Behaviour Research and Therapy 98 (2017) 91–102

Contents lists available at ScienceDirect

Behaviour Research and Therapy

journal homepage: www.elsevier.com/locate/brat

# Fitting latent variable mixture models

Gitta H. Lubke[*], Justin Luningham

*Department of Psychology, University of Notre Dame, Notre Dame, IN, United States*

ARTICLE INFO

*Article history:*
Received 26 November 2016
Received in revised form
3 April 2017
Accepted 11 April 2017
Available online 17 April 2017

*Keywords:*
Mixture modeling
Latent class analysis
Growth mixture models

ABSTRACT

Latent variable mixture models (LVMMs) are models for multivariate observed data from a potentially heterogeneous population. The responses on the observed variables are thought to be driven by one or more latent continuous factors (e.g. severity of a disorder) and/or latent categorical variables (e.g., subtypes of a disorder). Decomposing the observed covariances in the data into the effects of categorical group membership and the effects of continuous trait differences is not trivial, and requires the consideration of a number of different aspects of LVMMs. The first part of this paper provides the theoretical background of LVMMs and emphasizes their exploratory character, outlines the general framework together with assumptions and necessary constraints, highlights the difference between models with and without covariates, and discusses the interrelation between the number of classes and the complexity of the within-class model as well as the relevance of measurement invariance. The second part provides a growth mixture modeling example with simulated data and covers several practical issues when fitting LVMMs.

© 2017 Elsevier Ltd. All rights reserved.

## 1. Introduction

Latent variable mixture models (LVMMs) combine latent class analysis models and factor models or more complex structural equation models (Muthén, 2001). LVMMs are most commonly used to investigate population heterogeneity, which refers to the presence of subgroups in the population. LVMMs can serve to analyse data from heterogeneous populations without knowing beforehand which individual belongs to which of the subgroups.

The simplest types of mixture models are latent class analysis (LCA) models. These models are designed for multiple observed variables (e.g., symptom endorsements, of questionnaire items), and have a single latent class variable that groups the individuals in a sample into a user-specified number of latent groups (Lazarsfeld & Henry, 1968; McCutcheon, 1987). LCA models do not have factors within class, and the covariances between the observed variables *within class* are constrained to zero.[1] This is a very stringent assumption. Suppose we have 5 observed items measuring some disorder. Not allowing these items to covary within class means that there are no systematic severity differences between participants within a class in LCA models. The covariances between observed variables *in the total sample* only deviate from zero due to mean differences between the classes.

Factor models on the other hand are models for a single homogeneous population (i.e., no differences between subtypes), and observed variables in the sample are assumed to covary due to systematic differences along the underlying continuous latent factors (Bollen, 1989).

LVMMs can have one or more latent class variables, and permit the specification of factor models, growth models, or even more complex models within each class. If the within class model is a factor model, the resulting LVMM is often called factor mixture model. Covariances between observed variables in the total sample are attributed partially to mean differences between classes, and partially to continuous latent factors within each class. For example, consider data collected on several questionnaire items that measure anger. Suppose the population consists of two groups, a majority group of participants with very low levels of anger and a smaller group characterized by high scores on most of the items. The observed anger items in the total sample covary because of the mean differences between the two groups. In addition, the items can also covary if there are differences in the severity of anger *within* each group. These two sources of covariance are modeled in LVMMs by using latent categorical and latent continuous variables.

Latent class models are a special case of the LVMM where factor variances (or, alternatively, factor loadings) are zero. In the anger example this would mean that all participants within the low-

[*] Corresponding author. 110 Haggar Hall, Notre Dame, IN 46556, United States.
*E-mail address:* glubke@nd.edu (G.H. Lubke).

[1] This is called the “local independence assumption”.

http://dx.doi.org/10.1016/j.brat.2017.04.003
0005-7967/© 2017 Elsevier Ltd. All rights reserved.

scoring class do not differ in the severity of anger (i.e., zero anger factor variance within group). The same holds for the high scoring group: the assumption of the latent class model is no variability of anger within group because if there were systematic anger differences within class then the items would in fact covary. The observed covariances between the anger items in this model are modeled to be entirely due to mean differences between the groups. Factor models for a homogeneous population are also a special case: they are LVMMs with a single latent class. In the anger example this would boil down to neglecting the presence of two subgroups, and attributing all covariances to one underlying anger factor within a single homogenous population.

The LVMM framework is extremely flexible, and permits the specification of different types of mixture models. Models such as path models, factor models, survival models, growth curve models, and more general structural equation models can all be specified for multiple subgroups instead of for a single homogeneous population (see for instance Arminger, Stein, & Wittenberg, 1999; Dolan & van der Maas, 1998; Jedidi, Jagpal, & DeSarbo, 1997; Muthén & Shedden, 1999; Muthén & Muthén, 2000; Ram & Grimm, 2009; Varriale & Vermunt, 2012; Vermunt, 2008; Yung, 1997). The flexibility comes at a price. The framework is built on a set of assumptions that should be realistic for the data. Further, in order to estimate a model, all relations between observed variables, between observed variables and latent variables, and between latent variables have to be specified. It is therefore necessary to decide whether within-class model parameters are class specific, or are the same for all classes (i.e., class invariant). As will be discussed in this paper, the interpretation of the model depends on these decisions. It is important to note that different within-class parameterization can influence how many classes best fit the data (Lubke & Neale, 2008). However, comparing a set of carefully parameterized mixture models can provide great insight into the processes and interrelations between variables when the assumption of population homogeneity is unrealistic.

The paper is organized into two main parts. The first part provides the theoretical background. After discussing the generally exploratory character of mixture analyses, the modeling framework is presented together with some of the necessary assumptions and constraints. The first part concludes with the discussion of issues that deserve consideration prior to fitting models to data, such as the interrelation between number of classes and within-class model complexity, measurement invariance, and models with and without covariates. The second part consists of a growth mixture analysis with covariates, and illustrates some of the practical issues discussed in the first part of the paper.

## 2. Part I

### 2.1. Exploration of heterogeneity using mixture models

Latent variable mixture models (LVMMs) afford the possibility to detect groups of subjects in a sample, and to investigate the differences between the groups. LVMMs differ from other techniques to detect groups in data, such as taxometrics and cluster analysis, in that they require the user to specify all relations between observed and latent variables in the model (Lubke & Miller, 2014; Meehl, 1995). LVMMs are therefore prone to misspecifications. However, if there is sufficient *a priori* knowledge to specify these relations, then LVMMs usually have more power to detect groups in the data (Lubke & Tueller, 2010).

LVMMs differ from multi-group models in that it is not necessary to know which subject belongs to which group. Group membership is unobserved, or latent. Mixture models are therefore especially useful if the causes of the grouping are not known *a priori*. The grouping variable is formalized as a latent categorical variable, and the groups are called latent classes. In a cross-sectional setting, classes can consist of subjects with class-specific response profiles (e.g., high scores on some questionnaire items but low on others, or high on all), and in a longitudinal setting classes are characterized by class-specific trajectories over time (e.g., an increasing risk trajectory and a low constant trajectory).

If the process that causes the grouping is not well understood, then it is unlikely that the exact number of latent classes or the within-class structure are known. Mixture analyses are therefore often rather exploratory in character. Typically, a set of models with an increasing number of latent classes and different within-class structures is fitted to the data (e.g., more vs. less constrained models, see part 2, applied example). Model selection is based on measures such as the Bayesian Information Criterion (BIC) or the bootstrapped likelihood ratio test (McLachlan & Peel, 2000; Schwarz, 1978). Of course there is nothing wrong with exploratory analyses, quite the contrary. One can learn a lot from investigating heterogeneity, and such an analysis can be much more insightful about the structure in the data than incorrectly assuming that the data were sampled from a single homogeneous population. However, the exploratory character of a mixture analysis needs to be taken into account when best-fitting models are interpreted, and results need to be validated before specific conceptual conclusions concerning the class structure and within-class parameters can be drawn.

### 2.2. The modeling framework

This section provides a brief overview of the key aspects of the LVMM framework so that the practical challenges in an empirical mixture analysis, as illustrated in part 2 of the paper, can be fully appreciated.

Within the LVMM framework the population can consist of $k = 1, \ldots, K$ latent classes. If $K = 1$, then there is only a single class (i.e. a single homogeneous population). The $K = 1$ case therefore includes factor models, structural equation models, and growth models for a single homogeneous population. In case $K > 1$, then a model needs to be specified for each of the classes. These within-class models are estimated jointly using a mixture distribution. A mixture distribution is a weighted sum of $K$ component distributions, and is denoted as

$$f(Y) = \sum_{k=1}^{K} \pi_k f_k(Y; \theta_k) \tag{1}$$

where $Y$ is a vector of observed random variables, $\pi_k$ is a weight that quantifies the relative size of the $k^{th}$ component, and $\theta_k$ is a vector of model parameters for the $k^{th}$ component (see McLachlan & Peel, 2000, for more detail on mixture distributions). The most common choice for the component distributions $f_k$ is the multivariate normal distribution, although other distributions such as the Poisson distribution can be chosen to accommodate non-normal observed data (e.g., counts of cigarettes, etc.). In case each set of observed variables within class, $Y_k$, is multivariate normally distributed, we have $Y_k \sim MVN(\mu_{k,} \Sigma_k)$, where the parameter vector $\theta_k$ contains the parameters that structure the component specific means, $\mu_{k,}$ and covariance matrices, $\Sigma_k$ :

$$\mu_k = \nu_k + \Lambda_k (\mathrm{I} - \mathrm{B}_k)^{-1} \alpha_k, \tag{2}$$

$$\Sigma_k = \Lambda_k (\mathrm{I} - \mathrm{B}_k)^{-1} \Psi_k \left[ (\mathrm{I} - \mathrm{B}_k)^{-1} \right]^t \Lambda_k^t + \Theta_k, \tag{3}$$

where $\nu_k$ are the intercepts, $\Lambda_k$ is the factor loading matrix, $\mathrm{I}$ is an

Download English Version:

https://daneshyari.com/en/article/5038144

Download Persian Version:

https://daneshyari.com/article/5038144

Daneshyari.com