Bayesian clustering of sparsely observed tensors applied to surface weather data

Felipe Toledo Ferreira[1]

Instituto de Matemática e Estatística da Universidade de São Paulo

Clustering techniques are commonly applied for the purpose of detecting obscure patterns in data, such as subtle weather and neurological anomalies. Much of the information gathered surrounding these topics has since been characterized as ‘functional data’, due to its visual similarity to mathematical functions when plotted against arguments such as latitude or longitude (in weather analysis), wavelength (in neuroimaging) or time (in energy consumption). Since the inception of Functional Data Analysis (FDA) in [1], there have been several applications surrounding these fields, though there are important assumptions regarding the data under analysis which state that the sampled trajectories are measured without noise at dense and regularly spaced intervals. This greatly limits the study of datasets wherein these assumptions are false, as is the case of the Colorado Monthly Meteorological dataset, which spans the years 1895 through 1997. The weather stations which provide the measurements aren’t evenly distributed throughout the states’ area, and most weather stations were operable for only sections of the period spanned, resulting in an often ‘sparse’ number of measurements per month every year. When standard functional estimation methods are applied to these datasets, the results are prone to error.

Novel techniques have been presented to deal with these circumstances, such as the Principal Components Analysis through Conditional Expectation (PACE) proposed in [2], which is applicable to what is commonly referred to as ‘sparsely observed’ functional data. This approach has since been adapted to the Bayesian context and used to develop tools such as additive models for sparsely observed functional data, as in [3], in which the authors develop two distinct estimation methods steeped in Bayesian Variational Inference and MCMC sampling. In this paper, we propose a novel yet similar approach, focused on the problem of clustering sparsely observed tensor valued functional data. The use of Bayesian Variational Inference for clustering functional data has been previously proposed in [4], though in comparison the novelty of our approach is our focus on the PACE method, as well as the extension toward tensor valued data: functional data analysis is often restricted to trajectories sampled on the one-dimensional interval $[0,1] \subset \mathbb{R}$; in contrast, trajectories sampled on the three-dimensional cube $[0,1]^3 \subset \mathbb{R}^3$ are referred to as ‘tensor data’. We propose approaches based on Bayesian Variational Inference and MCMC sampling toward the issue of clustering sparsely observed tensors, and illustrate the performance of these approaches via simulated data. Ultimately, we apply our clustering proposal to the Colorado Monthly Meteorological (1895-1997) dataset

[1] felipe3@ime.usp.br

and verify its capacity to detect varying weather patterns.

Keywords: Weather data; Spatio-temporal data; Functional data analysis; Functional clustering; MCMC sampling; Variational inference; B-splines.

References

[1] Ramsay, James and Dalzell, C. J: *Some Tools for Functional Data Analysis*, Journal of the Royal Statistical Society. Series B (Methodological), vol. 53, no. 3, pp. 539–572 (1991).

[2] Yao, Fang and Müller Hans-Georg and Wang, Jane-Ling: *Functional Data Analysis for Sparse Longitudinal Data*, Journal of the American Statistical Association, vol. 100, no. 470, pp. 577–590 (2005).

[3] McLean, Mathew and Scheipl, Fabian and Hooker, Giles and Greven, Sonja and Ruppert, David: *Bayesian Functional Generalized Additive Models with Sparsely Observed Covariates*, submitted (2017).

[4] Xian, Chengqian and de Souza, Camila and Jewell, John and Dias, Ronaldo: *Clustering Functional Data via Variational Inference*, submitted (2023).

[5] Zambom, Adriano and Wang, Qing, and Dias, Ronaldo: *A Basis Approach to Surface Clustering*, Statistics Optimization & Information Computing, vol. 10, no. 2, pp. 339-351 (2022).

[6] Nychka, Douglas and Furrer, Reinhard and Paige, John and Sain, Stephan: *fields: Tools for spatial data*, R package version 14.1, https://github.com/dnychka/fieldsRPackage (2021).