Quadratic Majorisation of the Rating Scale Model

Pieter Schoonees, Patrick Groenen & Kathrin Gruber

gruber@ese.eur.nl

International Workshop on Psychometric Computing
February 29th – March 1st, 2024

Verbal Aggression (Vansteelandt, 2000)
$n = 316$ persons, $m = 24$ items, $r = 3$ responses each

	Would you **do**/**want** **curse**/**scold**/**shout** in this situation?		
S1: A bus fails to stop for me.	no	perhaps	yes
S2: I miss a train because a clerk gave me faulty information.	no	perhaps	yes
S3: The grocery store closes just as I am about to enter.	no	perhaps	yes
S4: The operator disconnects me when I had used up my last 10 cents for a call.	no	perhaps	yes

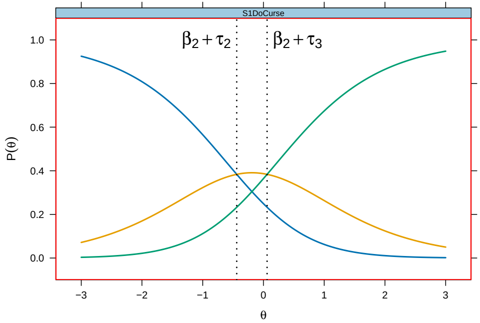
Verbal Aggression (Vansteelandt, 2000)
$n = 316$ persons, $m = 24$ items, $r = 3$ responses each

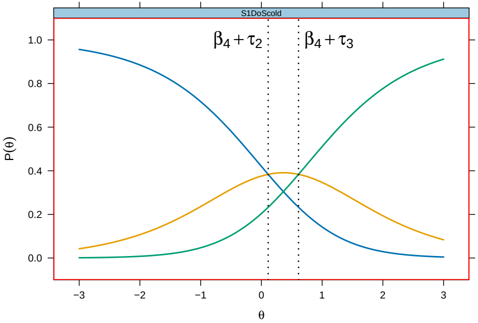
	Would you **do**/**want** **curse**/**scold**/**shout** in this situation?		
S1: A bus fails to stop for me.	no	perhaps	yes
S2: I miss a train because a clerk gave me faulty information.	no	perhaps	yes
S3: The grocery store closes just as I am about to enter.	no	perhaps	yes
S4: The operator disconnects me when I had used up my last 10 cents for a call.	no	perhaps	yes

Rating Scale Model

$$\text{logit } P(Y_{ij} = k \mid Y_{ij} \in \{k-1, k\}, \boldsymbol{\beta}, \theta_i) = \log \frac{P(Y_{ij} = k \mid \boldsymbol{\beta}, \theta_i)}{P(Y_{ij} = k-1 \mid \boldsymbol{\beta}, \theta_i)},$$

with person (trait) location θ_i and item locations $\boldsymbol{\beta}^\top = \{\beta_j + \tau_l\}_{j=1,l=1}^{m,r}$ (adjacent category probability formulation, Andrich, 1978).





Rating Scale Model (2)

$$\text{logit } P(Y_{ij} = k \mid Y_{ij} \in \{k-1, k\}, \theta_i, \boldsymbol{\beta}) = \log \frac{P(Y_{ij} = k \mid \theta_i, \boldsymbol{\beta})}{P(Y_{ij} = k-1 \mid \theta_i, \boldsymbol{\beta})},$$

with person (trait) location θ_i and item locations $\boldsymbol{\beta}^\top = \{\beta_j + \tau_l\}_{j=1,l=1}^{m,r}$ (adjacent-category logit formulation, Andrich, 1978).

Therefore,

$$P(Y_{ij} = k|\boldsymbol{\beta}, \theta_i) = \pi_{ijk}(\theta_i, \boldsymbol{\beta}) = \frac{\exp \sum_{l=1}^{r}(\theta_i - \beta_j - \tau_l)}{1 + \sum_{l=1}^{r} \exp \sum_{k=1}^{r}(\theta_i - \beta_j - \tau_k)}$$

subject to cornerpoint/identification constraint $\sum_{k=1}^{r}(\theta_i - \beta_j - \tau_k)$ for all i, j.

Joint Estimation

Full Likelihood

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r} y_{ijk} \log \pi_{ijk}(\theta_i, \boldsymbol{\beta})$$

Strategy (1): Maximise the full log-Likelihood **jointly** over all parameters (Wright & Panchapakesan, 1969; Wright & Douglas, 1977; Haberman, 1977).

- restriction (assumption) free
- (asymptotically) inconsistent item parameter estimates, problems in the normal approximation for person parameter estimates (Gilula & Haberman, 1994).
- mathematically convenient, relatively easy to implement

Conditional Estimation

Full Likelihood

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r} y_{ijk} \log \pi_{ijk}(\theta_i, \boldsymbol{\beta})$$

Strategy (2): Maximise the full log-Likelihood **conditional** on (a) $\theta_i = \sum_{j=1}^{m} Y_{ij}$ (Andersen, 1973; Fischer, 1981); (b) $\theta_i \sim \phi(\theta_i; 0, v)$ (Kiefer & Wolfowitz, 1956; Andersen & Madsen, 1977; Thissen, 1982).

- (strong) restrictive assumptions
- (asymptotically) consistent parameter estimates
- difficult to implement (computationally demanding)

Penalized Joint Estimation

Full Likelihood

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r} y_{ijk} \log \pi_{ijk}(\theta_i, \boldsymbol{\beta}) + \lambda \mathsf{Pen}(\cdot), \;\; \lambda \geq 0$$

Strategy (3): Maximise the joint log-Likelihood with $\mathbf{L}_2$**-penalty** $\mathsf{Pen}(\boldsymbol{\theta}) = \sum_i \theta_i^2$ (Hoerl & Kennard, 1970) and/or $\mathbf{L}_1$**-penalty** $\mathsf{Pen}(\boldsymbol{\beta}) = \sum_j \beta_j$ (Tibshirani, 1996; Zou & Hastie, 2005).

- moderate restrictions
- (asymptotically) consistent parameter estimates (comparable to marginal maximum likelihood estimation, Chen et al. 2019)

Penalized Joint Estimation (2)

Full Likelihood

$$\ell(\boldsymbol{\eta}) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r} y_{ijk}\log \pi_{ijk}(\boldsymbol{\eta}) + \lambda_{ij}\mathsf{Pen}(\boldsymbol{\eta}),\ \ \lambda_{ij} \geq 0$$

Extension: We maximise the joint log-Likelihood with item and person-specific **$\mathbf{L_2}$-penalty** $\mathsf{Pen}(\boldsymbol{\eta}) = \sum_{k=1}^{r} \eta_{ijk}^2$ with $\boldsymbol{\eta}^\top = \left\{\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top\right\}$.

- automatic incorporation of missing values
- fast converging optimization (majorization) algorithm

Iterative Majorization

$$\ell(\boldsymbol{\eta}) = \sum_i \sum_j \underbrace{\left(- \sum_k y_{ijk} \log \pi_{ijk}(\boldsymbol{\eta}) + \lambda_{ij} \sum_k \eta_{ijk}^2 \right)}_{h_{ij}(x)}, \;\; \lambda_{ij} \geq 0$$

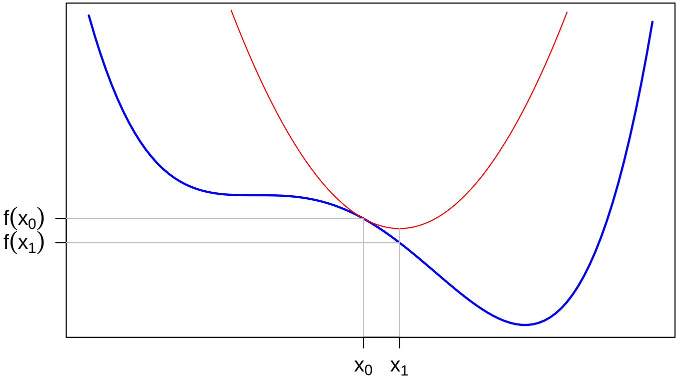
Majorize $\boldsymbol{f}(\boldsymbol{h}(\boldsymbol{x}))$ at support point $\boldsymbol{h}(\boldsymbol{x})$ with a simpler (quadratic) surrogate (De Leeuw & Heiser, 1980):

$$\boldsymbol{g}(\boldsymbol{h}(\boldsymbol{x}), \boldsymbol{h}(\boldsymbol{y})) = \boldsymbol{f}(\boldsymbol{h}(\boldsymbol{y})) + \frac{1}{2}(\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{x}_\star)^\top \boldsymbol{B}(\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{x}_\star) - \frac{1}{2}\partial \boldsymbol{f}(\boldsymbol{h}(\boldsymbol{y}))\boldsymbol{B}^{-1}\partial \boldsymbol{f}(\boldsymbol{h}(\boldsymbol{y}))$$

where $\boldsymbol{B} - \partial^2 \boldsymbol{f}(\boldsymbol{h}(\boldsymbol{y})) \geq 0$ and $\boldsymbol{x}_\star$ is the penalized least squares update:

$$\boldsymbol{x}_\star = \boldsymbol{h}(\boldsymbol{y}) - 2\boldsymbol{B}^{-1}\partial \boldsymbol{f}(\boldsymbol{h}(\boldsymbol{y})) \ .$$

Iterative Majorization (2)

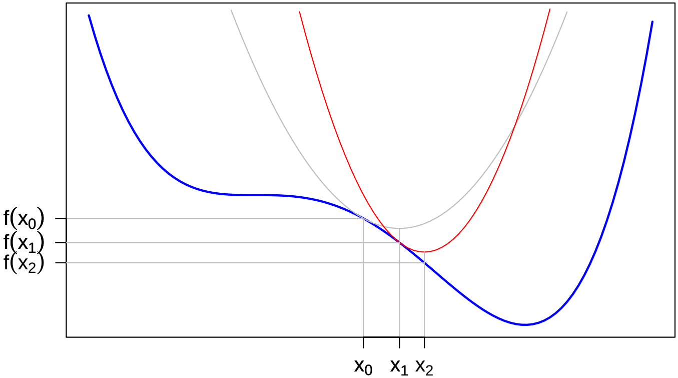


$\boldsymbol{f}(\boldsymbol{h}(\boldsymbol{y})) = \boldsymbol{g}(\boldsymbol{h}(\boldsymbol{y}), \boldsymbol{h}(\boldsymbol{y}))$
touch at support point $\boldsymbol{h}(\boldsymbol{y})$

$\boldsymbol{f}(\boldsymbol{h}(\boldsymbol{x})) \leq \boldsymbol{g}(\boldsymbol{h}(\boldsymbol{x}), \boldsymbol{h}(\boldsymbol{y}))$

(see also, Böhning & Lindsay, 1988;
Groenen, Mathar & Heiser, 1995)

Iterative Majorization (3)



Minimization succeeds with $\boldsymbol{g}(\boldsymbol{h}(\boldsymbol{x}), \boldsymbol{h}(\boldsymbol{x}_\star))$ over $\boldsymbol{h}(\boldsymbol{x})$

- (guaranteed) globally convergent
- no step-size specification required

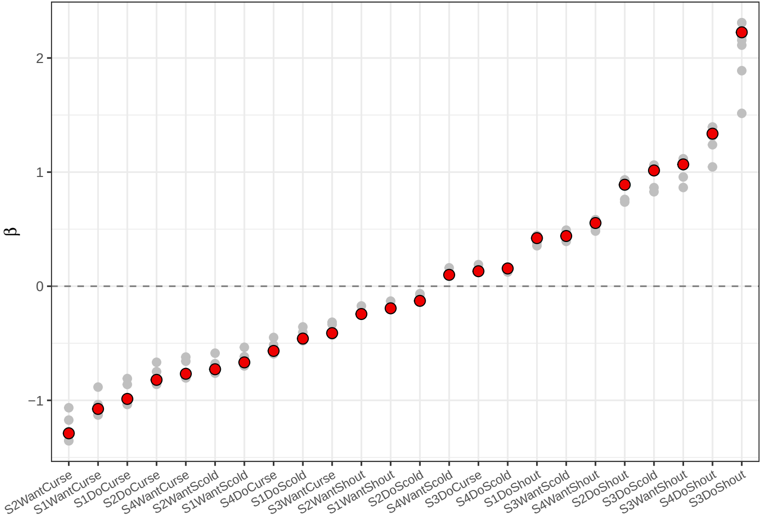
Polytomous IRT Models in R

- **eRm** (Mair & Hatzinger, 2007): Conditional maximum likelihood.
- **mirt** (Chalmers, 2012): Marginal maximum likelihood with Metropolis-Hastings Robbins-Monro integration algorithm.
- **brms** (Büerkner, 2021): Marginal maximum likelihood with exact (No U-Turn Sampler) and Variational Bayes integration.
- **irtmaj** (Schoonees, Groenen & Gruber, 2024): Penalized joint maximum likelihood with iterative majorisation.

Benchmark Results

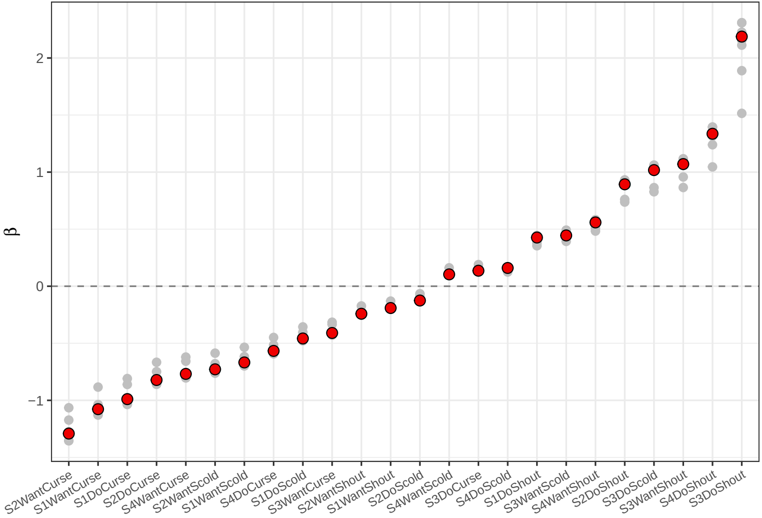
	Time (sec)	log-Likelihood	Settings
eRm	2.53	-5203.91	
mirt	1.75	-6345.84	
brms (NUTS)	1974.13	-6163.97	4 parallel chains, 4000 draws each
brms (VB)	50.55	-6233.80	meanfield
irtmaj	2.54	-5822.42	$\lambda = 0$, keep 0/full scores
	0.12	-5822.36	$\lambda = 0$, remove 0/full scores
	0.27	-5868.06	$\lambda = 0.001$, keep 0/full scores
	0.04	-6139.43	$\lambda = 0.01$, remove 0/full scores

Benchmark Results (2)



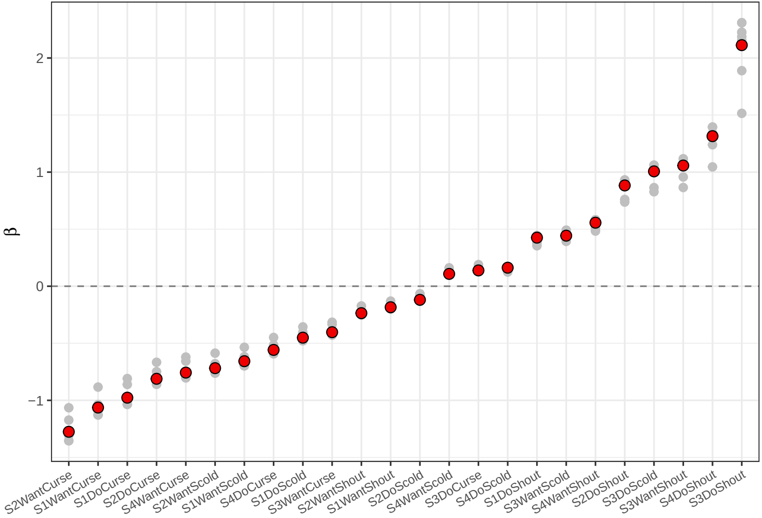
eRm

Benchmark Results (2)

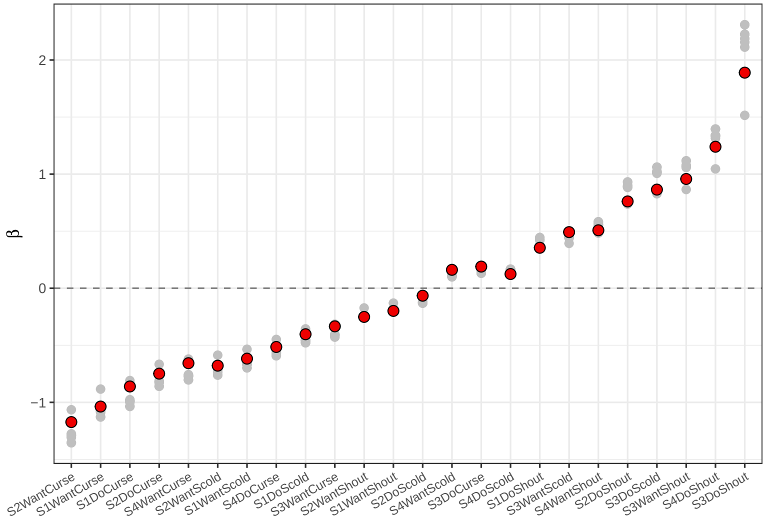


mirt

Benchmark Results (2)

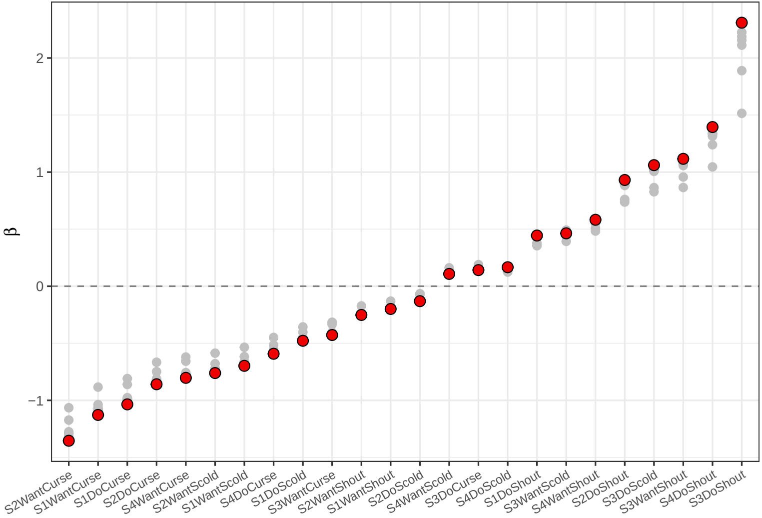


brms (NUTS)


β
2
1
0
−1
S2WantCurse
S1WantCurse
S1DoCurse
S2DoCurse
S4WantCurse
S2WantScold
S1WantScold
S4DoCurse
S1DoScold
S3WantCurse
S2WantShout
S1WantShout
S2DoScold
S4WantScold
S3DoCurse
S4DoScold
S1DoShout
S3WantScold
S4WantShout
S2DoShout
S3DoScold
S3WantShout
S4DoShout
S3DoShout

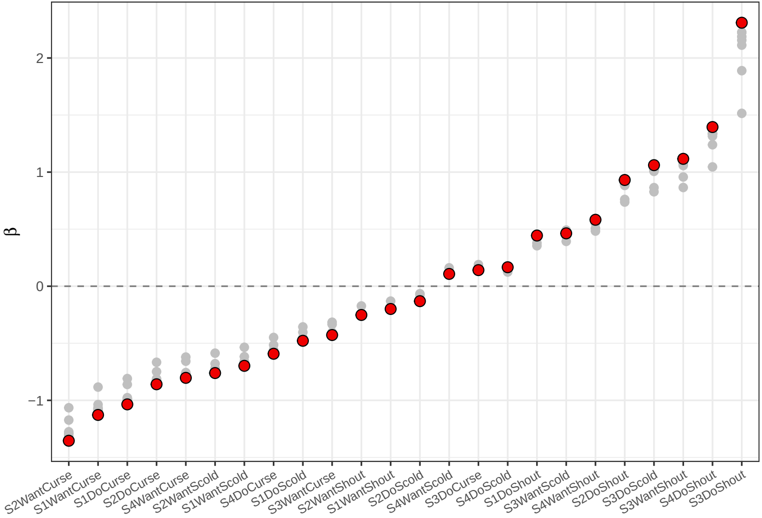
brms (VB)

Benchmark Results (2)



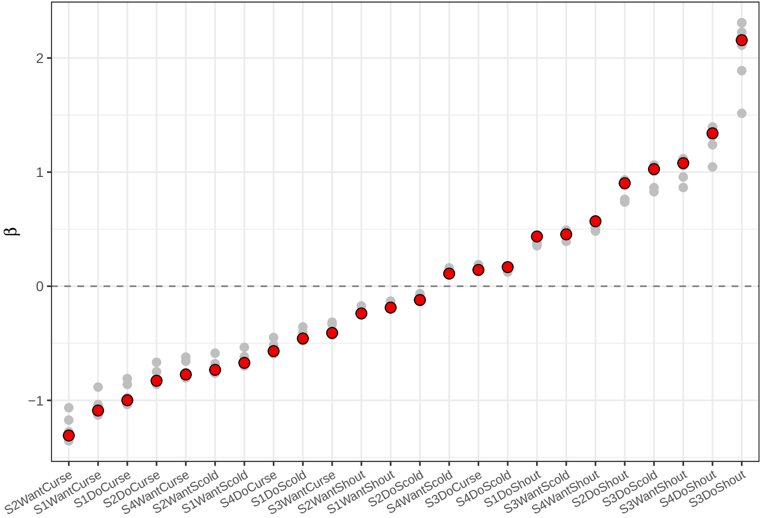
irtmaj
($\lambda = 0$,
keep 0/full scores)

Benchmark Results (2)



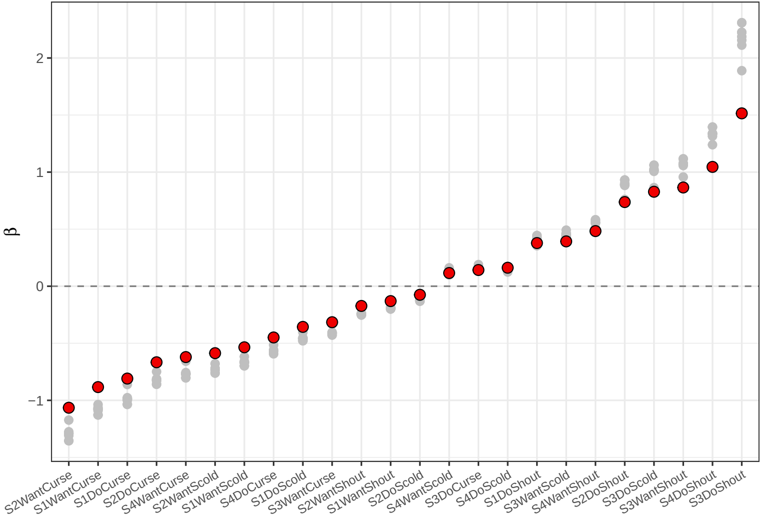
irtmaj
($\lambda = 0$,
remove 0/full scores)

Benchmark Results (2)



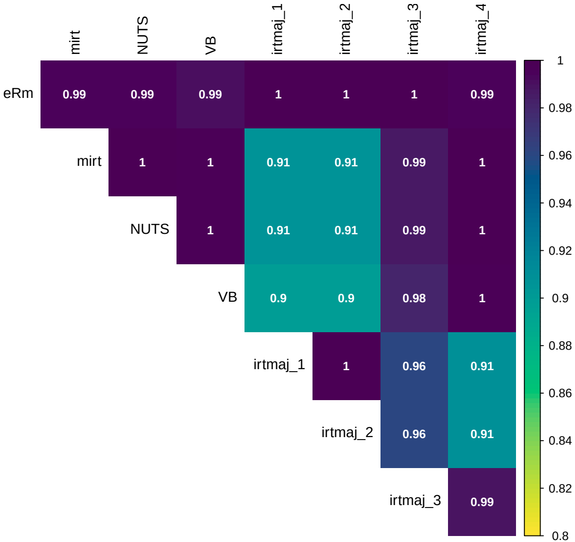
irtmaj
($\lambda = 0.001$,
keep 0/full scores)

Benchmark Results (2)



irtmaj
($\lambda = 0.01$,
remove 0/full scores)

Benchmark Results (3)



Pairwise $\widehat{\theta}_i$ correlations

Summary

Contribution

- Well-behaved, fast converging optimization algorithm for non-convex response functions.

Outlook

- Other models (binary outcomes, covariates for latent regression, multidimensional traits).
- Easy incorporation of constraints and thus, extensions to other penalties (e.g., for automatic dimensionality reduction).
- Pytorch for additional computational speed.

Literature

Andersen, E. B. (1973). Conditional inference and models for measuring. Copenhagen: Mental Hygiejnisk Forlag.

Andersen, E. & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika*, 42, 357-374.

Bürkner, P. (2021). Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5), 1–54.

Böhning, D. & Lindsay, B. G. (1988). Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4), 641-663.

Chalmers R.P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1–29.

Chen, Y., Li, X. & Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, 84, 124-146.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73, 533-559.

De Leeuw, J. & Heiser, W. J. (1980). Multidimensional scaling with restrictions on the configuration. In P. R. Krishnaiah (Ed.), *Multivariate Analysis* (Vol. V, pp. 501-522).

Literature

Fischer, G. H. (1981). On the existence and uniqueness of maximum likelihood estimates in the Rasch model. *Psychometrika*, 46, 59–77.

Gilula, Z. & Haberman, S. J. (1994). Models for analyzing categorical panel data. *Journal of the American Statistical Association*, 89, 645–656.

Groenen, P. J., Mathar, R., & Heiser, W. J. (1995). The majorization approach to multidimensional scaling for Minkowski distances. *Journal of Classification*, 12, 3-19.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, 5(5), 815-841.

Haberman, S. J. Joint and Conditional Maximum Likelihood Estimation for the Rasch Model for Binary Responses. ETS Research Report Series, 2004(1), i-63.

Hoerl, A. E. & Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 12, 55–67.

Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 887-906.

Literature

Mair, P. & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, 47, 175-186.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58 (1), 267–88.

Vansteelandt, K. (2000). Formal models for contextualized personality psychology. Unpublished doctoral dissertation.

Wright, B. D. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 29, 23-48.

Wright, B. D. & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement*, 1, 281-295.

Zou, H. & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B (Methodological)*, 67 (2), 301–320.

Appendix: Derivatives

First-order derivatives:

$$\frac{\partial \ell(\eta_{ij})}{\partial \eta_{ijk}} = \left(\pi_{ijk} \sum_{k=1}^{r} y_{ijk} - y_{ijk} \right) + 2\lambda_{ij}\eta_{ijk} \ .$$

Second-order derivatives (the $\boldsymbol{H}_{ij}$ entries):

$$\frac{\partial^2 \ell(\eta_{ij})}{\partial \eta_{ijk}} = \sum_{k=1}^{r} y_{ijk} \left(\delta^{kl}\pi_{ijk} - \pi_{ijk}\pi_{ijl} \right) + 2\lambda_{ij}\delta^{kl} \ ,$$

where δ^{kl} is the Kronecker delta.

Appendix: Majorization Strategy

Recall: $\boldsymbol{B}_{ij} - \boldsymbol{H}_{ij} \geq 0$. Therefore,

$$\boldsymbol{H}_{ij} \leq \frac{1}{2}\left(\sum_{k=1}^{r} y_{ijk} + 2\lambda_{ij}\right)\boldsymbol{I} = \boldsymbol{B}_{ij}$$

The diagonal matrix $\boldsymbol{B}$, with diagonal blocks $\{\boldsymbol{B}_{ij}\}_{i=1,j=1}^{n \quad m}$, then majorizes the full log-likelihood $\ell(\boldsymbol{\eta})$.