CLADAG 2021

BOOK OF ABSTRACTS AND SHORT PAPERS

13th Scientific Meeting of the Classification and Data Analysis Group
Firenze, September 9-11, 2021

edited by
Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

FIRENZE UNIVERSITY PRESS

ISSN 2704-601X (PRINT) - ISSN 2704-5846 (ONLINE)

SCIENTIFIC PROGRAM COMMITTEE

Giovanni C. Porzio (chair) (University of Cassino and Southern Lazio - Italy)

Silvia Bianconcini (University of Bologna - Italy)
Christophe Biernacki (University of Lille - France)
Paula Brito (University of Porto - Portugal)
Francesca Marta Lilja Di Lascio (Free University of Bozen-Bolzano - Italy)
Marco Di Marzio ("Gabriele d'Annunzio" University of Chieti-Pescara - Italy)
Alessio Farcomeni ("Tor Vergata" University of Rome - Italy)
Luca Frigau (University of Cagliari - Italy)
Luis Ángel García Escudero (University of Valladolid - Spain)
Bettina Grün (Vienna University of Economics and Business - Austria)
Salvatore Ingrassia (University of Catania - Italy)
Volodymyr Melnykov (University of Alabama - USA)
Brendan Murphy (University College Dublin -Ireland)
Maria Lucia Parrella (University of Salerno - Italy)
Carla Rampichini (University of Florence - Italy)
Monia Ranalli (Sapienza University of Rome - Italy)
J. Sunil Rao (University of Miami - USA)
Marco Riani (University of di Parma - Italy)
Nicola Salvati (University of Pisa - Italy)
Laura Maria Sangalli (Polytechnic University of Milan - Italy)
Bruno Scarpa (University of Padua - Italy)
Mariangela Sciandra (University of Palermo - Italy)
Luca Scrucca (University of Perugia - Italy)
Domenico Vistocco (Federico II University of Naples - Italy)
Mariangela Zenga (University of Milan-Bicocca - Italy)

LOCAL PROGRAM COMMITTEE

Carla Rampichini (chair) (University of Florence - Italy)

Chiara Bocci (University of Florence - Italy)
Anna Gottard (University of Florence - Italy)
Leonardo Grilli (University of Florence - Italy)
Monia Lupparelli (University of Florence - Italy)
Maria Francesca Marino (University of Florence - Italy)
Agnese Panzera (University of Florence - Italy)
Emilia Rocco (University of Florence - Italy)
Domenico Vistocco (Federico II University of Naples - Italy)

CLADAG 2021
BOOK OF ABSTRACTS AND SHORT PAPERS

13th Scientific Meeting of the Classification and Data Analysis Group
Firenze, September 9-11, 2021

edited by
Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

FIRENZE UNIVERSITY PRESS
2021

CLADAG 2021 BOOK OF ABSTRACTS AND SHORT PAPERS : 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021/ edited by Giovanni C. Porzio, Carla Rampichini, Chiara Bocci. — Firenze : Firenze University Press, 2021.
(Proceedings e report ; 128)

https://www.fupress.com/isbn/9788855183406

ISSN 2704-601X (print)
ISSN 2704-5846 (online)
ISBN 978-88-5518-340-6 (PDF)
ISBN 978-88-5518-341-3 (XML)
DOI 10.36253/978-88-5518-340-6

Graphic design: Alberto Pizarro Fernández, Lettera Meccanica SRLs
Front cover: Illustration of the statue by Giambologna, *Appennino* (1579-1580) by Anna Gottard

CLAssification and Data
Analysis Group (CLADAG)
of the Italian Statistical
Society (SIS)

FUP Best Practice in Scholarly Publishing (DOI https://doi.org/10.36253/fup_best_practice)
All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Boards of the series. The works published are evaluated and approved by the Editorial Board of the publishing house, and must be compliant with the Peer review policy, the Open Access, Copyright and Licensing policy and the Publication Ethics and Complaint policy.

The online digital edition is published in Open Access on www.fupress.com.

Published by Firenze University Press
Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

This book is printed on acid-free paper
Printed in Italy

BOOSTING MULTIDIMENSIONAL IRT MODELS

Michela Battauz [1] and Paolo Vidoni[1]

[1] Department of Economics and Statistics, University of Udine, (e-mail: michela.battauz@uniud.it, paolo.vidoni@uniud.it)

ABSTRACT: Multidimensional IRT models can be used to analyze the latent variables that underlay the responses given to a test or questionnaire. However, these models are not only difficult to estimate, but they also suffer of the rotational indeterminacy typical of factor analysis models. In this paper, we propose a boosting algorithm that, starting from a model that includes only the intercepts, sequentially updates a pair of coefficients in a component-wise approach. The solution provided by the algorithm tends to be sparse and to facilitate the interpretation without requiring a posterior rotation.

KEYWORDS: negative curvature direction, regularization, sparse solution.

1 Introduction

IRT models are commonly applied in educational assessment and they are also considered, with increasing frequency, in the field of health and psychological measurement studies. In these models, the probability of observing a categorical response is a function of a single latent trait (simple IRT models) or of multiple latent traits (multiple IRT models) and of some item parameters (see for example Reckase, 2009). Various methods have been proposed for model estimation. However, in the multidimensional setting, serious computational problems may occur if the number of items is large and many latent variables have to be considered. Moreover, in this context, the interpretability of the solution is very important.

In this paper, the new statistical boosting procedure introduced in Battauz & Vidoni (2021) is applied for estimating multiple IRT models. More precisely, we consider a suitable likelihood-based boosting algorithm which may escape from a region of local non-convexity of the objective function, improve the optimization procedure, provide a more interpretable sparse solution and regularize the estimates. We apply this new procedure to the multidimensional two-parameter logistic IRT model for dichotomously scored outcomes. An example concerning a sample from the 2017 Eurobarometer survey is presented.

2 Multidimensional IRT models: definition and inference

The response variable for the subject i on item j is a Bernoulli random variable Y_{ij}, $i = 1,\dots,n$, $j = 1,\dots,J$, with one denoting a positive response. The responses of subject i are collected in the vector $\mathbf{Y}_i = (Y_{i1},\dots,Y_{iJ})^\top$. Let $\theta_i = (\theta_{i1},\dots,\theta_{iD})^\top$, $i = 1,\dots,n$, be a latent random vector, composed of independent standard normal variables. Furthermore, it is assumed that $(\mathbf{Y}_i,\theta_i)$ are independent across subjects and that observations Y_{ij} are conditionally independent given θ_i. With particular attention to the multidimensional two-parameter logistic (2PL) IRT model, the conditional probability of giving a positive response to a specific item is defined as

$$P_{ij} = P(Y_{ij} = 1|\theta_i;\beta_j,\alpha_{1j},\dots,\alpha_{Dj}) = \frac{\exp(\beta_j + \alpha_{1j}\theta_{i1} + ,\cdots + \alpha_{Dj}\theta_{iD})}{1+\exp(\beta_j + \alpha_{1j}\theta_{i1} + ,\cdots + \alpha_{Dj}\theta_{iD})},$$

where β_j is the intercept and α_{dj}, $d = 1,\dots,D$, are the slope parameters. The vector of unknown model parameters is $\gamma = (\alpha_1^\top,\dots,\alpha_D^\top,\beta^\top)^\top$, with $\alpha_d = (\alpha_{d1},\dots,\alpha_{dJ})^\top$, $d = 1,\dots,D$, and $\beta = (\beta_1,\dots,\beta_J)^\top$; the vector γ has dimension $J + JD$, which, in some applications, can be very large.

Given the responses $\mathbf{y}$, realization of $\mathbf{Y} = (\mathbf{Y}_1^\top,\dots,\mathbf{Y}_n^\top)^\top$, the marginal likelihood for γ can be obtained by integrating out the unobserved θ values from the complete likelihood $L(\gamma;\mathbf{y}) = \prod_{i=1}^n f(\mathbf{y}_i|\theta_i;\gamma)\phi(\theta_i)$, where $f(\mathbf{y}_i|\theta_i;\gamma)$ is a Bernoulli-type probability function based on P_{ij} and $\phi(\cdot)$ denotes the density of a multivariate standard normal distribution with independent components. Thus, the marginal log-likelihood does not have a closed-form expression, since the D-dimensional integral does not have an analytic solution and requires numerical approximations. The most common methods for estimating the item parameters are based on the EM algorithm, approximating the integrals using Gaussian or adaptive quadrature procedures, or on suitable MCMC algorithms for handling with the high dimension of the integrals.

3 The boosting algorithm

We consider the boosting algorithm introduced in Battauz & Vidoni (2021), with the negative log-likelihood as objective function. Starting from a model that includes only the intercept terms, only two parameters are updated at each iteration of the algorithm, hence following a component-wise approach. The starting point of the algorithm poses a very challenging issue, since the gradient is null making any gradient descent method unable to move from it. A

peculiar feature of the method is that it exploits any local non-convexity of the objective function, since the gradient vector and the Hessian matrix are used to define two alternative directions. These are the classical Newton-type direction and a negative curvature direction given by the eigenvector associated with the most negative eigenvalue (if any) of a 2×2 submatrix of the Hessian matrix. More specifically, at step k of the boosting algorithm, the Newton-type direction for each pair of parameters indexed $b, c = 1, \ldots, J(D+1)$, $b < c$, is given by:

$$\mathbf{s}_{bc}^{(k)} = -\widehat{\mathbf{H}}_{bc}^{(k-1)^{-1}} \widehat{\mathbf{g}}_{bc}^{(k-1)}, \tag{1}$$

while the negative curvature direction is:

$$\mathbf{d}_{bc}^{(k)} = -sign\left\{ \left(\widehat{\mathbf{g}}_{bc}^{(k-1)} \right)^{\top} \widehat{\mathbf{u}}_{bc}^{(k-1)} \right\} \widehat{\mathbf{u}}_{bc}^{(k-1)}, \tag{2}$$

where $\widehat{\mathbf{g}}_{bc}^{(k-1)}$ and $\widehat{\mathbf{H}}_{bc}^{(k-1)}$ are the gradient and the Hessian computed at step $k-1$, and $\widehat{\mathbf{u}}_{bc}^{(k-1)}$ is the eigenvector corresponding to the minimum negative eigenvalue of $\widehat{\mathbf{H}}_{bc}^{(k-1)}$. The algorithm computes the variation of a quadratic approximation of the objective function for all the pairs of parameters in both the directions, and selects the one leading to the largest decrease. The algorithm represents a particular application of the optimization method proposed by Gould et al. (2000), who proved the convergence to second-order critical points. Since the algorithm converges to the maximum likelihood estimates, a suitable stopping criterion is necessary to obtain regularized estimates.

4 A real-data example

The proposal was applied to the responses of 1027 Italian citizens to some items of the 2017 Eurobarometer survey regarding the area that people thinks that the decisions should be made at the European level. Table 1 reports the items and the estimated parameters. The number of iterations of the algorithm as well as the number of latent variables were selected by 5-fold cross-validation. The table also reports the maximum likelihood estimates (MLEs) obtained with the R package mirt and using the quartimax rotation, which was chosen for the higher similarity of the solution. It is possible to observe that the MLEs tend to assume more extreme values, while the boosting procedure provides regularized estimates. Both the methods identify a first dimension strongly related to all the items. The interpretation of the second dimension seems a bit more clear using the boosting algorithm, since it reveals a positive

correlation between the areas of terrorism, immigration, democracy and peace (that present the highest estimated discrimination parameters). However, the areas of energy supply, environment, investment and job creation are also related to this dimension.

Table 1. *Items of the Eurobarometer survey included in the analysis and parameter estimates.*

		boosting			MLE		
QC7	Areas where more decision-making should take place at a European level	β_j	α_{1j}	α_{2j}	β_j	α_{1j}	α_{2j}
1	Fighting terrorism	3.07	4.02	1.61	6.10	-8.80	2.83
2	Dealing with health and social security issues	1.02	3.37	0.00	1.10	-3.55	-0.68
3	Promoting equal treatment of men and women	1.41	3.34	0.00	1.45	-3.37	-0.54
4	Promoting democracy and peace	1.95	2.99	0.92	2.08	-3.37	0.21
5	Securing energy supply	1.78	3.27	0.44	1.87	-3.49	-0.06
6	Dealing with migration issues from outside the EU	2.36	3.32	1.06	2.49	-3.73	0.33
7	Protecting the environment	2.41	4.74	0.58	2.46	-4.87	-0.38
8	Stimulating investment and job creation	1.80	4.41	0.74	2.02	-5.06	-0.24

References

GOULD, N. I. M., LUCIDI, S., ROMA, M., & TOINT, PH. L. 2000. Exploiting negative curvature directions in linesearch methods for unconstrained optimization. *Optimization Methods and Software*, **14**(1-2), 75–98.

MICHELA BATTAUZ, PAOLO VIDONI. 2021. A new likelihood-based boosting algorithm for factor analysis models with binary data. *Submitted*.

RECKASE, MARK D. 2009. *Multidimensional Item Response Theory Models*. New York, NY: Springer Verlag.