


Smart Statistics for Smart Applications

Book of Short Papers SIS2019



Editors: Giuseppe Arbia, Stefano Peluso,
Alessia Pini and Giulia Rivellini

Copyright © 2019

PUBLISHED BY PEARSON

WWW.PEARSON.COM

Giugno 2019 ISBN 9788891915108

A statistical learning approach to group response categories in questionnaires

Un approccio basato sull'apprendimento statistico per raggruppare le categorie di risposta nei questionari

Michela Battauz

Abstract Questionnaires still represent one of the main sources of data collection. The possibility of administering them online, eventually through mobile devices, has increased the amount of information collected. This requires the use of novel statistical tools, able to treat large amounts of data. In this paper we present a procedure to group the response categories of the items of a questionnaire, thus leading to a more parsimonious model. To this end, the nominal response model with a lasso-type penalty is employed. The proposal is illustrated through an application to data on the satisfaction with town services.

Abstract *I questionari rappresentano ancora una delle principali fonti di raccolta di dati. La possibilità di somministrazione online, eventualmente attraverso dispositivi mobili, ha aumentato la quantità di informazione raccolta. Questo richiede l'uso di nuove tecniche statistiche, capaci di trattare grandi quantità di dati. In questo articolo presentiamo una procedura per raggruppare le categorie di risposta delle domande di un questionario, portando quindi a un modello più parsimonioso. A tal fine, si impiegherà il modello di risposta nominale con una penalizzazione di tipo lasso. La proposta è illustrata attraverso un'applicazione a dati sulla soddisfazione dei servizi offerti da una città.*

Key words: fused lasso, item response theory, lasso, penalized likelihood, regularization.

1 Introduction

Item response theory (IRT) includes various statistical models for the analysis of the responses given to a test or questionnaire [1]. In these models, the probability of

Michela Battauz

Department of Economics and Statistics, University of Udine, via Tomadini 30/A Udine e-mail: michela.battauz@uniud.it

giving a certain response depends on a latent variable of interest and some parameters related to the items. Some IRT models are suited for the analysis of polytomous items. Among these, the graded response model [6] and the generalized partial credit model [4] assume that the response categories have a predetermined order. Instead, the nominal response model [2] does not require the ordering of the response options. However, it is not only used for modelling unordered responses, but it is also useful to check the expected order of the categories [7]. It is certainly the most flexible IRT model for polytomous items, but involves the estimation of many parameters that can be very unstable in small samples. In this paper, we propose the use of a lasso-type penalty [3, 9] to group the response categories and provide regularized estimates. The methodology will be presented in Section 2, and it will be illustrated through a real-data example in Section 3. Some concluding remarks will be given in Section 4.

2 Regularized estimation of the nominal response model

Suppose that item j has m_j possible responses, which are indicated with $k = 0, \dots, m_j - 1$. In the nominal response model, the probability of giving the response k to item j for subject i is given by:

$$P(Y_{ij} = k | \theta_i) = \frac{e^{\alpha_{jk}\theta_i + \beta_{jk}}}{\sum_{h=0}^{m_j-1} e^{\alpha_{jh}\theta_i + \beta_{jh}}}, \quad (1)$$

where θ_i represents a latent variable, while α_{jk} and β_{jk} are parameters. The slope parameters α_{jk} capture the relation between the latent variable and the probability of giving response k , and higher values of this parameter indicate that a certain response is more likely to be given by subjects with higher values of the latent variable. When the slope parameters of two response categories of the same item are equal, these can be collapsed [8]. Instead, the intercept parameters β_{jk} are related to the number of subject giving response k for item j . The parameters of the model are usually estimated by means of the marginal maximum likelihood method, maximizing the following log-likelihood function:

$$\ell(\alpha, \beta) = \sum_{i=1}^n \sum_{j=1}^J \log \int_{\mathbb{R}} \prod_{k=0}^{m_j-1} P(Y_{ij} = k | \theta_i)^{I(Y_{ij}=k)} \phi(\theta_i) d\theta_i, \quad (2)$$

where n is the number of subjects, J is the number of items, $I(\cdot)$ is an indicator function, and $\phi(\cdot)$ is the density of a standard normal variable. In order to achieve a more parsimonious model and to limit the variability of the parameter estimates, the proposal is to include a penalty term in the log-likelihood function that forces the slope parameters of the same item to assume the same value. The penalized log-likelihood function is as follows:

A statistical learning approach to group response categories in questionnaires

$$\ell_p(\alpha, \beta) = \ell(\alpha, \beta) - \lambda \sum_{j=1}^J \sum_{k=0}^{m_j-2} \sum_{h=k+1}^{m_j-1} |\alpha_{jk} - \alpha_{jh}|. \quad (3)$$

The penalty in (3) is similar to a fused-lasso penalty [10]. However, in this case, there is not a natural order of the parameters, and hence all the pairs of slope parameters of the same item should be considered. Similarly to [11], we explored also a adaptive version of the penalty that includes weights that depend on the data:

$$\ell_p(\alpha, \beta) = \ell(\alpha, \beta) - \lambda \sum_{j=1}^J \sum_{k=0}^{m_j-2} \sum_{h=k+1}^{m_j-1} |\alpha_{jk} - \alpha_{jh}| w_{jkh}, \quad (4)$$

with

$$w_{jkh} = |\hat{\alpha}_{jk}^{MLE} - \hat{\alpha}_{jh}^{MLE}|^{-1}, \quad (5)$$

where $\hat{\alpha}_{jk}^{MLE}$ denotes the maximum likelihood estimate. When the maximum likelihood estimates are very close to each other, these weights become very high thus encouraging a fusion between the parameters.

3 A real-data example

The proposed methodology was applied to the 2015 Chapel Hill Community Survey¹, designed to investigate the satisfaction of the residents with the town services. The questionnaire is composed of 18 items with 5 possible response options, which are: *very dissatisfied*, *dissatisfied*, *neutral*, *satisfied*, *very satisfied*. The sample size is equal to 407. The survey explores the satisfaction of the residents with respect to a variety of aspects of the public services including, for example, safety, parks and recreation programs, library services, maintenance of streets, maintenance of buildings, or flow of traffic. Overall, the satisfaction with the public services is quite good: the median is equal to *satisfied* for most of the items; it is equal to *neutral* for only 4 items and it is equal to *very satisfied* for 1 item.

All analyses were performed with the R software [5], employing the `mirt` package for the estimation of the nominal model with the maximum likelihood method, and functions written by the authors to implement the maximum penalized likelihood estimation proposed in this paper. The value of the tuning parameter λ was selected through 5-fold cross-validation, and it is represented by the vertical dotted line in Figure 1. The Figure shows the regularization path of two items taken as example, where the parameter estimates are plotted against increasing values of λ . Category *very dissatisfied* is not represented in the graph because it was the reference category, with coefficients set to zero to assure the identification of the parameters. Using the non-adaptive penalization, only the slope parameter of category *dissatisfied* of the item on the left panel is fused with the slope parameter of category

¹ <https://catalog.data.gov/dataset/community-survey-q1-overall-satisfaction-with-town-services>

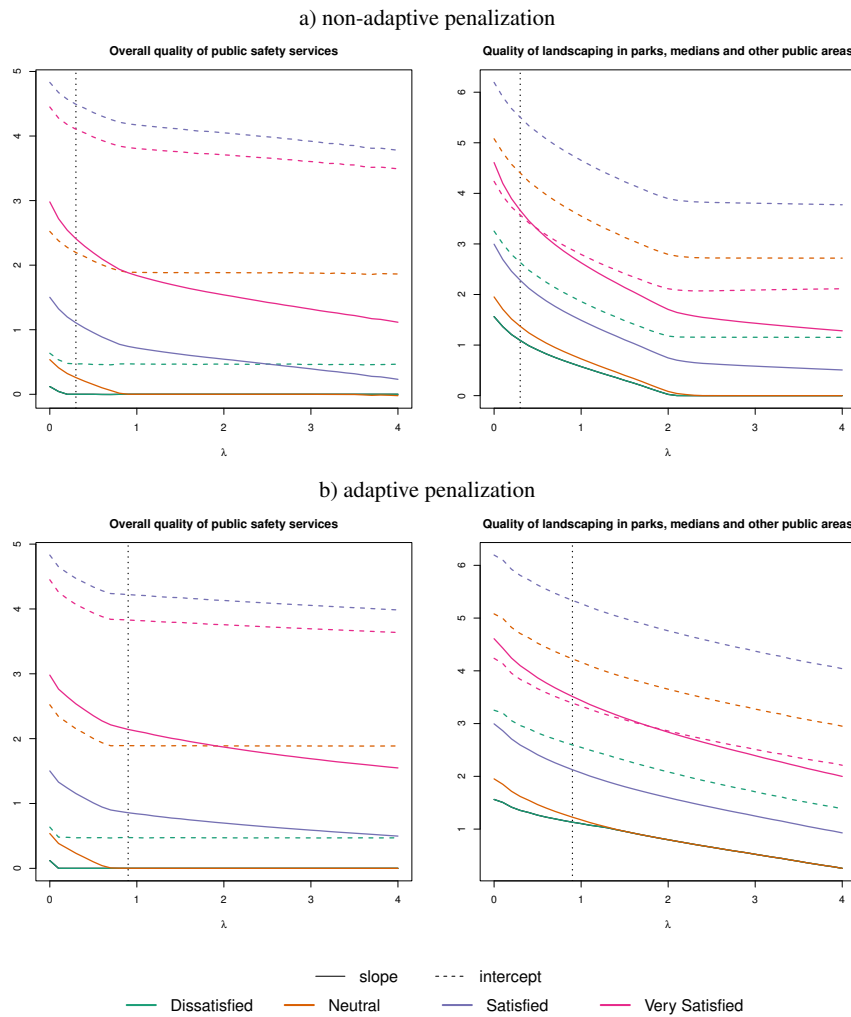


Fig. 1 Regularization path of two items.

very dissatisfied at the selected value of λ . Applying the adaptive penalization, the slope parameters of both categories *dissatisfied* and *neutral* are fused with the slope parameter of category *very dissatisfied* at the selected value of λ for the item on the left panel. Hence, these three categories of this item can be collapsed. The item on the right panel does not present slope parameters fused at the selected value of λ . However, all the parameters present shrunk values.

Figures 2 and 3 show the probability curves for the same two items when $\lambda = 0$, i.e. using the maximum likelihood estimates, and when $\lambda = 0.9$, which is the value selected by cross-validation using the adaptive penalization.

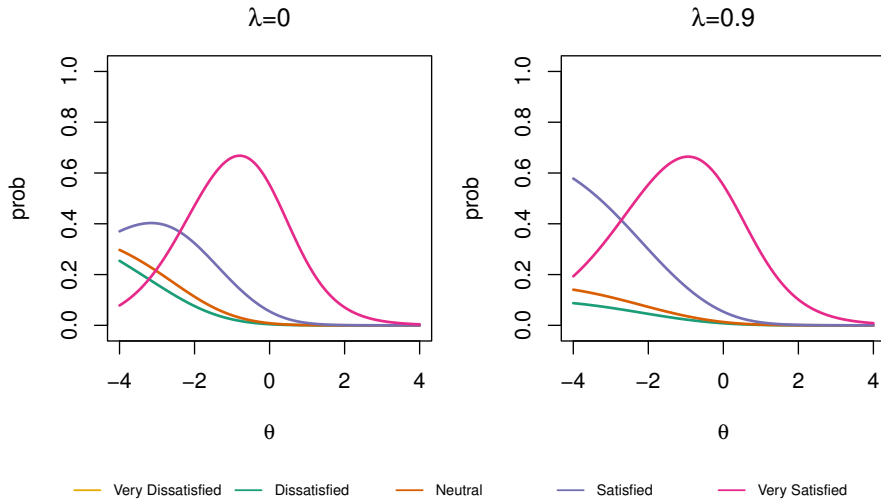


Fig. 2 Probability curves of item *Overall quality of public safety services*.

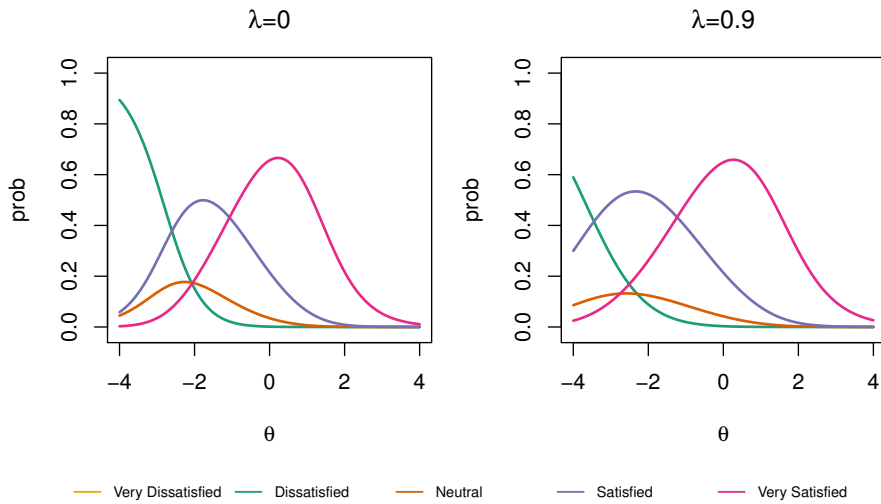


Fig. 3 Probability curves of item *Quality of landscaping in parks, medians and other public areas*.

4 Conclusions

The procedure proposed is quite effective in grouping the response categories that present a similar relation with the latent variable, especially using the adaptive version of the penalty. These categories can be collapsed, thus leading to a more parsimonious model. Since this procedure is meant to remove the noise eventually

present in the data, it does not lead to a loss of information when the response categories are grouped. In this respect, it is very important the selection of the tuning parameter λ , and cross-validation is a method suited to select a model that describes the data generating process. Besides of indicating the categories that can be collapsed, the method provides also regularized estimates and hence it represents a promising method for improving the efficiency of the estimators. A simulation study to better understand the performance of the method is under study.

Acknowledgements This work was partially supported by PRIN 2015 prot. 2015EASZFS.003 and partially by PRID 2017, University of Udine.

References

1. Bartolucci, F., Bacci, S., Gnaldi, M.: Statistical analysis of questionnaires: A unified approach based on R and Stata. Chapman and Hall/CRC, Boca Raton (2015)
2. Bock, D. R.: Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, **37**, 29–51 (1972)
3. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman and Hall/CRC, New York (2015)
4. Muraki, E.: A generalized partial credit model: Application of an EM algorithm. *Appl. Psychol. Meas.*, **16**, 159–176 (1992)
5. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018)
6. Samejima, F.: Estimation of ability using a response pattern of graded scores. *Psychometrika*, **34**, 1–97 (1969)
7. Thissen, D., Cai, L., & Bock, R. D.: The nominal categories item response model. In: Nering, M. L., Ostini, R. (eds.) *Handbook of Polytomous Item Response Theory Models*, pp. 43–75. Routledge (2010).
8. Thissen, D. & Cai, L.: Nominal categories models. In: van der Linden, W. J. (ed.) *Handbook of Item Response Theory, Volume One: Models*, pp. 51–73. Chapman and Hall/CRC (2016).
9. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., B*, **58**, 267–288 (1996)
10. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K.: Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc., B*, **67**, 91–108 (2005)
11. Zou, H.: The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.*, **101**, 1418–1429 (2006)