Modelling the latent class structure of multiple Likert items: a paired comparison approach

Modellazione della struttura a classi latenti di item multipli su scala Likert: un approccio basato sul confronto a coppie

Brian Francis, Lancaster University, UK

Abstract The modelling of the latent class structure of multiple Likert items measured on the same response scale can be challenging. The standard latent class approach is to model the absolute Likert ratings, where the logits of the profile probabilities for each item have an adjacent category formulation (DeSantis et al., 2008). We instead propose modelling the relative orderings, using a mixture model of the *relative* differences between pairs of Likert items. This produces a paired comparison adjacent category log-linear model (Dittrich et al., 2007; Francis and Dittrich, 2017), with item estimates placed on a (0,1) "worth" scale for each latent class. The two approaches are compared using data on environmental risk from the International Social Survey Programme, and conclusions are presented.

Abstract *La modellazione della struttura a classi latenti di item multipli misurati sulla stessa scala Likert pu essere problematica. Lapproccio classico a classi latenti si basa sulla modellazione dei punteggi assoluti tramite una formulazione per categorie adiacenti dei logit delle probabilit dei vari profili di risposta per ciascun item (De Santis et al, 2008). In questo contributo si propone, invece, di modellare gli ordinamenti degli item tramite un modello mistura specificato sulle differenze relative tra coppie di item. Il modello risultante un modello log-lineare per categorie adiacenti basato su confronti a coppie (Dittrich et al, 2007; Francis and Dittrich,2017), dove le stime degli item sono poste su una scala di merito (0, 1) per ogni classe latente. I due approcci vengono confrontati e debitamente commentati utilizzando dati relativi al rischio ambientale, provenienti dallInternational Social Survey Programme.*

Key words: latent class, ordinal data, multiple Likert items, paired comparisons

Brian Francis

Department of Maths and Statistics, Fylde College. Lancaster University, Lancaster, LA1 4YF, UK
e-mail: B.Francis@Lancaster.ac.uk

1 Introduction

Collections of multiple Likert items in questionnaires are very common, and are usually used to measure underlying constructs. Scale from the Likert items can be built either through simply adding the item score or through using an IRT model such as a graded response model to build a score. This approach assumes that there is a single underlying construct to the items. The current paper, in contrast, takes a different view. It proposes that there is a latent class structure to the Likert items, with different classes having different patterns of high and low responses. In this approach, score building is not the aim; instead the aim is to understand the various patterns of responses that might exist in the population.

The standard latent class approach to multiple ordinal indicators essentially constructs a polytomous latent class model (Linzer and Lewis, 2011), and constrains the latent class profile probabilities, imposing a linear score ordinal model on them (Magidson and Vermunt, 2004; DeSantis et al., 2008). This results in a latent class adjacent category ordinal model. The method however uses the *absolute* responses, and this has been criticised by some authors, as they state that each respondent has their own way of interpreting the Likert scale. Such interpretation may itself be culturally determined, or may depend on other covariates such as age, gender and so on. For example younger people and males may be more likely to express a firm opinion, using the end categories of a unipolar Likert scale, than older people and females. The alternative is to take a relative approach. While one method of doing this is to standardise the items for each respondent, subtracting the respondent mean. This is unsatisfactory as it ignores the categorical nature of the data. In this paper we instead develop a paired comparisons approach, which produces a worth scale for each latent class, ranking the items in order of preference. The paper compares the two methods and discusses the advantages and disadvantages of each method.

Some common notation is introduced which will be used to develop both models. The Likert items are assumed to be measured on the same response scale with identical labelling; it is assumed that there are H possible ordered response categories taking the values $1,\ldots,H$ for each of the J Likert items indexed by j, and with N respondents indexed by i. $y_{ij};\quad y_{ij} \in 1,2,\ldots,H$ is defined to be the (ordinal) response given by respondent i to item j. A set of H indicators for each item and respondent with the indicator z_{ijh} taking the value 1 if $y_{ij} = h$ and 0 otherwise.

2 The ordinal latent class model

We first introduce the ordinal latent class model, which models the absolute responses. Let y_{ij} be the ordinal response of respondent i to item j. It is assumed that there are K latent classes. The item response vector for respondent i is

$$\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots, y_{iJ}),$$

Then the ordinal latent class model is defined by:

$$P(\mathbf{y_i}) = \sum_{k=1}^{K} \pi(k) P(\mathbf{y_i}|\mathbf{k})$$

$$= \sum_{k=1}^{K} \pi(k) \prod_{j} P(y_{ij}|k) \qquad \text{under conditional independence.}$$

We write

$$P(y_{ij}|k) = \prod_{h=1}^{H} p_{jkh}^{z_{ijh}}$$

where p_{jkh} is the probability of observing the ordinal response h for indicator j given membership of latent class k - these are sometimes called the latent class profile probabilities.

Ordinality is imposed by using an adjacent categories ordinal model and we parameterise the model through regression parameters on the logit scale, which separates out the intercept parameter β_{jh} and the class specific parameters β_{jkh} for each item and response category.

$$\text{logit}(p_{jkh}) = \beta_{jh} + \beta_{jkh}$$

$$= \beta_{jh} + h\beta_{jk} \qquad \text{under a linear score model}$$

The likelihood L is then given by

$$L = \prod_{i} \sum_{k=1}^{K} \pi(k) P(\mathbf{y_i}|\mathbf{k}).$$

Model fitting is usually carried out by using the EM algorithm - details are given in Francis et al. (2010) and Aitkin et al. (2014). Determination of the optimal number of classes is commonly achieved by choosing that model which minimises an information criterion, although a wide variety of other methods have been proposed. We have used the BIC in this paper.

3 The latent class ordinal paired comparison model

An alternative to the absolute latent class approach is to work on a **relative scale**. This perhaps is of greater interest. We take a paired comparison approach, using the difference in the ordinal likert responses. This allows the development of a "worth" scale between 0 and 1 awith items placed on this scale. The sum of the item scores is defined to be 1. This section proceeds by developing the ordinal paired comparison

model, and then extends that model by adding a mixture or latent class process to the model.

3.1 The ordinal paired comparison model

This model starts by constructing a set of paired comparisons - taking all possible pairs of items and comparing them in turn (Dittrich et al., 2007). For respondent i and for any two items $j = a$ and $j = b$, let

$$x_{i,(ab)} = \begin{cases} h & \text{if item } a \text{ preferred by } h \text{ steps to item } b = y_{ia} - y_{ib} \\ 0 & \text{if Likert ratings are equal} = 0 \\ -h & \text{if item } b \text{ preferred by } h \text{ steps to item } a = y_{ia} - y_{ib} \end{cases}$$

The probability for a single PC response $x_{i,(ab)}$ is then defined by

$$p(x_{i,(ab)}) = \begin{cases} \mu_{ab} \left(\frac{\pi_a}{\pi_b} \right)^{x_{i,(ab)}} & : & \text{if } x_{i,(ab)} \neq 0 \\ \mu_{ab} \, c_{ab} & : & \text{if } x_{i,(ab)} = 0 \end{cases}$$

The πs represent the worths or importances of the items, c_{ab} represents the probability of no preference between items a and b and μ_{ab} is a normalising quantity for the comparison ab. Over all items, we now form a pattern vector $\boldsymbol{x}_i$ for observation i with $\boldsymbol{x}_i = (x_{i,(12)}, x_{i,(13)}, \ldots, x_{i,(ab)}, \ldots, x_{i,(J-1,J)})$ and count up the number of responses n_ℓ with that pattern. The **probability** for a certain pattern ℓ is

$$p_\ell = \triangle^* \prod_{a<b} p(x_{ab})$$

where $\triangle^*$ is a constant (the same for all patterns). A log-linear model can now be constructed with observed counts n_ℓ. The **expected counts for a pattern** ℓ are defined as $\boxed{m_\ell = n \, p_\ell}$ where n is the total number of respondents defined by $n = n_1 + n_2 + \cdots + n_\ell + \cdots + n_L$ and where L is the number of all possible patterns.

Taking natural logs, the **log expected counts** are obtained by

$$\ln m_\ell = \alpha + \sum_{a<b} x_{ab}(\lambda_a - \lambda_b) + \mathbf{1}_{x_{ab}=0} \, \gamma_{ab}$$

For $x_{ab} = h$ this is $h(\lambda_a - \lambda_b)$, for $x_{ab} = -h$ this is $h(-\lambda_a + \lambda_b)$ and for $x_{ab} = 0$ this is γ_{ab}.

To show that this is an adjacent categories model, the log odds of a pair for any two adjacent categories on the ordinal scale can be examined - say h and $h+1$. Then, as $m_\ell = n p_\ell$, we have

$$\ln \left(\frac{m_\ell(h)}{m_\ell(h+1)} \right) = \ln(\mu_{ab}) + h(\lambda_a - \lambda_b) - \ln(\mu_{ab}) - (h+1)(\lambda_a - \lambda_b)$$
$$= \lambda_a - \lambda_b$$

which is true for any h as long as h or $h+1$ are not zero.

The worths π_j are calculated from the λ_j through the formula

$$\pi_j = \frac{\exp(2\lambda_j)}{\sum_{j=1}^{J} \exp(2\lambda_j)}$$

.

3.2 *Extending the model to incorporate latent classes*

As before, we assume that there are K latent classes with different preference patterns (the lambdas). The likelihood L becomes:

$$L = \prod_{\ell} \left(\sum_{k=1}^{K} q_k \, n \, p_{\ell k} \right) \quad \text{where} \sum_{\ell} p_{\ell k} = 1 \quad \forall k \text{ and } \sum_{k} q_k = 1.$$

$$\ln p_{\ell k} = \alpha + \sum_{a<b} x_{ab}(\lambda_{ak} - \lambda_{bk}) + \mathbf{1}_{x_{ab}=0} \, \gamma_{ab}$$

λ_j is replaced in the model by λ_{jk}, and we now have to additionally estimate the q_k. q_k is the probability of belonging to class k (the mass points or class sizes). Again, we use the EM algorithm to maximise the likelihood, and use the BIC to determine the number of classes. Typically, we need to use a range of starting values to ensure an optimal solution.

4 An Example

Six question items on the topic of environmental danger were taken from the 2000 sweep of the International Social Survey Programme , which focused on issues relating to the environment. As part of this survey, the respondents assessed the environmental danger of a number of different activities and items. The question is reproduced below; each question used the same response scale. The six Likert items are:

c air pollution caused by **c**ars (CAR)
t a rise in the world's **t**emperature (TEMP)
g modifying the **g**enes of certain crops (GENE)
i pollution caused by **i**ndustry (IND)
f pesticides and chemicals used in **f**arming (FARM)
w pollution of **w**ater (rivers, lakes, …) (WATER)

with the response scale for each of the items as follows:

In general, do you think *item* is

4. extremely dangerous for the environment
3. very dangerous
2. somewhat dangerous
1. not very dangerous
0. not dangerous at all for the environment

Table 1 BIC values from fitting latent class models (a) the standard ordinal LC model and (b) the ordinal PC LC model

	(a) standard ordinal LC model absolute		(b) Ordinal PC LC model relative	
No. of classes K	BIC	no of parameters	BIC	no of parameters
1	24207.04	24		
2	22680.48	31	6823.11	26
3	22153.75	38	6359.56	32
4	22112.70	45	6204.76	38
5	22097.07	52	6303.71	44
6	22084.99	59		
7	22083.33	66		

Both absolute and relative latent class models are fitted to this data. The standard ordinal latent class model (absolute) was fitted using Latent Gold 5.1 (Vermunt and Magidson, 2013), and the paired comparison ordinal latent class model (relative) was fitted using an extension to the `prefmod` package in R (Hatzinger and Maier, 2017). Both approaches used 20 different starting values to ensure that the global maximum of the likelihood was reached. Table 1 shows the BIC values for both models, for a range of values of K. It can be seen that the standard latent class approach needs either six or seven classes (six classes is chosen here), whereas the paired comparison latent class model gives a minimum BIC for $K = 4$. The smaller number of classes found for the paired comparison approach is perhaps to be expected, as the standard approach needs to model both the absolute level of the Likert responses as well as the differences.

We examine the mean Likert rating for each of the items within each of the latent classes for the standard ordinal latent class model. In contrast, the worths provide the interpretation of the latent classes in the paired comparison LC model. Both plots are shown in Figure 1, which are oriented so that greater dangerousness (or greater danger worth) is towards the top of the plots.

It can be seen that for the standard ordinal latent class model, the first three classes - Class 1 (51%), Class 2 (24%) and Class 3 (12%) - all show little difference between the items, but differ according to their absolute level. The three remaining classes, in contrast, show considerable differences between the items. The paired comparison solution gives a similar story. The largest class shows little difference between the items, with the three remaining classes showing large differences in
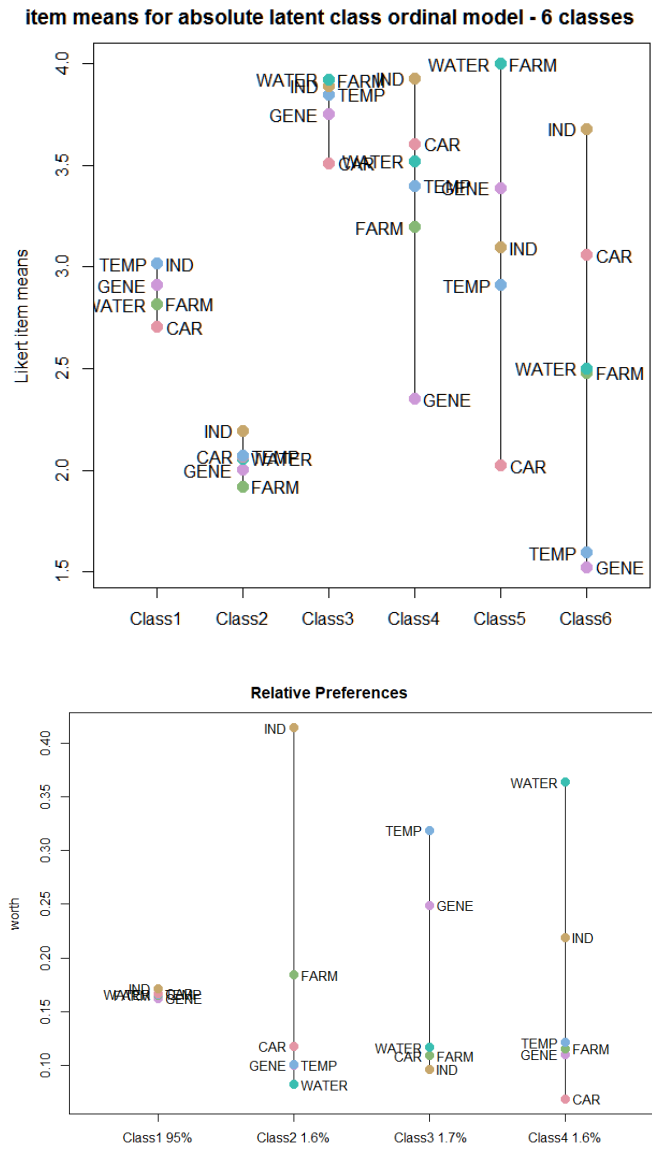
item means for absolute latent class ordinal model - 6 classes

Relative Preferences

Fig. 1 Item worths for (top) standard ordinal LC model and (bottom) ordinal paired comparison LC model

dangerousness between items. Although the item rankings show some minor differences between the two methods, the results are similar.

5 Discussion and conclusions

This paper has demonstrated that the paired comparison ordinal model can be useful to understand the relative ordering of items in multiple Likert responses when the absolute level of the response is not of interest. The method leads to simpler models, which makes interpretation simpler. There are however some restrictions in using the model. The most important is that all Likert items must be measured on the same response scale. Differences between Likert items only make sense when this is true, and the paired comparison method relies on that. The PC method as currently implemented also assumes equidistance between the Likert categories, and further work is needed to relax this assumption.

References

Aitkin, M., D. Vu, and B. Francis (2014). Statistical modelling of the group structure of social networks. *Social Networks 38*, 74–87.

DeSantis, S. M., E. A. Houseman, B. A. Coull, A. Stemmer-Rachamimov, and R. A. Betensky (2008). A penalized latent class model for ordinal data. *Biostatistics 9*(2), 249–262.

Dittrich, R., B. Francis, R. Hatzinger, and W. Katzenbeisser (2007). A Paired Comparison Approach for the Analysis of Sets of Likert Scale Responses. *Statistical Modelling 7*, 3–28.

Francis, B. and R. Dittrich (2017). Modelling multiple likert items through an adjacent categories ordinal paired comparison model. Presented at the 10th ERCIM comference on computational and Methodological Statistics, London; 16-18 December 2017.

Francis, B., R. Dittrich, and R. Hatzinger (2010). Modeling heterogeneity in ranked responses by nonparametric maximum likelihood: How do europeans get their scientific knowledge? *The Annals of Applied Statistics 4*(4), 2181–2202.

Hatzinger, R. and M. J. Maier (2017). *prefmod: Utilities to Fit Paired Comparison Models for Preferences*. R package version 0.8-34.

Linzer, D. and J. Lewis (2011). poLCA: An R Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software 42*(10).

Magidson, J. and J. Vermunt (2004). Latent class analysis. In D. Kaplan (Ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences*, Chapter 10, pp. 175–198. Thousand Oaks, CA.: Sage Publications.

Vermunt, J. and J. Magidson (2013). *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*. Belmont, MA: Statistical Innovations Inc.