Sujet PostDoc MIAI

Modèle de Markov caché et mélange de distributions : application sur des patients atteints d'apnée du sommeil

Encadrants: Vincent Brault et Adeline Leclercq-Samson (LJK, Université Grenoble Alpes)

1 Présentation de la structure

Le Laboratoire Jean Kuntzmann est un laboratoire CNRS en mathématiques appliquées et informatique, basé au bâtiment IMAG sur le campus universitaire de Grenoble.

Description de l'équipe : Le LJK est composé de 250 membres, son directeur est Jean-Guillaume Dumas.

`https://www-ljk.imag.fr/`.

2 Contexte du projet

2.1 Motivations

Dans de nombreuses applications où on recueille des données au cours du temps chez plusieurs individus, que les données soient continues ou de type comptage, elles peuvent comporter un **excès de zéros**, pour non utilisation de l'appareil ou dysfonctionnement de la transmission des données. C'est par exemple le cas en génomique (voir par exemple Aubert et al. (2014)), pour des données en clinique d'utilisation de dispositifs médicaux, etc.

A titre illustratif, on représente dans la figure 1 des profils d'observance de dispositifs médicaux chez des patients atteints du syndrome d'apnées du sommeil, avec en abscisse le jour d'observation depuis le début de la prise en charge et en ordonnée le nombre d'heures d'observance.

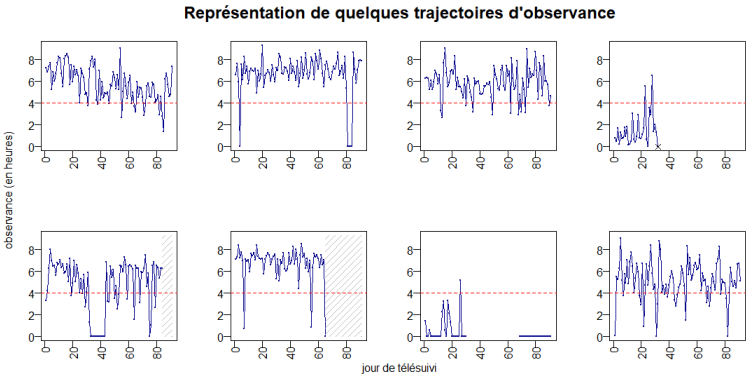

Figure 1: Courbes d'observance de 8 patients: nombre d'heures d'utilisation de l'appareil en ordonnée pour chaque jour (abscisse) depuis le début de la prise en charge médicale.

On peut remarquer la variété des profils d'observance, avec des patients très réguliers, des patients dont l'observance augmente ou diminue, et enfin des patients avec des "zéros", c'est à dire des non observances (l'appareil n'est pas du tout utilisé par le patient).

2.2 Objectifs statistiques

L'objectif du projet est de proposer une méthode de **classification des individus** en fonction de leur comportement temporel. Ces excès de zéros rendent l'analyse statistique difficile. Une approche couramment utilisée dans la littérature est de supprimer les zéros puis d'utiliser des méthodes de classification. Cependant, on comprend aisément que ces zéros sont porteurs d'une information importante et que les omettre fournit une classification qui perd parfois de son sens. Une autre difficulté est que les trajectoires peuvent avoir des longueurs différentes. Un certain nombre de méthodes (K-means, SVM, etc), qui fonctionnent sur des matrices de données (les individus ont tous le même nombre d'observations) ne sont alors plus utilisables.

Notre proposition est d'utiliser une modélisation mélangeant une approche par **Modèle de Markov Caché et mélange de distributions** avec une dirac en zéro pour résoudre simultanément le problème de la prise en compte des zéros et l'objectif de classification. Le modèle est décrit ci-dessous.

3 Modélisation des trajectoires

On note Y_{ij} l'observation du sujet i au temps t_j. On suppose que les trajectoires des individus peuvent être regroupées en K groupes. Pour tenir compte de la présence de zeros (en excès), on modélise l'observance comme un mélange entre une Dirac en 0 et une Gaussienne (voir extensions). Le passage entre ces deux états, pas d'utilisation (0 heure) ou utilisation (autre valeur) de l'appareil, est modélisé par une chaine de Markov (non observée, donc cachée) à deux états.

Le modèle est le suivant pour le sujet i. Soit Z_i la variable indiquant l'appartenance de l'individu i au groupe k.

$$P(Z_i = k) = \pi_k$$

On note $(S_{ij}, j \geq 1)$ la chaine de Markov de l'individu i au temps t_j, à valeurs dans $\{0, 1\}$ correspondant aux deux états {non utilisation, utilisation}.

$$\begin{aligned} P(S_{ij} = 1 | S_{ij-1} = 0, Z_i = k) &= p_{0k} \\ P(S_{ij} = 1 | S_{ij-1} = 1, Z_i = k) &= p_{1k} \end{aligned}$$

Puis on modélise le nombre d'heures Y par le mélange de lois suivant :

$$\begin{aligned} Y_{ij} | S_{ij} = 0, Z_i = k &\sim \delta_0 \\ Y_{ij} | S_{ij} = 1, Z_i = k &\sim \mathcal{N}(a_k + b_k t_j, \sigma_k^2) \end{aligned}$$

Les groupes sont définis par leur probabilité à entrer et sortir de l'état "zero" (p_{0k}, p_{1k}), ainsi que par leur comportement d'utilisation de l'appareil au travers des paramètres de pente (a_{ik}, b_{ik}).

D'autres extensions du modèle seront envisagées.

4 Etapes du post-doc

Le post-doc pourra travailler parmi les directions suivantes :

- estimation des paramètres du modèle par maximum de vraisemblance via le développement d'un algorithme EM (voir par exemple Dempster et al. (1977)) ;
- sélection du nombre de groupes K par critère BIC (voir par exemple Schwarz et al. (1978)) ;

- développement d'un package permettant le passage à l'échelle (programmation en C++, parallélisation) ;
- étude théorique de l'estimateur : identifiabilité, propriétés asymptotiques ;
- extension du modèle en remplacant la loi Normale par la loi Gamma, une loi de comptage (Poisson), des effets aléatoires individuels sur les a_k, b_k ;
- extension à une redondance de clustering en introduisant une autre variable de groupes $W_i \in \{1, \ldots, L\}$ telle que

$$\begin{aligned} Y_{ij}|S_{ij} = 0, W_i = \ell, Z_i = k &\sim \delta_0 \\ Y_{ij}|S_{ij} = 1, W_i = \ell, Z_i = k &\sim \mathcal{N}(a_\ell + b_\ell t_j, \sigma_\ell^2) \end{aligned}$$

- extension à la prise en compte des données manquantes.

5 Compétences attendues

- Savoir calculer et développer une méthode d'estimation.
- Savoir coder un algorithme d'inférence.
- Savoir analyser des données réelles.
- Autonomie.
- Organisation.
- Capacité à communiquer
- Anglais : niveau B1 à B2

Conditions de diplôme : être titulaire d'un Doctorat

6 Contacts

Pour plus d'information sur le poste vous pouvez contacter : Madame Adeline LECLERCQ-SAMSON, Enseignante-Chercheuse en statistique, par mail : `adeline.leclercq-samson@univ-grenoble-alpes.fr` et Monsieur Vincent Brault `vincent.brault@univ-grenoble-alpes.fr`.

7 Pourquoi travailler à l'UGA ?

8 Comment candidater ?

Envoyer dès que possible CV et lettre de motivation avec le rappel de la référence *SO-0521-LJKPOSTDOC* à `vincent.brault@univ-grenoble-alpes.fr`, `adeline.leclercq-samson@univ-grenoble-alpes.fr` et `dgdrh-recrutement@univ-grenoble-alpes.fr`. Les entretiens auront lieu au fil de l'eau en fonction des dossiers reçus.

References

J. Aubert, S. Schbath, et B. Laroche. Metagenomics data analysis using a latent block model: application to plant-microbial communities interactions in the rhizosphere. Dans *ECMTB14-European Conference on Mathematical and Theoretical Biology*, page 1, 2014.

A. P. Dempster, N. M. Laird, et D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

G. Schwarz et al. Estimating the dimension of a model. *Annals of statistics*, 6(2):461–464, 1978.