Reprinted from the Proceedings of the COMPSTAT 2000 conference.

Graphical Displays for Latent Class Cluster and Latent Class Factor Models

Jeroen K. Vermunt[1] and Jay Magidson[2]

[1] Dep. of Methodology, Tilburg University, P.O. Box 90153, 5000LE Tilburg, NL

[2] Statistical Innovations Inc., P.O. Box 1, Belmont, MA 02478, USA

Keywords. latent class analysis, factor analysis, cluster analysis, mixture models, categorical data, graphical displays, bi-plot, tri-plot, statistical software

Latent class (LC) analysis is becoming one of the standard data analysis tools in social, biomedical, and marketing research. This paper discusses two different forms of (exploratory) LC analysis which are implemented in a new computer program called Latent GOLD® (Vermunt and Magidson, 2000). We will not only focus on the model formulations, but also on (partially new) graphical displays (uni-, bi-, and tri-plots) that facilitate the interpretation of results from a LC analysis. As an illustration, we use a small data set containing five dichotomous indicators measuring certain political attitudes and three covariates (Hagenaars, 1993). First we present models without covariates; later we show how to include covariates in a LC model.

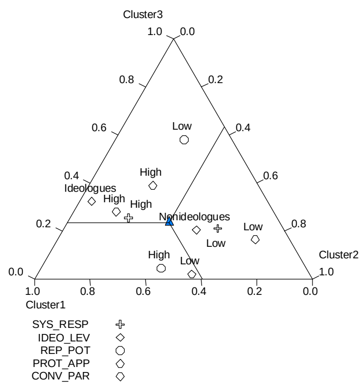
A standard exploratory approach to fitting LC models to data is to begin with a one-class (independence) model, followed by a two-class model, a three-class model, etc., and continuing until a model is found that provides an adequate fit to the data. We refer to such models as LC cluster models since the k classes serve the same function as the k clusters obtained in cluster analysis.

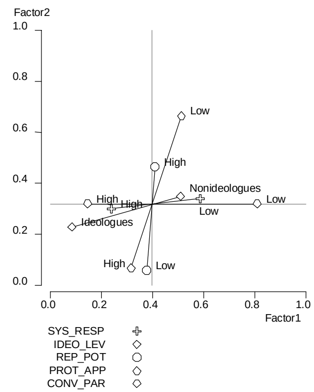
We propose an alternative sequence for fitting models which involves increasing the number of latent variables (factors) rather than the number of classes (clusters). We call this approach LC factor analysis because of the natural analogy to standard factor analysis. A LC factor model contains t dichotomous and mutually independent latent variables. To exclude higher-order interactions, logit models are specified for the response probabilities. An interesting feature of the resulting t-factor model is that it has exactly the same number of parameters as a model with $(t+1)$ clusters.

Model	L^2	BIC	df	p-value
1 Cluster	296.56	113.19	26	0.00
2 Cluster (or 1 Factor)	95.82	-45.24	200	0.00
3 Cluster	24.80	-73.94	14	0.04

4 Cluster	7.95	-48.47	8	0.44
2 Factor	12.30	-86.44	14	0.58

The table shows the test results for the one- to four-cluster models and the one- and two-factor models for the political attitudes data. As can be seen, the two-factor model fits much better than the three-cluster model, even though they have the same number of parameters. This is something we have encountered for all

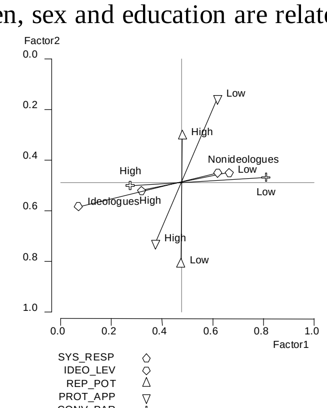


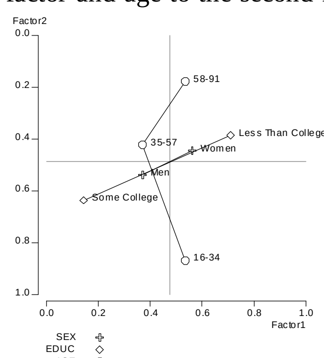


data set we analyzed thus far using the two types of LC models.

The figure presents the tri-plot for the three-cluster and the bi-plot for the two-factor model. The tri-plot depicts the probability of belonging to cluster *k* for each indicator level (see also van der Ark and van der Heijden 1998) and the bi-plot shows the probability of being in level 2 of factor *t*. These probabilities are obtained by collapsing the individual class-membership probabilities over the other variables. Both plots show clearly that the five indicators measure two different dimensions.

It is also possible to include covariates as predictors of the latent variable(s) in a LC model. Below is the bi-plot for the two-factor model with sex, education, and age as covariates (indicators and covariates are plotted separately). As can be seen, sex and education are related to the first factor and age to the second factor.





The LC models implemented in Latent GOLD® are more general than the models presented in this paper: indicators and covariates not only can be nominal but also ordinal, continuous, or counts; it is possible to include local dependencies between indicators; and factors can have more than two ordered levels and can be correlated.

References

Hagenaars, J.A. (1993). *Loglinear models with latent variables*. Newbury Park: Sage.

van der Ark, L.A. & van der Heijden, P.G.M. (1998). Graphical display of latent budget and latent class analysis. In: J. Blasius & M.Greenacre (eds.) *Visualization of categorical data*, 489-509. Boston: Academic Press.

Vermunt, J.K. & Magidson, J. (2000). *Latent GOLD 2.0 User's Guide*. Belmont, MA: Statistical Innovations Inc.