

## 1 Latent class model

The latent class model using notation from (2.14) in Heinen is

$$p_v = \sum_{t=1}^T p_{v|\theta_t} p_{\theta_t}$$

where

- $p_{v|\theta_t}$  is the probability of a particular response pattern given that the individual is in latent class  $t$ , i.e. given  $\theta_t$
- $p_{\theta_t}$  is the probability of being in latent class  $t$ .

NOTE using notation from (2.20) in Heinen (Note I use  $p$  to represent the total number of observed variables whereas Heinen uses  $n$ )

$$p_{v|\theta_t} = \prod_{j=1}^p \prod_{g=0}^{m_j} p_{jg|\theta_t}^{x_{vjg}}$$

where

- $x_{vjg} = 1$  if in response pattern  $v$ , item  $j$  is responded to with category  $g$
- $x_{vjg} = 0$  otherwise.
- $p_{jg|\theta_t}$  is the probability of answering question  $j$  with category  $g$  given that the individual is in latent class  $t$ .

## 2 Likelihood and Estimation

So, the likelihood should be

$$L = \prod_v \sum_{t=1}^T p_{\theta_t} \prod_{j=1}^p \prod_{g=0}^{m_j} p_{jg|\theta_t}^{x_{v j g}}$$

where  $\prod_v$

means take the product over all the observed responses found in a particular data set.

Note  $x_{v j g}$  is the observed data in this likelihood, AND

$p_{jg|\theta_t}$  and  $p_{\theta_t}$  that are the parameters of interest in the model.

Maximization of this Likelihood can be performed using: Newton Raphson, scoring algorithm, or the EM algorithm (Expectation, Maximization Algorithm). EM is natural here because the latent classes can be considered as missing data. That is, if we knew what class each individual was in the problem would be easy, but here the classes are “missing” or unobserved. (See pages 54-60 in Heinen for sketch of the idea for EM for this model)

MPLUS uses the E-M algorithm.

### 3 Degrees of freedom - identifiability

Degrees of freedom for the unrestricted latent class model with  $T$  latent classes with  $j = 1 \dots n$  observed variables where each has  $m_j$  categories is

$$df = \prod_{j=1}^p m_j - \left( T + \sum_{j=1}^p T(m_j - 1) \right)$$

It is necessary but not sufficient that  $df \geq 0$  for the model to be identified.

page 60 of Heinen discusses a sufficient condition that tests whether a matrix containing the first derivatives of each of the  $p_v$  with respect to model parameters is of full column rank or not. Not easy to implement with MPLUS (as far as I can tell).

Non identified models can often be made identified by placing restrictions on model parameters. We'll see later how to place restrictions. Chapter 3 in Heinen.

## Goodness of Fit for Latent Class Models

- $\chi^2$  test compares observed frequencies with expected frequencies under the model of  $T$  latent classes
- if  $n \gg 2^p$  Chi Square test should be good, but how much bigger???
  - Literature has mixed responses
  - Some say  $n = 2 * 2^p$  is ok,
  - Others say  $n = 16 * 2^p$  is necessary.
- 1 technique of dealing with small counts for particular responses is to “group the responses” and expected responses so that each “group” has at least 5 responses
  - reduces the degrees of freedom
  - BK very briefly mention this on page 91 and their software outputs  $\chi^2$  associated with “groups” but unfortunately the grouping algorithm is not explained
- To compare models with different numbers of latent classes, it is not correct to do a chi-square difference test because the models are not actually nested. Common to use AIC to compare models with different numbers of classes. Choose the one with the smallest AIC.