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** [4]

**1** Division of Psychiatry, UCL, London, U.K. **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 **4** Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, U.S.A ¶ Corresponding author

**DOI:** 10.21105/joss.08382

**Software**

- Review
- Repository
- Archive

**Editor:** Sehrish Kanwal

**Reviewers:**

- @CalebAbhulimhen
- @larryshamalama

**Submitted:** 09 April 2025
**Published:** 11 March 2026

**License**
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

## 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.

Interest may also lie in relating the identified clusters to an outcome, either through a secondary regression step or an all-in-one "supervised" approach that uses information from the outcome to inform the clustering process. 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 implements the methods proposed in Stephenson, Wu, & Dominici (2023) and Wu, Williams, Savitsky, & Stephenson (2024), providing functionality to allow for Bayesian clustering analyses – both unsupervised and supervised – to be performed while incorporating survey weights and design features that account for complex survey sampling designs. Asymptotically correct point estimates and credible intervals are produced with respect to the underlying population from which the observed sample was generated. This novel feature allows for application of latent class analysis (LCA) to datasets realized from surveys administered by government statistical agencies. The package uses methods derived from the 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, as well as 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).

Detailed information regarding package usage and functionality can be found in the vignette titled, "An introduction to the baysc package," available on the package GitHub repo located at https://github.com/smwu/baysc/.

## 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 Stephenson et al. (2023) and Wu et al. (2024) that can integrate survey sampling weights when performing cluster analysis using categorical or binary 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 additional variables, such as a binary outcome, users can either: 1) adopt a two-step approach where unsupervised clustering is performed and then a secondary regression is fit, or 2) use the all-in-one supervised approach that jointly models all variables, precisely propagating measurement error from the clustering analysis into the outcome regression model. The supervised approach also uses a mixture reference coding scheme to allow interaction effects between the clusters and the outcome to be captured. Statistical details can be found in the vignette, "An introduction to the baysc package," and Wu et al. (2024).

## 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 conflicts 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. doi:10.18637/jss.v081.i13

Lanza, S. T., & Cooper, B. R. (2016). Latent class analysis for developmental research. *Child Development Perspectives*, *10*(1), 59–64. doi:10.1111/cdep.12163

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*, 1–29. doi:10.18637/jss.v042.i10

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. doi:10.18637/jss.v064.i07

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. doi:10.32614/RJ-2017-022

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. doi:10.21105/joss.00978

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.

doi:10.1201/9781003277965

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. doi:10.1093/jn/nxaa208

Stephenson, B. J., Wu, S. M., & Dominici, F. (2023). Identifying dietary consumption patterns from survey data: a Bayesian nonparametric latent class model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnad135. doi:10.1093/jrsssa/qnad135

White, A., & Murphy, T. B. (2014). BayesLCA: An R package for Bayesian latent class analysis. doi:10.18637/jss.v061.i13

Wu, S. M., Williams, M. R., Savitsky, T. D., & Stephenson, B. J. (2024). Derivation of outcome-dependent dietary patterns for low-income women obtained from survey data using a supervised weighted overfitted latent class analysis. *Biometrics*, *80*(4), ujae122. doi:10.1093/biomtc/ujae122