Journal of Mathematical Psychology 67 (2015) 8–25

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

The flexibility of models of recognition memory: The case of confidence ratings

Karl Christoph Klauer[a,*,1], David Kellen[b,1]

[a] Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, D-79085 Freiburg, Germany

[b] Center for Cognitive and Decision Sciences, Faculty of Psychology, University of Basel, Switzerland

HIGHLIGHTS

- A method for computing NML for models with categorical data is presented.
- NML penalties are given for confidence-rating based models of recognition memory.
- In simulation studies, NML outperforms AIC and BIC in model recovery.
- A meta-analysis based on NML supports dual-process signal-detection models.

ARTICLE INFO

Article history:
Received 17 December 2014
Received in revised form
7 May 2015
Available online 10 June 2015

Keywords:
Model selection
Minimum-description length
Normalized maximum likelihood
Recognition memory
Confidence ratings

ABSTRACT

The normalized maximum likelihood (NML) index is a model-selection index derived from the minimum-description length principle. In contrast to traditional model-selection indices, it also quantifies differences in flexibility between models related to their functional form. We present a new method for computing the NML index for models of categorical data that parameterize multinomial or product-multinomial distributions and apply it to comparing the flexibility of major models of recognition memory for confidence-rating based receiver-operating-characteristic (ROC) data. NML penalties are tabulated for datasets of typical sizes and interpolation functions are fitted that allow one to interpolate NML penalties for datasets with sizes between the tabulated ones. Recovery studies suggest that the NML index performs better than traditional model-selection indices in model selection from ROC data. In an NML-based meta-analysis of 850 ROC datasets, versions of the dual-process signal detection models received most support followed by the finite mixture signal detection model and constrained versions of two-high threshold models.

Recognition memory has frequently been studied by means of mathematical models (for a review, see Malmberg, 2008 and Yonelinas & Parks, 2007). A range of models has been proposed. In some models, information from memory is represented in terms of discrete states; in others, a continuous representation of evidence is postulated. Discrete-state models are variants of the so-called threshold models (e.g., Blackwell, 1963 and Snodgrass & Corwin, 1988); the continuous models are variants of the so-called signal-detection models (Macmillan & Creelman, 2005). Finally, hybrid models implement combinations of both ideas. Model fits and comparisons are frequently based on the shape of the observed receiver operating characteristic (ROC) functions.

1. Receiver operating characteristics

In the most basic recognition experiment, participants study items to be remembered later. In a subsequent test phase, they are shown the studied items mixed with new items, so-called distractors, and their task is to discriminate studied items from new items. These data are typically modeled in terms of two probabilities, the probability to respond OLD given an old item and the probability to respond OLD given a new item, also called the hit and false-alarm rates, respectively. An important line of research is to obtain hit and false-alarm rates at different levels of response bias and to plot hit rates against false alarm rates across levels of response bias resulting in a so-called ROC function. Fig. 1 shows examples of typical ROCs: From top to bottom, they are typical of an ROC generated by a threshold model, a signal-detection model with higher variability of the memory response for old items than for new items, and a simpler signal detection model with equal

* Corresponding author.
E-mail address: christoph.klauer@psychologie.uni-freiburg.de (K.C. Klauer).

1 Both authors contributed equally to this paper.

http://dx.doi.org/10.1016/j.jmp.2015.05.002

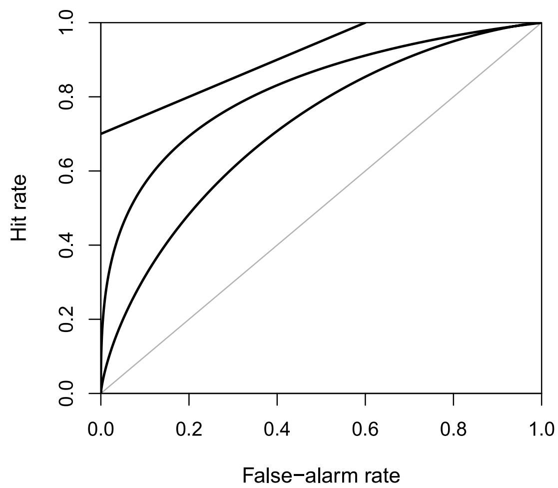

Fig. 1. Example ROC functions.

variability for old and new items. All of these fall above the chance-level diagonal that is also shown.

Different levels of response bias are traditionally produced via manipulations of base rates of old relative to new items in the test phase or by payoff manipulations (Bröder & Schütz, 2009). The vast majority of studies in the field has, however, relied on a less expensive method of generating ROC data via confidence ratings (see Wixted, 2007 and Yonelinas & Parks, 2007, for reviews). That is, ratings of confidence in the OLD or NEW response, as the case may be, are collected and the different levels of confidence interpreted as expressing different levels of response bias. Although most widely used in research on recognition memory, confidence-rating data and (some of) the above models also play an important role in perception (e.g., Swets, Tanner, & Birdsall, 1961) and reasoning (e.g., Dube, Rotello, & Heit, 2010).

Despite decades of research, the question which of the above models provide the best description of the data in recognition memory is still under debate (e.g., Bröder & Schütz, 2009; Dube & Rotello, 2012; Kellen & Klauer, 2014; Kellen, Klauer, & Bröder, 2013; Kellen, Singmann, Vogt, & Klauer, 2015; Onyper, Zhang, & Howard, 2010; Province & Rouder, 2012; Wixted, 2007 and Yonelinas & Parks, 2007). To our minds, several factors have prevented this very productive field from reaching a clear and non-contested decision on the most adequate model. Among these are the distorting, but often ignored influences of individual differences in memory performance and response-bias settings and of analogous differences between items (Klauer & Kellen, 2010; Rouder & Lu, 2005), an over-reliance on one method, the confidence-rating paradigm, and the absence of model-selection measures that take into account differences between models in flexibility related to functional form. The purpose of the present manuscript is to address this last problem for the important case of confidence-rating data in a similar fashion as described by Kellen et al. (2013) and Klauer and Kellen (2011a) for binary OLD/NEW ROC data (see also Kellen & Klauer, 2011). It turns out that some of the solutions developed here are even more widely applicable than defined by this original purpose as elaborated on below.

2. Model selection

Given data and a range of models, the task to select the model that most parsimoniously accounts for the data is discussed under the heading "model selection". There is a growing awareness in psychology that goodness of fit and model flexibility should both be weighed when evaluating mathematical models. For example, two special issues of the Journal of Mathematical Psychology (Myung, Forster, & Brown, 2000; Wagenmakers & Waldorf, 2006) were recently devoted to this topic.

In recognition memory, model selection has usually relied on the Akaike information criterion (AIC) and the Bayesian information criterion (BIC; see, e.g., Burnham & Anderson, 2005). Let f be the model's probability function, $\boldsymbol{x}$ the observed data, and $\hat{\theta}(\boldsymbol{x})$ the maximum-likelihood estimate of the p model parameters, $\theta = (\theta_1, \ldots, \theta_p)$. AIC and BIC are given by

$$\text{AIC} = -2\log f(\boldsymbol{x} \mid \hat{\theta}(\boldsymbol{x})) + 2p, \quad \text{and}$$

$$\text{BIC} = -2\log f(\boldsymbol{x} \mid \hat{\theta}(\boldsymbol{x})) + p * \log(N),$$

where N is the number of data points. In both indices, the first term quantifies the model's goodness of fit (minus twice the maximum log-likelihood), and the second term is the flexibility penalty. AIC and BIC thus gauge model flexibility basically in terms of the number of parameters. This is both crude and not very helpful in the present context. It is crude because an extra parameter can have anything between very little effect on the model's flexibility (i.e., on its capability to fit diverse datasets) and a tremendous effect depending on the functional form via which it enters the model equations. It is also not very helpful given that many of the major models employ the same number of parameters in fitting ROC data (plus or minus one or two). This problem has repeatedly been noted in the literature, without definitive solution so far (Macho, 2004; Onyper et al., 2010 and Wixted, 2007; see also Cohen, Rotello, & Macmillan, 2008).

The purpose of the present project is to bring functional form into consideration using recent developments in the model-selection field based on the minimum-description-length (MDL) principle (Myung, Navarro, & Pitt, 2006). Roughly, a model reduces the complexity of a code (e.g., a string of binary values) needed to describe the data, because only the model, the estimated parameter values as well as the residuals have to be encoded, once the model has been fit. Inasmuch as the residuals show less variability than the original data, a good model thereby reduces the code needed to describe the dataset. The code length is a function of both the model's complexity and its ability to account for the data (Grünwald, 2007).

Model selection based on the minimum-description-length principle has a strong track record in psychology. These modern methods have to date been applied to the class of multinomial processing tree models (Wu, Myung, & Batchelder, 2010a,b), to models of human categorization (Myung, Pitt, & Navarro, 2007), to clustering models (Navarro & Lee, 2005), to models of recognition memory (Kellen & Klauer, 2011; Kellen et al., 2013; Klauer & Kellen, 2011a), to decision making (Davis-Stober & Brown, 2011; Moshagen & Hilbig, 2014), and to structural equation models (Preacher, 2006), among others.

In the present context, the principle leads to the normalized maximum likelihood index (NML) for model selection. Much like AIC and BIC, the index adds minus the maximum log-likelihood of the data given the model and a penalty term for the model's flexibility. The penalty is given by the logarithm of the sum of maximum likelihood values summed over the entire set of possible data patterns $\boldsymbol{y}$ that might in principle occur in the experimental setting. NML[2] is given by

$$\text{NML} = -\log f(\boldsymbol{x} \mid \hat{\theta}(\boldsymbol{x})) + \log \sum_{\boldsymbol{y}} f(\boldsymbol{y} \mid \hat{\theta}(\boldsymbol{y})).$$

[2] Strictly speaking, this expression gives the logarithm of NML. NML cannot be computed for the case of continuous data as the penalty term is not finite for continuous data (e.g., Karabatsos & Walker, 2006). This problem can be sidestepped through the introduction of a (informative) prior distribution for the data in the computation of the penalty (see Zhang, 2011).