Qual Life Res (2012) 21:637–650
DOI 10.1007/s11136-011-9976-6

Latent variable mixture models: a promising approach for the validation of patient reported outcomes

Richard Sawatzky · Pamela A. Ratner · Jacek A. Kopec · Bruno D. Zumbo

Supplementary Technical Appendix

This Appendix provides additional information for the paper, *Latent Variable Mixture Models: A Promising Approach for the Validation of Patient Reported Outcomes.* It elaborates the IRT and latent factor analysis representations of the GRM, the computation of predicted probabilities based on a latent variable mixture model (LVMM), notes about model estimation of a LVMM, notes about explaining latent class membership, and annotated MPlus 5.2 syntax for fitting a LVMM (i.e., the syntax for the example described in the manuscript).

IRT and latent factor analysis representations of the GRM

Using the conventional IRT notation for the GRM [1], the cumulative probability (P_{ij}) of an item *i* response at or above category *j* is expressed as follows:

$$P_{ij}\left(Y \geq j \mid \theta\right) = \frac{\exp(\alpha_i\left(\theta - \beta_{ij}\right))}{1+\exp(\alpha_i\left(\theta - \beta_{ij}\right))}, \quad (1)$$

where α denotes the discrimination parameter for item *i*, β denotes the difficulty parameter for the response categories less one, for each item *i*, and θ denotes the (predicted) latent factor score. This is equivalent to a factor analysis representation of the GRM based on Muthén's general latent variable modeling framework [2], where the relationships between a latent factor (a.k.a. "theta") and ordinal indicators are represented in a form equivalent to a logistic proportional

odds model. The corresponding formula for the cumulative probability (P_{ij}) of an item *i* response at or above category *j* is expressed as follows [3]:

$$P_{ij}\left(Y \geq j \mid \theta\right) = \frac{\exp(-\tau_{ij} + \lambda_i\theta)}{1+\exp(-\tau_{ij} + \lambda_i\theta)}, \quad (2)$$

where τ_{ij} denotes the thresholds between the categories of item *i*, and λ_i denotes the factor loading for item *i*. If θ is normally distributed with a mean of zero and variance of one, none of the thresholds or factor loadings are constrained, and a logistic link function with maximum likelihood estimation is used, the following transformation can be applied to convert the Mplus thresholds (τ) and factor loadings (λ) of Equation 2 into the difficulty (β) and discrimination (α) parameters of the GRM [4]:

$$\beta_{ij} = \frac{\tau_{ij}}{\lambda_i} \text{ and } \alpha_i = \lambda_i . \quad (3)$$

The computation of predicted probabilities based on a LVMM

A LVMM based on the GRM can be obtained by allowing the factor loadings and the thresholds to vary across two or more latent classes. The cumulative probability of an item response at or above category *j* within a latent class can be computed as follows [3]:

$$P_{ijk}\left(Y \geq j \mid \theta, C = k\right) = \frac{\exp(-\tau_{ijk} + \lambda_{ik}\theta)}{1+\exp(-\tau_{ijk} + \lambda_{ik}\theta)}, \quad (4)$$

where *C* is the latent class variable with *k* classes. The cumulative probability of an item response at or above category *j*, for an individual, within the combined (heterogeneous) population is obtained by summing the product of the individual's item response probabilities within the latent classes and the posterior probability of latent class membership, as is shown in the following equation:

$$P_{ij}(Y \geq j \mid \theta) = \sum_{k=1}^{K}(X_k * P_{ijk}(Y \geq j \mid \theta)), \quad (5)$$

where X_k is the posterior probability of an individual being in class *k*, which when summed across all classes equals 1.0. The posterior probability of latent class membership can be obtained using Bayes' theorem by multiplying the likelihood of the model-predicted factor score (theta), for an individual, by the probability of theta given a normal distribution (i.e., a normal prior) [5, 6].

The probability of an item response corresponding with the predicted theta score for an individual is obtained in a similar way as would be done for a proportional odds logistic regression model, which is achieved as follows [see 6, 7, 8]:

$$\text{if } j \text{ is the first category: } P_{ij}(Y = j|\theta) = 1 - P_{ij+1}(Y \geq j+1|\theta), \quad (6a)$$

$$\text{if } j \text{ is a middle category: } P_{ij}(Y = j|\theta) = P_{ij}(Y \geq j|\theta) - P_{ij+1}(Y \geq j+1|\theta), \quad (6b)$$

$$\text{if } j \text{ is the last category: } P_{ij}(Y = j|\theta) = P_{ij}(Y \geq j|\theta), \quad (6c)$$

where $P_{ij}(Y \geq j|\theta)$ is the cumulative probability obtained from Equations 2 (one-class model) or 5 (mixture model).

Notes about model estimation of a LVMM

The expectation maximization (EM) algorithm and various extensions thereof are widely used to obtain the maximum likelihood parameter estimates in mixture models and in many other applications where there are unknown elements (e.g., missing data) [9, 10]. This is an iterative process that begins by computing the posterior probabilities based on the expectation of the log likelihood initially using arbitrary starting values (the E-step). The resulting information is used in the M-step to produce new maximum likelihood parameter estimates consistent with the observed data that are used in the next iteration. These steps are repeated until the likelihood no

longer improves beyond a predetermined small increment. It is important to recognize that the estimation procedures for mixture models, in general, are susceptible to local maxima resulting in the model's convergence on a sub-optimal solution [11]. That is, a local maximum in the likelihood value could be obtained as a result of the parameter starting values (a.k.a. initial values). It is therefore recommended to attempt to replicate the maximum likelihood using many sets of different starting values. The plausibility of having found the optimal maximum likelihood is increased when different starting values result in the same (replicated) maximum likelihood. To further assess whether the optimal maximum likelihood has been obtained, it is recommended to compare the stability of the parameter estimates, the predicted factor scores, and the posterior probabilities of the latent class memberships across neighboring solutions. If these are similar, then the model is deemed to be adequately defined for the data. In the absence of replicating the maximum likelihood, and to increase the chances of obtaining an optimal solution, researchers can increase the number of starting value sets, or the number of iterations, or address potential sources of under-identification (such as very sparse frequencies in the cross-tabulations of the item responses).

Notes about explaining latent class membership

We demonstrate a two-step approach that involves first identifying the individuals' most likely latent class membership (based on the estimated model) and subsequently regressing class membership on exogenous variables (using logistic regression for categorical data). This approach, however, may lead to inaccurate results if, for example, the latent classes are poorly discriminating leading to greater uncertainty in the prediction of latent class membership (i.e., a low entropy value is obtained). In this situation it is desirable to use pseudo-class draws to more accurately estimate the parameters and variances of the variables explaining latent class

membership [12-14]. As nicely exemplified in the description provided by Petras and Masyn [15], pseudo-class draws involve taking random draws from the discrete posterior latent class probability distribution of class membership for each individual in the sample. Typically it is recommended that one take 20 draws [12]. The logistic regression model with the variables explaining latent class membership is estimated repeatedly for the 20 draws and the obtained parameters are averaged and variances estimated.

Mplus 5.2 syntax for a 3-class LVMM

Syntax	Explanation
USEVARIABLES ARE SFRC_03-SFRC_10 SFRC_11R SFRC_12R; CATEGORICAL ARE SFRC_03-SFRC_10 SFRC_11R SFRC_12R;	Specifies 10 ordinal categorical variables.
AUXILIARY ARE (r) CCCC_031 CCCC_051 CCCC_061 CCCC_081 CCCC_101 CCCC_121 CCCC_191R CCCC_280 CCCC_901 CCCC F1 DHHC_SEX DHHC_AGE;	Specifies the variables that are saved and subsequently used in a multinomial model using pseudo-class draws of the posterior probabilities of latent class membership [12-14]. These variables are not included in the LVMM. Rather, "auxiliary (r)" instructs the software to run a separate model.
CLASSES = C(3);	Specifies three latent classes.
ANALYSIS: TYPE = MIXTURE;	Specifies a mixture model.
ESTIMATOR = MLR;	Specifies a robust maximum likelihood estimator.
ALGORITHM = INTEGRATION; INTEGRATION = STANDARD(50); ADAPTIVE = ON; CHOLESKY = ON;	Specifies adaptive integration with 50 integration points for the latent factor.
LINK = LOGIT;	Specifies a logistic link function (proportional odds in the case of ordinal variables).
MITERATIONS = 1000;	Specifies the maximum number of iterations.
STARTS = 5000 1000; STITERATIONS = 20;	Specifies 5,000 sets of stage 1 random starting values with 20 iterations, followed by 1,000 sets of stage 2 random starting values with the greatest log-likelihood in stage 1.
K-1STARTS = 5000 1000;	Specifies 5,000 and 1,000 sets of stage 1 and 2 starting values respectively for the k-1 class model used for the bootstrapped likelihood ratio test and the Vuong-Lo-Mendell-Rubin likelihood ratio test [16-18].

Syntax	Explanation
LRTSTARTS = 1000 200 5000 1000;	Specifies the stage 1 and 2 starting values for the k-1 class model (first two numbers respectively) and k class model (last two numbers) for the simulated data used for the bootstrapped likelihood ratio test [9, 19].
MODEL: %OVERALL% physfun by SFRC_03* SFRC_04-SFRC_10 SFRC_11R SFRC_12R; [physfun@0]; physfun@1;	Specifies the overall population model, with all slopes and thresholds to be estimated, and a standardized latent factor (mean of 0 and variance of 1) (the thresholds are estimated as the default and are therefore not specified in the syntax).
%C#1% physfun by SFRC_03* SFRC_04-SFRC_10 SFRC_11R SFRC_12R; [SFRC_03$1-SFRC_10$2]; [SFRC_11R$1-SFRC_12R$1]; %C#2% physfun by SFRC_03* SFRC_04-SFRC_10 SFRC_11R SFRC_12R; [SFRC_03$1-SFRC_10$2]; [SFRC_11R$1-SFRC_12R$1]; %C#3% physfun by SFRC_03* SFRC_04-SFRC_10 SFRC_11R SFRC_12R; [SFRC_03$1-SFRC_10$2]; [SFRC_11R$1-SFRC_12R$1];	Specifies the within-class models, with all slopes and thresholds to be estimated within each of the latent classes (the thresholds are indicated using square brackets). The mean and variance of the within class latent factors are constrained to be equivalent to the population mean and variance of 0 and 1 by default.
OUTPUT:	
RESIDUAL	Produces univariate and bivariate residuals within each of the latent classes.
STANDARDIZED	Produces the standardized model results.
TECH1	Produces the parameter specification details and starting values.
TECH2	Produces the parameter derivatives.
TECH3	Produces the correlation and covariance matrices of the estimated parameters.

Syntax	Explanation
TECH7	Produces the within-class frequency distributions using pseudo-class draws to determine latent class membership.
TECH10	Produces the univariate and bivariate residuals of the overall model, including the standardized difference scores for each categorical comparison, and the likelihood ratio test for each variable [20].
TECH11	Produces the Vuong-Lo-Mendell-Rubin likelihood ratio test [16-18].
TECH14;	Produces the bootstrapped likelihood ratio test [9, 19].

* Default settings. The following additional default settings were adopted (not explicitly specified): EM convergence criteria; Maximum and minimum values of logit thresholds; Random seed specification.

References

1. Samejima, F. (1997). Graded response model. In W. J. v. d. Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.
2. Muthén, B. (2008). Latent variables hybrids. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 1-24). Charlotte, NC: Information Age Publishing.
3. Muthén, B., & Muthén, L. (2010). IRT in Mplus. http://www.statmodel.com/download/MplusIRT2.pdf. Accessed 15 January 2011.
4. Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory. *Structural Equation Modeling - A Multidisciplinary Journal, 15*, 136-153.
5. Dayton, C. M. (1998). *Latent class scaling analysis*. Thousand Oaks, CA: Sage.
6. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum.
7. O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks, CA: Sage.
8. Wu, A. D., & Zumbo, B. D. (2007). Thinking about item response theory from a logistic regression perspective: A focus on polytomous models. In S. S. Sawilowsky (Ed.), *Real data analysis* (pp. 241-269). Charlotte, NC: Information Age Publishing.
9. McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
10. Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge, NY: Cambridge University Press.
11. Asparouhov, T., & Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27-51). Charlotte, NC: Information Age Publishing.
12. Wang, C. P., Brown, C. H., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association, 100*, 1054-1076.
13. Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association, 92*(440), 1375-1386.
14. Muthén, B., & Muthén, L. (2007). Wald test of mean equality for potential latent class predictors in mixture modeling. http://www.statmodel.com/download/MeanTest1.pdf. Accessed 20 October 2010.
15. Petras, H., & Masyn, K. (2010). General growth mixture analysis with antecedents and consequences of change. In A. R. Piquero, D. Weisburd & ebrary Inc. (Eds.), *Handbook of quantitative criminology* (pp. 69-100). New York: Springer.
16. Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57*, 307.
17. Lo, Y. T., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767-778.
18. Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling - A Multidisciplinary Journal, 14*, 202-226.
19. Nylund, K. L., Asparoutiov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling - A Multidisciplinary Journal, 14*, 535-569.
20. Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley-Interscience.