

Tutorial For Inequality Constrained Latent Class Modeling (LCM version 1.0)

Rens van de Schoot and Herbert Hoijtink
Department of Methods and Statistics
Utrecht University, The Netherlands *

August 5, 2008

1 Installation and Preparation of The Dataset

This tutorial refers to the software that is described and used in the chapter about latent class models specified using inequality constraints that is enclosed in the folder *literature* in the zip-file. Both tutorial and software are based on previously published articles Hoijtink (2001), Hoijtink and Molenaar (1997), Hoijtink (1998), and Hoijtink (2000).

The data that can be analyzed with this software are the responses $x_{ij} \in \{0, 1\}$ of $I = 1, \dots, N$ persons to $j = 1, \dots, J$ items. The item responses are dichotomous and can represent responses like *no*, *yes* or *incorrect*, *correct*, etc.. The main goal of latent class analysis is to determine groups of persons with similar items responses. The maximum number of cases is 4000, the maximum number of items is 25, the maximum number of latent classes is 8 and the maximum of models is 20.

1.1 Installation Of The Software

The software is available at "<http://www.fss.uu.nl/ms/hoijtink>". Download the file "LCM_v01.zip" and copy this zip-file to your desktop. Extract the files to a predetermined folder. The following subfolders should appear: *software*, *examples* and *literature*.

*Correspondence should be addressed to the second author: Department of Methods and Statistics, Utrecht University, P.O. Box 80.140, 3508TC, Utrecht, The Netherlands; E-mail address: h.hoijtink@uu.nl. Supported by a grant from the Dutch organization for scientific research: NWO-VICI-453-05-002

1.2 Preparing Data

The first step is to prepare your data and save it with a *.txt* extension. Every row should consist of separate cases and every column should represent a dichotomous answer to an item. No missing values are allowed and responses should be separated using a space or tabs, use for example fixed ASCII or Tab delimited. Next, copy the data file to the software folder and rename the data file into *data.txt* (warning: information in the original *data.txt* will be lost).

For teaching purposes the file *data.txt* in the folder software consists of responses of 10 cases to 3 imaginary items. Examples below are computed using this data set.

2 Analysis

2.1 Specifications

Data analysis starts with the the following specifications. Open the file *input.txt* with Notepad. The first step is to specify:

- the number of cases by changing the number below N
- the number of items by changing the number below $Nitem$
- the number of classes you want the program to estimate by changing the number below $Nclass$

In our example there are 10 cases, 3 items and 2 classes, the first lines of the input file should look like this (see *input.txt*):

N	$Nitem$	$Nclass$
10	3	2

The next step is to specify starting values for each class-specific probability, representing the probability to score 1 on an item in a class. In our example with 3 items and 2 classes the inequality constraints can look like this, where π_{qj} represents the class specific probability for item j in class q :

$$\{\pi_{11}, \pi_{12}, \pi_{13}\} > \{\pi_{21}, \pi_{22}, \pi_{23}\},$$

in words this means that persons who belong to the first class are expected to have higher probabilities to respond 1 to each of the three items than persons who belong to the second class.

For specification of the starting values, the only rule is that they should be coherent with the pre-specified inequality constraints for the model under investigation. Note that these values may be arbitrarily chosen, as long as they are in agreement with the inequality constraints. In our example the initial probability to score 1 on item 1 for a person in the first class is set to .7 and for a person in the second class to .2. All values between 0 and 1 are permitted

as long as the initial values for the first class are higher than the values for the second class. The starting values for our example could look like this (rows indicate classes and columns indicate items):

Starting values

.7	.7	.7
.2	.2	.2

Do not forget to save any changes you may make to the file *input.txt*, do not change the name of this file.

2.2 Inequality Constraints

After the specification discussed above, the inequality constraints need to be specified. In section 3 more elaborate examples will be discussed, but for now we will use our simple example.

To specify the inequality constraints, begin with the first class specific probability in the first class π_{11} . Specify how often π_{11} is larger and smaller than a probability in this or another class. In our example π_{11} is larger than 3 other probabilities and smaller than 0 other probabilities. The input file for the first probability in the first class looks like this (see *constraints.txt*):

<i>Class</i>	<i>Item</i>	<i>Larger than</i>	<i>Smaller than</i>
1	1	3	0

Next, specify the probabilities that are smaller than π_{11} (that is, π_{11} is larger) followed by that probabilities that are larger than π_{11} (that is, π_{11} is smaller). In our example all probabilities in the second class are smaller and none are larger than π_{11} and the input file looks like this:

Larger than :

<i>Class</i>	<i>Item</i>
2	1
2	2
2	3

Smaller than

<i>Class</i>	<i>Item</i>
--------------	-------------

Repeat the same steps for every item in this class and then proceed with the next class. Leave one blank line after the last constraint and put the following text on the last row of *constraints.txt*: END OF FILE.

The last constraint in our example is for the third item within the second class (π_{23}) and looks like this:

<i>Class</i>	<i>Item</i>	<i>Larger than</i>	<i>Smaller than</i>
2	3	0	3
<i>Larger than :</i>			
<i>Class</i>	<i>Item</i>		
<i>Smaller than</i>			
<i>Class</i>	<i>Item</i>		
2	1		
2	2		
2	3		

END OF FILE

Do not forget to save any changes you may make to *constraints.txt*, and do not change the name of this file.

2.3 Computing Estimates

To compute the marginal likelihood (see, the chapter about latent class models specified using inequality constraints that is enclosed in the folder *literature* in the zip-file) run the file *run.bat*. A pop-up window will appear in which you will see the iteration steps. Depending on your model complexity and the specification of your computer, the computation can take a while.

2.4 Output

After completing the computation two output files will be generated that contain the following information which is explained in the chapter about latent class models specified using inequality constraints that is enclosed in the folder *literature* in the zip-file:

output.txt

- the log marginal likelihood
- the weight per class
- the class specific probabilities

person.txt

- the first column gives the most likely class for the person at hand
- subsequent columns give posterior class probabilities for the person at hand

2.5 Model Comparison

To compare 2 different models you can calculate the posterior model probability (PMP) for each model, see the chapter about latent class models specified using

inequality constraints that is enclosed in the folder *literature* in the zip-file for more information. You can compare more than 2 models, but for now we compare the the model described above with a model that contains only 1 class and no inequality constraints between items (see subfolder *1_class_solution* for the files used).

Proceed as follows. First compute for each model the log marginal likelihood. In our example this is -21.085 for the 2-class solution and -21.745 for the 1-class solution (see subfolder *1_class_solution*).

Next, start the file *postprob.exe* and fill in the number of models under investigation and for each model the marginal likelihood. The results are the PMP's for every model. In our example the PMP for the 2-class model is 0.659 and for the 1-class model 0.340. These PMP's show that there is more support for the a 2-class model with the specified inequality constraints than the unconstrained 1-class solution.

3 Examples of the Book Chapter

In the following paragraphs the computation for some of the examples, used in the chapter about latent class models specified using inequality constraints that is enclosed in the folder *literature* in the zip-file, will be explained. Every model used in this manual has its own subfolder with all parameters pre-specified.

3.1 Example 1: Masculinity and Femininity

This example can be found in the folder *bipolar*. Traditionally masculinity and femininity are considered to be different extremes of the same (bipolar) dimension. However, this was criticized because there were no relevant variables that were related to either this bipolar dimension, masculinity or femininity. It was therefore suggested that masculinity and femininity should be treated as independent concepts.

One theory is the so called bipolar model. In Table 1 of the chapter about latent class models specified using inequality constraints that is enclosed in the folder *literature* in the zip-file, the bipolar-model is specified as followed: Class specific probabilities are represented by --, -, + and ++ such that $-- < - < + < ++$. The restrictions are only within items (that is, within a row of the table) and not between items (that is, not between rows of the table). The numbers 1 through 4 refer to four latent classes that are ordered along a bipolar dimension ranging from 'feminine' to 'masculine'. The class specific probabilities are restricted such that the probability of choosing feminine characteristics is decreasing from class 1 to 4, and the probability of choosing masculine characteristics is increasing from class 1 to 4. Starting values and inequality constraints should represent this ordering of this prior knowledge.

The file *input.txt* contains the following information:

<i>N</i>	<i>Nitem</i>	<i>Nclass</i>
166	8	4
<i>Starting values</i>		
.9	.9	.9
.7	.7	.7
.5	.5	.5
.3	.3	.3
.9	.3	.9
.7	.9	.5
.5	.7	.7
.3	.9	.9

The inequality constraints are specified as follows. The first item in the first class should have a higher probability than the first items in the other three classes. The input for this constraint is:

<i>Class</i>	<i>Item</i>	<i>Larger than</i>	<i>Smaller than</i>
1	1	3	0
<i>Larger than :</i>			
<i>Class</i>	<i>Item</i>		
2	1		
3	1		
4	1		
<i>Smaller than</i>			
<i>Class</i>	<i>Item</i>		

The same holds for items 2, 3 and 4. The first item in the second class should have a higher probability than the first items in classes 3 and 4, but a lower probability than the first item in class 1. The input for this second constraint is:

<i>Class</i>	<i>Item</i>	<i>Larger than</i>	<i>Smaller than</i>
2	1	2	1
<i>Larger than :</i>			
<i>Class</i>	<i>Item</i>		
3	1		
4	1		
<i>Smaller than</i>			
<i>Class</i>	<i>Item</i>		
1	1		

All other inequality constraints are specified following the same procedure. Do not forget to end the file with a blank line and with on the last row "END OF FILE".

3.2 Example 2: Balance Task

The files for the analysis of the example of the balance scale task can be found in the subfolder *twodim*.

References

Hoijtink, H. (1998). Constrained latent class analysis using the gibbs sampler and posterior predictive *p*-values: Applications to educational testing. *Statistica Sinica*, 8, 691-712.

Hoijtink, H. (2000). Posterior inference in the random intercept model based on samples obtained with markov chain monte carlo methods. *Computational Statistics*, 3, 315-336.

Hoijtink, H. (2001). Confirmatory latent class analysis: Model selection using bayes factors and (pseudo) likelihood ratio statistics. *Multivariate Behavioral Research*, 36, 563-588.

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the gibbs sampler and posterior predictive checks. *Psychometrika*, 62, 171-190.