the Presence of Interaction Effects." *Computational Statistics & Data Analysis* 72: 92–104. doi:10.1016/j.csda.2013.10.025.

Hennig, Christian. 2002. "Fixed Point Clusters for Linear Regression: Computation and Comparison." *Journal of Classification* 19 (2). Springer Science; Business Media: 249.

Preda, Cristian, and Gilbert Saporta. 2005. "Clusterwise PLS regression on a stochastic process." *Computational Statistics and Data Analysis* 49: 99–108. doi:10.1016/j.csda.2004.05.002.

Rubin, D. 1987. *Multiple Imputation for Non-Response in Survey*. New-York: Wiley.

Späth, H. 1979. "Clusterwise Linear Regression." *Computing* 22: 367–73.

Suk, Hye Won, and Heungsun Hwang. 2010. "Regularized Fuzzy Clusterwise Ridge Regression." *Advances in Data Analysis and Classification* 4 (1). Springer: 35–51.

van Buuren, S. 2018. *Flexible Imputation of Missing Data (Chapman & Hall/Crc Interdisciplinary Statistics)*. Hardcover; Chapman; Hall/CRC.

Zhao, Yize, and Qi Long. 2016. "Multiple Imputation in the Presence of High-Dimensional Data." *Statistical Methods in Medical Research* 25 (5). SAGE Publications Sage UK: London, England: 2021–35.