The Journal of Open Source Software

# tidyLPA: An R Package to Easily Carry Out Latent Profile Analysis (LPA) Using Open-Source or Commercial Software

**Joshua M. Rosenberg**[1], **Patrick N. Beymer**[2], **Daniel J. Anderson**[3], **and Jennifer A. Schmidt**[2]

**1** University of Tennessee, Knoxville **2** Michigan State University **3** University of Oregon

**DOI:** 10.21105/joss.00958

**Software**

- Review
- Repository
- Archive

**Submitted:** 16 September 2018
**Published:** 20 September 2018

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

## Summary

Researchers are often interested in identifying homogeneous subgroups within heterogeneous samples on the basis of a set of measures, such as profiles of individuals' motivation (i.e., their values, competence beliefs, and achievement goals). Latent Profile Analysis (LPA) is a statistical method for identifying such groups, or *latent profiles*, and is a special case of the general mixture model where all measured variables are continuous (Harring & Hodis, 2016; Pastor, Barron, Miller, & Davis, 2007). The **tidyLPA** package allows users to specify different models that determine whether and how different parameters (i.e., means, variances, and covariances) are estimated, and to specify and compare different solutions based on the number of profiles extracted.

The aim of the **tidyLPA** package is to provide a simple interface for conducting and evaluating LPA models. Given that LPA is only one type of mixture model, we do not expect it to replace the more general functionality of other tools that allow for the estimation of wider range of models. Nevertheless, this package provides convenient methods for conducting LPA using both open-source and commercial software, while aligning with a widely used coding framework (i.e., *tidy* data, described more below). In doing so, **tidyLPA** allows researchers with and without access to proprietary tools, such as MPlus, to conduct LPA.

## A *tidy* user-interface

The input for *tidyLPA* assumes a tidy data structure (see Wickham & others, 2014), and all output are returned in a tidy from, which aligns with the broad array of tools within the **tidyverse** collection of R packages. The data can be efficiently used to create plots, explore model results, or used in subsequent analyses. The interface is also designed to work efficiently with the *pipe* operator, `%>%`, and **dplyr** helper functions that can be used to select variables, e.g.:

```
{code} data %>% tidyLPA::estimate_profiles(dplyr::starts_with())
```

The package is designed and documented to be easy to use, especially for beginners to LPA, but with fine-grained options available for estimating models and evaluating specific output as part of more complex analyses.

## Functionality through both open-source and commercial software

The *tidyLPA* package provides an interface to two different tools for estimating models, one from the open-source **mclust** R package (Scrucca, Fop, Murphy, & Raftery, 2017) and the other the commercial **MPlus** (L. Muthen & Muthen, 2017) software (via the **MplusAutomation** R package [hallquist_et_al_2018]). The packages are benchmarked to one another; the benchmarks are checked when **tidyLPA** is deployed through automated tests.

Both the open-source and commercial tools allow for the specification of four model parameterizations:

- Equal variances and covariances fixed to 0 (Model 1)
- Varying variances and covariances fixed to 0 (Model 2)
- Equal variances and equal covariances (Model 3)
- Varying variances and varying covariances (Model 6)

Two additional model parameterizations (Models 4 and 5) are only available through MPlus.

The two primary functions in the package are `estimate_profiles()` and `compare_solutions()`, with the former used to estimate a given model and the latter used evaluate differences in the fit of alternative models and number of profiles extracted. The `estimate_profiles()` function returns the predicted probability of membership in each profile for each case in the dataset, and allows for simple interpretation of the model output, particularly when combined with the `plot_profiles()` function, which displays the mean values (and their standard errors) on each measure for each profile.

The `compare_solutions()` function fits a wide range of models and returns various fit indices, including likelihood ratio tests and other statistics (e.g., entropy) for each parameterization. All three functions use **mclust**; corresponding functions with `_mplus()` appended use the **MPlus** software.

## References

Harring, J. R., & Hodis, F. A. (2016). Mixture modeling: Applications in educational psychology. *Educational Psychologist*, *51*(3-4), 354–367.

Muthen, L., & Muthen, B. (2017). Mplus user's guide (8th ed.). *Muthen & Muthen*.

Pastor, D. A., Barron, K. E., Miller, B., & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology*, *32*(1), 8–47.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2017). Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 205–233.

Wickham, H., & others. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23.