`fdasynthesis`: Génération de données synthétiques pour l'analyse de données fonctionnelles

Arianna Burzacchi
DMat, Politecnico di Milano, Milan, Italy &
MoST, Fondazione Bruno Kessler, Trento, Italy
aburzacchi@fbk.eu

Aymeric Stamm
Laboratoire de Mathématiques Jean Leray,
UMR CNRS 6629, Nantes Université, France
aymeric.stamm@cnrs.fr

Résumé

Avec l'augmentation de données disponibles et de leur utilisation, la confidentialité des informations sensibles qui y sont associées est devenue un sujet clé. La génération de données synthétiques (GDS) se présente comme une solution possible au problème de la fuite de confidentialité, car elle fournit des jeux de données synthétiques, non associés à aucun utilisateur spécifique, mais tout aussi informatifs. D'autres paquets R ont déjà implémenté des méthodes de GDS, dont `synthpop` (Nowok et al. 2016), pour la gestion de données tabulaires. Nous présentons un algorithme de GDS pour des données non structurées de type fonctionnel et son implémentation dans le paquet R `fdasynthesis`.

Le paquet permet de générer des données synthétiques fonctionnelles cohérentes avec une distribution d'origine, via une approche inspirée de la méthode *k-Nearest Neighbors*. Plus précisément, chaque nouvelle fonction est synthétisée en médiant une observation de référence réelle et les k courbes les plus proches. Afin de garantir la scalabilité et garantir les proprietes d'invariance géométriques, tout en prenant en compte les problématiques d'alignement inhérentes aux données fonctionnelles, la méthode s'inscrit dans le cadre de *Functional Shape Analysis* (Kurtek et al. 2012; Tucker et al. 2013; Srivastava et Klassen 2016). La synthèse des données se fait donc en opérant dans l'espace des *Square-Root Velocity Functions* (SRVFs), où la moyenne est définie comme la *Karcher mean*, en exploitant et en intégrant les fonctionnalités du paquet R `fdasrvf` (Tucker 2024).

Une description plus détaillée du cadre mathématique et méthodologique est disponible dans le manuscrit (Burzacchi et al. 2024). Le paquet `fdasynthesis` est actuellement disponible sur GitHub (Burzacchi et Stamm 2024).

Mots-clefs : Génération de données synthétiques; Analyse de données fonctionelles; Square Root Velocity Functions; `fdasynthesis`

Bibliographie

Burzacchi, A., L. Bellanger, K. Le Gall, A. Stamm, et S. Vantini. 2024. *Generating Synthetic Functional Data for Privacy-Preserving GPS Trajectories*. https://doi.org/10.48550/arXiv.2410.12514.

Burzacchi, A., et A. Stamm. 2024. *fdasynthesis*. Https://github.com/araiari/fdasynthesis.

Kurtek, S., A. Srivastava, E. Klassen, et Z. Ding. 2012. « Statistical Modeling of Curves Using Shapes and Related Features ». *Journal of the American Statistical Association* 107 (499): 1152-65. https://doi.org/10.1080/01621459.2012.699770.

Nowok, B., G. M. Raab, et C. Dibben. 2016. « synthpop: Bespoke Creation of Synthetic Data in R ». *Journal of Statistical Software* 74 (11): 1-26. https://doi.org/10.18637/jss.v074.i11.

Srivastava, A., et E. P. Klassen. 2016. *Functional and shape data analysis*. Springer.

Tucker, J. D. 2024. *fdasrvf: Elastic Functional Data Analysis*. https://github.com/jdtuck/fdasrvf_R.

Tucker, J. Derek, W. Wu, et A. Srivastava. 2013. « Generative models for functional data using phase and amplitude separation ». *Computational Statistics and Data Analysis* 61: 50-66. https://doi.org/10.1016/j.csda.2012.12.001.