FECHNERIAN SCALING OF IRT MODELS FOR DICHOTOMOUS DATA

Thomas Kiefer and Ali Ünlü
Fakultät Statistik, Technische Universität Dortmund, D-44221 Dortmund, Germany
{kiefer, uenlue}@statistik.tu-dortmund.de

Abstract

Fechnerian scaling as developed by Dzhafarov and Colonius aims at imposing a metric on a set of objects based on their pairwise dissimilarities, e.g., discrimination probabilities. The objects may be perceptual stimuli or abstract categories. In this paper we apply Fechnerian scaling to a space of uni- and multidimensional logistic models used in item response theory for dichotomous data. The space of models is created by assigning to each ordered pair of models (A,B) *a discrimination probability, taken to be the probability with which model* B *fits, by some statistical criterion, a data set randomly generated by model* A *at least as well as* A *fits this data set itself. We then use (metric) multidimensional scaling to (isometrically) embed, for visualization purposes, the set of the item response theory models with pairwise Fechnerian distances in the Euclidean* $2D$ *space.*

Fechnerian scaling (FS; Dzhafarov & Colonius, 2006b, 2007) provides a theoretical framework for deriving Fechnerian distances among objects (e.g., colors or signals) from discrimination probabilities or other measures showing the degree with which objects are discriminated from each other by what is generically referred to as a perceiving system (e.g., person or technical device). As described in Dzhafarov and Colonius (2006a,b) "nonpsychophysical" interpretations of the terms "object" and "perceiving system" are possible, designating, for instance, such purely conceptual entities as a statistical model and a computational model comparison procedure, respectively. Following these authors' suggestion, in this paper we take as the object set a set of item response theory (IRT; Reckase, 2009; Van der Linden & Hambleton, 1997) psychometric models, and a computational model comparison procedure based on the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van der Linde, 2002) as the perceiving system. In the statistical model selection literature (Burnham & Anderson, 2002; Myung, Forster, & Browne, 2000) such a sort of "behavioristic" approach to comparing statistical models has not been studied so far. Therefore, the general aim of this paper is to outline the scope of FS, an originally psychophysical "metric from discriminability" theory, for the exploratory and graphical analysis of logistic IRT models. All computations in this paper have been performed in the freely available and powerful computing environment R (www.r-project.org).

Method

In order to be able to compute a Fechnerian metric on an abstract set of statistical models, what are the "discrimination probabilities" to actually begin with? Conceptually speaking, the entries of a data matrix of pairwise discrimination probabilities for a collection of statistical models are to be representing the probabilities with which a model comparison procedure (such as the DIC criterion) "perceives" an ordered pair of a data-generating row model (belonging to what is called in FS the first observation area) and a data-fitting column model (in the second observation area) as comprised of two "different" models. The Fechnerian

distances computed from such a discrimination probability matrix then can be interpreted as dissimilarities among the statistical models "from the point of view" of the computational model comparison procedure (the DIC criterion).

However, the application of FS to model comparison is not straightforward as it may seem. There is a crucial difference between the classical paradigm of model selection, which we subsequently refer to as the "best-of" paradigm, and the experimental paradigm of pairwise presentations and same-different judgments in FS, which we briefly call the "same-different" paradigm. In the best-of paradigm, stimuli (competing models) from a selection set are presented mutually, all at a time, and the perceiver (model comparison criterion) has to select a stimulus (model) that yields the "best" value "from the point of view" of the perceiver (the "best" model selection criterion, e.g., smallest DIC, value). This paradigm is different from and must not be confused with the "same-different" paradigm of FS.

Same-Different Discrimination Probability Matrix

A data matrix of same-different discrimination probabilities among competing statistical models can be derived as follows (for an example see the Results and Discussion section):

(1) Simulate a large number of data sets for each of the row models (to attain reliable estimates of the discrimination probabilities).

(2) For any simulated data set, estimate the parameters for each of the column models (e.g., by Markov chain Monte Carlo (MCMC)) and compute the corresponding selection criterion (e.g., DIC) values.

(3) For any data-generating row model and each of the data sets simulated under this model, fix the data-fitting column model representing the data-generating row model as the *baseline model*, and increment the counter for any of the other column models whenever it outperforms the baseline model in selection criterion (DIC) value computed for the simulated data set – in case of a tie (equal DIC value), flip a coin. Divide these counts by the numbers of simulated data sets and form a corresponding row-models-by-column-models matrix of relative frequencies, where the entries on the main diagonal for the (column) baseline models are set to 0.50.

(4) This matrix of relative frequencies is obtained under a "greater-less" paradigm of pairwise presentations with greater-less judgments. In "response" to every pair of stimuli (models) the perceiver (model comparison procedure) "judges" which of the two stimuli (models) is "greater" (has a smaller model selection, e.g., DIC, value). The obtained matrix of probabilities $\gamma(s,x)$ for greater-less comparisons among the row models s and column models x then can be transformed into a matrix of same-different discrimination probabilities $\psi(s,x)$ through $\psi(s,x) = \left|\gamma(s,x) - 0.50\right|$. This choice of transformation can be deemed reasonable if one compares the principle of regular mediality for greater-less judgments with the regular minimality principle for same-different judgments (see Dzhafarov, 2003; Dzhafarov & Colonius, 2006a).

It turns out that the same-different discrimination probabilities among the statistical models so obtained, in general satisfy the only property of the data required by FS, the property of regular minimality, in its canonical form (the minima, 0, on the main diagonal; see the section Fechnerian Distances).

Item Response Theory

IRT provides statistical models that link dichotomous response data to a latent trait. In an achievement tests, for example, a number of examinees with unknown abilities may respond to a set of binary items, producing a "correct" or "incorrect" response, coded by 1 or 0,

respectively. The subjects' answers $x_{ij} \in \{0,1\}$ can be modeled by what is called the item response function (IRF) of an item, which gives the probability of a correct response to the item as a function of person ability. In this paper we consider a set of (dichotomous) logistic IRT models. The IRF for the k-dimensional three-parameter logistic model, $\mathrm{M_k\,3PL}$, is $\mathrm{P}_{\mathrm{M_k\,3PL}}(x_{ij}=1 \mid \vec{\theta}_i) = \gamma_j + (1-\gamma_j)\cdot\frac{\exp(\vec{\alpha}_j'\vec{\theta}_i+\beta_j)}{1+\exp(\vec{\alpha}_j'\vec{\theta}_i+\beta_j)}$, where $\vec{\theta}_i$ and $\vec{\alpha}_j$ are the k-dimensional (column) vectors of person i's ability and item j's discrimination (with scalar product $\vec{\alpha}_j'\vec{\theta}_i$), respectively, and the scalars β_j and γ_j denote the difficulty and pseudo-guessing parameters of item j, respectively. Restricting the parameters of this model yields the following special cases: k-dimensional two-parameter ($\gamma_j = 0$) and one-parameter ($\gamma_j = 0$ and $\vec{\alpha}_j = \vec{1}$) logistic models. We refer to the unidimensional versions of these models as $\mathrm{1PL}$, $\mathrm{2PL}$, and $\mathrm{3PL}$.

Parameter Estimation and Model Fit

The parameters of the logistic IRT models can be estimated using the MCMC Gibbs sampling technique (Casella & George, 1992; Geman & Geman, 1984; Patz & Junker, 1999a,b; Robert & Casella, 2004). Values representing the model parameters are sampled repeatedly from their full conditional posterior distributions. After the "burn in" phase the generated Markov chain attains its stationary distribution. The value taken as the MCMC estimate is the mean over a large number of successive iterations sampled. The model selection criterion that comes as a by-product of the Gibbs sampling is the DIC (Spiegelhalter et al., 2002). In accordance with classical model selection indices such as AIC and BIC the DIC penalizes a good fit by a value representing the number of used parameters. A model with the smallest DIC value is selected among the competing models (cf. Kang & Cohen, 2007).

Fechnerian Distances

The Fechnerian distances among the IRT models are derived based on the theory of FS as developed by Dzhafarov and Colonius (2006b, 2007). We provide a brief and by necessity schematic overview of the main concepts of FS. The only property of the data matrix ψ of (same-different) discrimination measures required by FS is regular minimality (RM). This property states that every row and every column of the ψ-matrix contains a single minimal entry, and an entry minimal in its row is minimal in its column. Given RM is satisfied, FS imposes a metric on the set of objects as follows. Let $a \to X \to b$ denote a chain, a finite sequence, $(a = x_0, x_1, \ldots, x_k, x_{k+1} = b)$ of stimuli leading from stimulus a to stimulus b. For such a chain we can compute what is called its psychometric length (of the first kind) $L^{(1)}[a \to X \to b] = \sum_{i=0}^{k} (\psi(x_i, x_{i+1}) - \psi(x_i, x_i))$. The quantities $\psi(x_i, x_{i+1}) - \psi(x_i, x_i)$ are referred to as psychometric increments of the first kind. Among all such chains find a chain with minimal psychometric length, and take this minimal value of $L^{(1)}$ for the quasidistance $G_{ab}^{(1)}$ from a to b (referred to as the oriented Fechnerian distance of the first kind). Analogously, we can define the psychometric increments, $\psi(x_{i+1}, x_i) - \psi(x_i, x_i)$, psychometric lengths, $L^{(2)}$, and the oriented Fechnerian distances, $G_{ab}^{(2)}$, of the second kind. It can be shown that $G_{ab}^{(1)} + G_{ba}^{(1)} = G_{ab}^{(2)} + G_{ba}^{(2)} =: G_{ab}$, and this metric G_{ab} is taken for the "true" or "overall" Fechnerian distance between the stimuli a and b.

Results and Discussion

Discrimination Probabilities and Fechnerian Distances

Having simulated a data set, for example, under the 5-dimensional two-parameter logistic model, as the data-generating row model, we fix the column model $M_5$2PL as the baseline model. The baseline model fitted to the simulated data set gives a DIC value, which the DIC values of the other column models are compared to. The comparison relates to whether or not a column model yields a smaller DIC value than the baseline model (see the section Method).

We present the matrix of probabilities for greater-less comparisons among the logistic IRT models (for 50 data sets simulated per row model, 20 test items, and a sample size of 200):

	1PL	2PL	3PL	$M_2$1PL	$M_2$2PL	$M_2$3PL	$M_5$1PL	$M_5$2PL	$M_5$3PL
1PL	0.50	0.24	0.02	0.36	1	1	0.06	1	0
2PL	0	0.50	0	0	1	1	0	1	0
3PL	0.02	0.52	0.50	0	1	1	0	1	0
$M_2$1PL	0	0.02	0	0.50	1	1	0.78	1	0
$M_2$2PL	0	0	0	0	0.50	0.26	0	1	1
$M_2$3PL	0	0	0	0	0.45	0.50	0	1	1
$M_5$1PL	0	0	0	0	1	0.98	0.50	1	1
$M_5$2PL	0	0	0	0	0	0	0	0.50	0.30
$M_5$3PL	0	0	0	0	0	0	0	0.63	0.50

This probability matrix can be transformed into a matrix of same-different discrimination probabilities by the transformation described in the Method section:

	1PL	2PL	3PL	$M_2$1PL	$M_2$2PL	$M_2$3PL	$M_5$1PL	$M_5$2PL	$M_5$3PL
1PL	0	0.26	0.48	0.14	0.50	0.50	0.44	0.50	0.50
2PL	0.50	0	0.50	0.50	0.50	0.50	0.50	0.50	0.50
3PL	0.48	0.02	0	0.50	0.50	0.50	0.50	0.50	0.50
$M_2$1PL	0.50	0.48	0.50	0	0.50	0.50	0.28	0.50	0.50
$M_2$2PL	0.50	0.50	0.50	0.50	0	0.24	0.50	0.50	0.50
$M_2$3PL	0.50	0.50	0.50	0.50	0.05	0	0.50	0.50	0.50
$M_5$1PL	0.50	0.50	0.50	0.50	0.50	0.48	0	0.50	0.50
$M_5$2PL	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0	0.2
$M_5$3PL	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.13	0

This matrix satisfies RM (in the canonical form) and FS can be performed yielding the Fechnerian distances shown below:

	1PL	2PL	3PL	$M_2$1PL	$M_2$2PL	$M_2$3PL	$M_5$1PL	$M_5$2PL	$M_5$3PL
1PL	0	0.76	0.96	0.64	1	1	0.92	1	1
2PL	0.76	0	0.52	0.98	1	1	1	1	1
3PL	0.96	0.52	0	1	1	1	1	1	1
$M_2$1PL	0.64	0.98	1	0	1	1	0.78	1	1
$M_2$2PL	1	1	1	1	0	0.29	1	1	1
$M_2$3PL	1	1	1	1	0.29	0	0.98	1	1
$M_5$1PL	0.90	1	1	0.78	1	0.98	0	1	1
$M_5$2PL	1	1	1	1	1	1	1	0	0.33
$M_5$3PL	1	1	1	1	1	1	1	0.33	0

Graphical Representation

Classical multidimensional scaling (MDS; Borg & Groenen, 2005) can serve as a reference against which to consider FS (Dzhafarov & Colonius, 2006b). In the present paper we apply metric MDS to the computed Fechnerian distances to isometrically embed, for visualization, the space of the logistic IRT models in two-dimensional Euclidean space (Figure 1, left plot). Moreover, metric MDS on the Fechnerian distances can be compared with nonmetric MDS performed on the computed same-different discrimination probabilities (Figure 1, right plot).

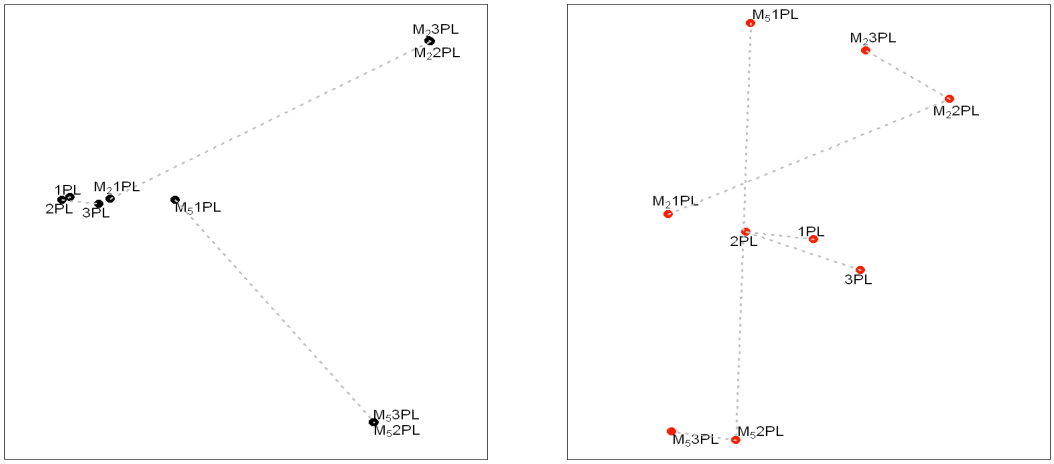

Fig. 1. Two-dimensional Euclidean representations of the logistic IRT models for the Fechnerian distances (metric MDS, left plot) and for the same-different discrimination probabilities (nonmetric MDS, right plot). "Special case" relation for nested models of same dimensionality is displayed using dashed lines.

As can be seen from the left plot of Figure 1, the models M_k2PL and M_k3PL are very close to each other. In the right plot, the two-parameter logistic models are more separated from the three-parameter logistic models of corresponding same dimensionality. In contrast, the multidimensional models M_k1PL are located closer to the unidimensional models than to the more complex models M_k2PL and M_k3PL in the left plot. In the right plot, instead, there is the "outlier" M_51PL, which is located closer to the two-dimensional two- and three-parameter logistic models than to any of the other models. In both plots, the graphical representations reveal the hierarchy in (difference of) model complexity among the one-parameter logistic models; in terms of complexity, 1PL ought to be "closer" to M_21PL than to M_51PL. This is also reflected by the fact that a geodesic (of minimal psychometric length) chain in the set of all nine IRT models connecting object 1PL to object M_51PL is given by the sequence of models $(1PL, M_21PL, M_51PL)$. Altogether, in this application FS, as a method of preprocessing the dissimilarity data for the logistic IRT models, seems to stabilize and improve on the MDS results (for a discussion of FS as a data-analytic tool see Dzhafarov, 2010).

Applying the FS procedure to evaluate dissimilarities among statistical models is an interesting new approach that proves valuable in being further pursued and investigated. In-depth simulation studies have to be conducted, for instance including broader or different classes of statistical models and other model comparison criteria. In combination with MDS, moreover, the presented approach can allow for interesting applications of interactive graphical methods to be used for the exploratory analysis of multivariate statistical models.

Acknowledgements

We are deeply indebted to Prof. Dr. Ehtibar Dzhafarov for calling attention to this topic and for his helpful comments on a first draft of the manuscript.

References

Borg, I., & Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Berlin: Springer.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer.

Casella, G. & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician, 46*, 167-174.

Dzhafarov, E. N. (2003). Thurstonian-type representations for "same-different" discriminations: Deterministic decisions and independent images. *Journal of Mathematical Psychology, 47*, 184-204.

Dzhafarov, E. N. (2010). Dissimilarity cumulation as a procedure correcting for violations of triangle inequality. *Journal of Mathematical Psychology, 54*, 284-287.

Dzhafarov, E. N., & Colonius, H. (2006a). Regular minimality: A fundamental law of discrimination. In H. Colonius & E. N. Dzhafarov (Eds.), *Measurement and Representation of Sensations* (pp. 1-46). Mahwah, NJ: Erlbaum.

Dzhafarov, E. N., & Colonius, H. (2006b). Reconstructing distances among objects from their discriminability. *Psychometrika, 71*, 365-386.

Dzhafarov, E. N., & Colonius, H. (2007). Dissimilarity cumulation theory and subjective metrics. *Journal of Mathematical Psychology, 51*, 290-304.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 6*, 721-741.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*, 331-358.

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Model selection [Special issue]. *Journal of Mathematical Psychology, 44*, 1-2.

Patz, R. J., Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146-178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342-366.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York: Springer.

Robert, C. P., & Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*, 583-639.

Van der Linden, W. K., & Hambleton, R. (1997). *Handbook of Modern Item Response Theory*. New York: Springer.