The Journal of Open Source Software

baysc: An R package for Bayesian survey clustering

Stephanie M. Wu[1], **Matthew R. Williams**[2], **Terrance D. Savitsky**[3], **and Briana J.K. Stephenson**[1]

1 Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, U.S.A **2** RTI International, Research Triangle Park, North Carolina, U.S.A **3** Office of Survey Methods Research, U.S. Bureau of Labor Statistics, Washington, DC, U.S.A

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

Model-based clustering methods allow a large number of correlated variables to be summarized into underlying patterns, where each pattern describes a cluster and each individual is assigned to a cluster. Example applications include identifying dietary patterns from dietary intake data (Stephenson et al., 2020) and creating profiles of health and development among children (Lanza & Cooper, 2016). Bayesian formulations of such analyses allow the number of clusters to be determined by the data rather than through researcher post-hoc analyses. Clustering methods can also be extended to the hybrid supervised setting where interest lies in the association between the identified clusters and an outcome. When such clustering methods are applied to survey data, failure to account for the complex survey design and incorporate survey weights into the estimation leads to biased estimation and inference when the results are generalized to the population outside of the survey data.

The `baysc` R package provides functionality to allow for Bayesian clustering analyses, both unsupervised and supervised, to be performed while incorporating survey weights and accounting for complex survey design. The package uses methods derived from the latent class analysis (LCA) literature and focuses on clustering in the setting where the correlated variables are categorical and the outcome, where applicable, is binary. The package includes additional functions for plotting and summarizing output and creating simulated data, and an example dataset from the National Health and Nutrition Examination Survey (NHANES) containing dietary intake and hypertension data among low-income women in the United States (National Center for Health Statistics, 2023).

Statement of Need

A number of R packages provide functionality for model-based clustering in R. Frequentist approaches include `poLCA` (Linzer & Lewis, 2011) for classical LCA, `randomLCA` (Beath, 2017) for LCA with individual-specific random effects, and `mclust` (Scrucca, Fraley, Murphy, & Raftery, 2023) and `tidyLPA` (Rosenberg, Beymer, Anderson, Van Lissa, & Schmidt, 2019) for clustering of continuous variables. `BayesLCA` (White & Murphy, 2014) and `BayesBinMix` (Papastamoulis & Rattray, 2017) use Bayesian approaches for categorical and binary data, respectively. `PReMiuM` (Liverani, Hastie, Azizi, Papathomas, & Richardson, 2015) fits a wide variety of supervised models that handle various types of discrete and continuous exposure and outcome data. However, these packages do not allow for survey weights and complex survey design to be incorporated to ensure valid estimation and inference when using survey data.

The `baysc` package implements a weighted pseudo-likelihood approach proposed in Wu, Williams, Savitsky, & Stephenson (2023) that can integrate sampling weights when creating patterns using categorical data. The models adjust for stratification, clustering, and informative sampling to provide accurate point and variance estimation. When interest lies in how clusters are related to an outcome, supervised methods are available that jointly model the exposure and outcome, capturing the exposure-outcome association with more precision than two-step approaches that perform the clustering and the regression analyses sequentially. In addition, interaction effects between the clusters and the outcome are able to be captured through a mixture reference coding scheme.

Acknowledgement

This work was supported in part by the National Institute of Allergy and Infectious Diseases (NIAID: T32 AI007358), the National Heart, Lung, and Blood Institute (NHLBI: R25 HL105400), and the Harvard Data Science Initiative Bias^2 Program. The authors declare no conficts of interest.

References

Beath, K. J. (2017). randomLCA: An r package for latent class with random effects analysis. *Journal of Statistical Software*, *81*, 1–25.

Lanza, S. T., & Cooper, B. R. (2016). Latent class analysis for developmental research. *Child Development Perspectives*, *10*(1), 59–64.

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An r package for polytomous variable latent class analysis. *Journal of statistical software*, *42*, 1–29.

Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M., & Richardson, S. (2015). PReMiuM: An r package for profile regression mixture models using dirichlet processes. *Journal of statistical software*, *64*(7), 1.

National Center for Health Statistics. (2023). National health and nutrition examination survey home page. Centers for Disease Control; Prevention, National Center for Health Statistics, Atlanta, GA. Retrieved from https://www.cdc.gov/nchs/nhanes.htm

Papastamoulis, P., & Rattray, M. (2017). BayesBinMix: An r package for model based clustering of multivariate binary data. *R J.*, *9*(1), 403.

Rosenberg, J. M., Beymer, P. N., Anderson, D. J., Van Lissa, C., & Schmidt, J. A. (2019). tidyLPA: An r package to easily carry out latent profile analysis (LPA) using open-source or commercial software. *Journal of Open Source Software*, *3*(30), 978.

Scrucca, L., Fraley, C., Murphy, T. B., & Raftery, A. E. (2023). *Model-based clustering, classification, and density estimation using mclust in r*. Chapman; Hall/CRC.

Stephenson, B. J., Sotres-Alvarez, D., Siega-Riz, A.-M., Mossavar-Rahmani, Y., Daviglus, M. L., Van Horn, L., Herring, A. H., et al. (2020). Empirically derived dietary patterns using robust profile clustering in the hispanic community health study/study of latinos. *The Journal of Nutrition*, *150*(10), 2825–2834.

White, A., & Murphy, T. B. (2014). BayesLCA: An r package for bayesian latent class analysis.

Wu, S. M., Williams, M. R., Savitsky, T. D., & Stephenson, B. J. (2023). Derivation of outcome-dependent dietary patterns for low-income women obtained from survey data using a supervised weighted overfitted latent class analysis. *arXiv preprint arXiv:2310.01575*.