Régression non-paramétrique sur données incomplètes

Projet de thèse proposé par F. Bouhadjera, V. Audigier et N. Niang

CNAM, Laboratoire CEDRIC, équipe MSDMA

Keywords : régression non-paramétrique, grande dimension, données incomplètes, océanographie, santé publique.

1 Introduction

Nombreuses sont les problématiques de régression nécessitant d'expliquer une variable cible Y à valeurs réelles en fonction d'un ensemble de variables explicatives X_1, X_2, ..., X_p. Il peut, par exemple, s'agir de modéliser le lien entre la concentration en chlorophylle - a dans du phytoplancton, pigment essentiel à la photosynthèse, et d'autres pigments ; modéliser la concentration en formaldéhyde dans des logements, en fonction des habitudes du ménage (indice d'utilisation de produit d'hygiène, indice d'entretien, indice de soins, etc). Ces liaisons sont classiquement modélisées en effectuant des hypothèses sur la distribution de la variable cible. Parmi eux, le modèle linéaire Gaussien est probablement le plus populaire.

Toutefois, en pratique, identifier la distribution adéquate n'est pas toujours évident, et pourtant, ce choix influence grandement les résultats de la modélisation et leur interprétation. Pour s'affranchir de cette difficulté, une façon de procéder est de recourir à des méthodes de régression dites *non-paramétriques*. L'idée est d'apprendre la relation entre les variables directement à partir des données, sans faire d'hypothèses de distribution. Parmi ces méthodes, certaines sont issues du machine learning comme les arbres de régression, les k plus proches voisins et d'autres adoptent un point de vue plus statistique comme la régression à noyau, quantile, spline, etc.

Ces approches sont séduisantes, mais à l'ère des données massives, elles deviennent difficilement applicables en raison des données manquantes. Dans la littérature sur les données manquantes, on retrouve deux principales méthodes pour gérer les données incomplètes : l'imputation multiple et les approches directes. L'imputation multiple [1, 2] 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* (voir Figure 1).

Quant aux méthodes directes, elles consistent à "adapter" la technique de régression pour qu'elle puisse être mise en oeuvre "directement" sur des données incomplètes. Dans un cadre de régression paramétrique, ceci consiste à employer des algorithmes itératifs spécifiques comme l'algorithme *Expectation-Maximization* (EM) ou celui de *Data-Augmentation*.

Malheureusement, ces deux principales méthodes de gestion des valeurs manquantes n'ont pas été conçues pour un cadre de régression non-paramétrique. En effet, les méthodes d'imputation multiple nécessitent de définir une stratégie d'agrégation des M modèles de régression obtenus à partir de chaque tableau imputé, or les règles de Rubin ne s'appliquent plus dans ce cadre. Aussi, les algorithmes itératifs des méthodes directes reposent généralement sur une hypothèse de distribution et sont donc inopérants dans un cadre non-paramétrique. Ainsi, l'objet de cette thèse est de développer de nouvelles approches d'imputation multiple et de méthodes directes pour pouvoir mettre en oeuvre des méthodes de régression non-paramétriques sur des données incomplètes.

2 Objectifs et approches proposées

2.1 Imputation multiple

Dans le cadre non-paramétrique, l'étape d'analyse (voir Figure 1) consiste à estimer une fonction de régression r telle que

$$Y = r(X_1, \ldots, X_p) + \varepsilon$$

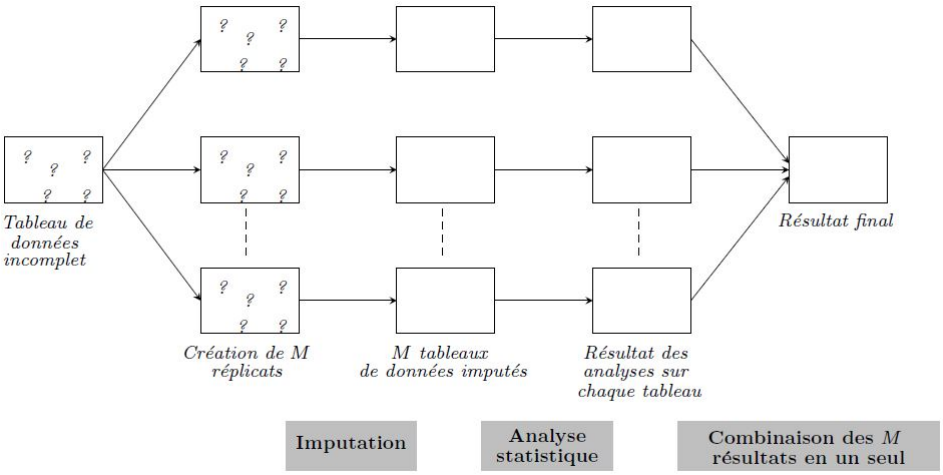

Figure 1: Représentation schématique des trois étapes de l'imputation multiple (de gauche à droite) : imputation, analyse et agrégation

où ε correspond aux termes d'erreur. Dans le contexte de l'imputation multiple, on obtient alors pour chaque tableau imputé, une estimation $\widehat{r}_m$ de r ainsi qu'une variance associée $\widehat{Var}\left(\widehat{r}_m\right)$ évaluée en tout point de l'espace des entrées. Ceci permet de construire des intervalles de confiance autour de chacune des fonctions de régression estimée (cf Figure 2)

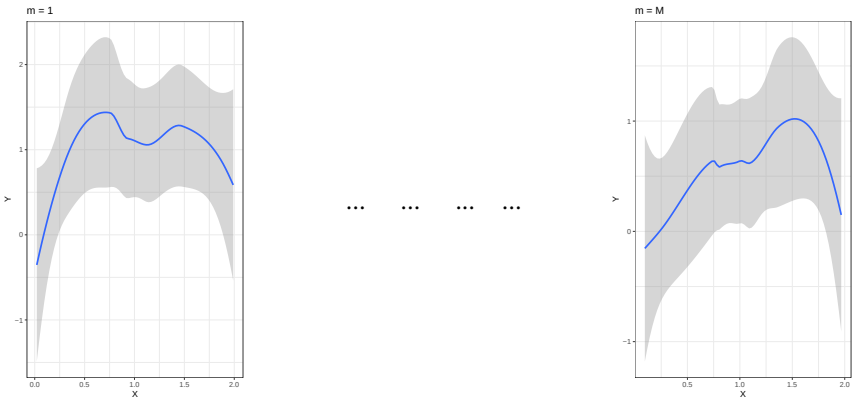

Figure 2: Illustration de l'étape d'analyse dans le cas non-paramétrique

La question de l'agrégation a déjà fait l'objet de différents travaux dans plusieurs contextes, comme par exemple celui de l'analyse factorielle [3], du clustering [4], ou encore des réseaux [5], mais pas de la régression non-paramétrique. Une première extension naturelle des règles de Rubin pourrait être effectuée selon les équations suivantes :

$$\bar{r}(x) = \frac{1}{M}\sum_{m=1}^{M}\widehat{r}_m(x) \tag{1}$$

$$\widehat{Var}(x) = \frac{1}{M}\sum_{m=1}^{M}\widehat{Var}(\widehat{r}_m(x)) + \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{r}_m(x)-\bar{r}(x)\right)^2$$

Notons que [6] propose une agrégation des fonctions selon l'équation (1) sans que celle-ci ait été évaluée. Aussi, l'agrégation des termes de variance reste un problème ouvert.

2.2 Méthode directe

La régression non-paramétrique sur des données complètes, consiste généralement en la minimisation du critère suivant :

$$\arg\min_{r} \sum_{i=1}^{n} \left(Y_i - r(X_{i1}, \ldots, X_{ip})\right)^2 . \tag{2}$$

La solution de ce problème de minimisation peut être obtenue par différentes approches : k-plus proches voisins, noyaux, spline, etc. Par exemple, pour la méthode à noyau, on construit un estimateur $\widehat{r}$ au point $x = (x_1, \ldots, x_p)$ tel que

$$\widehat{r}(x_1, \ldots, x_p) = \frac{\sum_i Y_i K\left(\frac{X_{i1}-x_1}{h_1}, \ldots, \frac{X_{ip}-x_p}{h_p}\right)}{\sum_i K\left(\frac{X_{i1}-x_1}{h_1}, \ldots, \frac{X_{ip}-x_p}{h_p}\right)} \tag{3}$$

où $K : \mathbb{R}^p \longrightarrow \mathbb{R}$ est une densité de probabilité (appelée noyau) et $h = (h_1, \ldots, h_p)$ est un vecteur de fenêtre. Plusieurs travaux ont été initiés pour construire un tel estimateur dans le cadre incomplet (par exemple [7]). Il s'agit de minimiser le même critère (2) sur la base des valeurs observées seulement. Pour cela, on définit $w_{ij} = 0$ si x_{ij} est manquant et $w_{ij} = 1$ sinon, et on réécrit le critère (2) comme suit :

$$\arg\min_{r} \sum_{i} \left(Y_i - r(w_{i1}X_{i1}, \ldots, w_{ip}X_{ip})\right)^2 .$$

La solution de ce problème de minimisation est moins évidente. On pourra envisager de développer un nouvel algorithme d'optimisation à l'image de l'algorithme EM évoqué en introduction. Dans le cadre de l'estimateur à noyau, une première solution pourrait être de remplacer dans l'équation (3) les X_{ij} par :

$$X_{ij}^{\star} = w_{i,j}X_{ij} + (1 - w_{i,j})\mathbb{E}[X_{i,j}|X_{i,-j}]$$

pour $i = 1, \ldots, n$ et $j = 1, \ldots, p$. L'algorithme EM consistera à itérer entre évaluation de $\widehat{r}$ et mise à jour des $X_{ij}^{\star}$.

2.3 Grande dimension

Le principal défaut d'une utilisation directe des estimateurs non-paramétriques d'une fonction de régression est celui du fléau de la dimension. En effet, ces méthodes sont généralement employées pour un petit nombre de variables. Ainsi, pour pallier à ce défaut, une des solutions suggérées dans la littérature consiste à supposer que la fonction de régression r se décompose en une somme de p fonctions r_p définies sur $\mathbb{R}$, c-à-d que pour tout $(x_1, \ldots, x_p)$ de $\mathbb{R}^p$ on a :

$$r(x_1, \ldots, x_p) = \mu + \sum_{j=1}^{p} r_j(x_j).$$

Le modèle est réduit mais reste néanmoins purement non-paramétrique. D'autres techniques, basées sur de la réduction de dimension via des variables latentes, ont été proposées dans le cadre non-paramétrique [8, 9].

On envisagera ce type de techniques pour pouvoir mettre en oeuvre la méthodologie proposée sur des jeux de données comportant davantage de variables.

3 Application

La méthodologie développée pourra être mise en oeuvre dans différents domaines d'application. Dans le domaine de la biologie marine, on pourra s'intéresser à des données phytoplanctoniques issues de 9210 stations du globe [10]. L'objectif sera d'étudier la relation entre la cholorophylle - a, qui est le pigment essentiel à la photosynthèse des phytoplanctons, et une dizaine d'autres pigments. Une première analyse exploratoire de ces données a permis de mettre en évidence le caractère complexe de la liaison entre ces pigments. Aussi, ces données sont très impactées par les problèmes de valeurs manquantes. En effet, ces données sont notamment récoltées par satellite et les mesures ne peuvent pas être effectuées en présence de nuages.

Dans le domaine de la santé publique, on pourra s'intéresser à des données de qualité de l'air intérieur portant sur 516 logements. L'objectif ici sera de comprendre le lien entre la concentration en formaldéhyde et 21 variables caractérisants les habitudes des occupants du logements. L'hétérogénéité des logements (appartements ou maisons par exemple) rend difficile la modélisation de cette liaison et justifie le recours à des approches de régression non-paramétriques. Aussi, ces données sont incomplètes, comme souvent dans les enquêtes statistiques (longueur des questionnaires, difficulté à répondre, etc).

4 Calendrier prévisionnel

Le travail de thèse se déroulera selon les étapes suivantes

- Étape 1 (6 mois) : étude bibliographique détaillée sur la gestion des données manquantes et la régression non-paramétrique.
- Étape 2 (9 mois) : développement d'une méthodologie d'agrégation des modèles de régression non-paramétriques issus d'une procédure d'imputation multiple.
- Étape 3 (9 mois) : proposer un estimateur non-paramétrique d'une fonction de régression sur données incomplètes et étudier ses propriétés, tout d'abord dans le cadre où le nombre de variables explicatives est petit, puis en grande dimension.
- Étape 4 (6 mois) : applications des méthodes proposées aux données phytoplanctoniques et de santé publique.

Toutes les méthodes proposées feront chacune l'objet d'une publication et d'un développement logiciel librement accessible.

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

References

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

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

[3] J. Josse, J. Pagès, and F. Husson. Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification*, 5:231–246, 2011.

[4] V. Audigier and N. Niang. Clustering with missing data: which equivalent for Rubin's rules? *Advances in Data Analysis and Classification*, 2022.

[5] A. Imbert, A. Valsesia, C. L. Gall, C. Armenise, G. Lefebvre, P. A. Gourraud, N. Viguerie, and N. Villa-Vialaneix. Multiple hot-deck imputation for network inference from rna sequencing data. *Bioinform.*, 34(10):1726–1732, 2018.

[6] M. Geraci and A. C. McLain. Multiple imputation for bounded variables. *Psychometrika*, 83:919–940, 2018.

[7] F Bouhadjera, M. Lemdani, and E. Ould Saïd. Strong uniform consistency of the local linear relative error regression estimator under left truncation. *Statistical Papers*, (64):421–447, 2023.

[8] R. Rosipal. Nonlinear partial least squares an overview. *Chemoinformatics and advanced machine learning perspectives: complex computational methods and collaborative techniques*, pages 169–189, 2011.

[9] A. Tenenhaus, A. Giron, E. Viennet, M. Béra, G. Saporta, and B. Fertil. Kernel logistic pls: A tool for supervised nonlinear dimensionality reduction and binary classification. *Computational Statistics & Data Analysis*, 51(9):4083–4100, 2007.

[10] J. Peloquin *et al*. The maredat global database of high performance liquid chromatography marine pigment measurements. *Earth System Science Data*, 5(1):109–123, 2013.