Journal of Mathematical Psychology 55 (2011) 430–450

Contents lists available at SciVerse ScienceDirect

Journal of Mathematical Psychology

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

The flexibility of models of recognition memory: An analysis by the minimum-description length principle

Karl Christoph Klauer*, David Kellen

Institut für Psychologie, Albert–Ludwigs-Universität Freiburg, Germany

ARTICLE INFO

Article history:
Received 27 June 2011
Received in revised form
5 September 2011
Available online 19 October 2011

Keywords:
Minimum description length
Normalized maximum likelihood
Fisher information approximation
Signal detection theory
Recognition memory

ABSTRACT

Ten continuous, discrete, and hybrid models of recognition memory are considered in the traditional paradigm with manipulation of response bias via baserates or payoff schedules. We present an efficient method for computing the Fisher information approximation (FIA) to the normalized maximum likelihood index (NML) for these models, and a relatively efficient method for computing NML itself. This leads to a comparative evaluation of the complexity of the different models from the minimum-description-length perspective. Furthermore, we evaluate the goodness of the approximation of FIA to NML. Finally, model-recovery studies reveal that use of the minimum-description-length principle consistently identifies the true model more frequently than AIC and BIC. These results should be useful for research in recognition memory, but also in other fields (such as perception, reasoning, working memory, and so forth) in which these models play a role.

The testing of quantitative models is one of the most important aspects of scientific inquiry. The goal is to find the most parsimonious model accounting for psychological phenomena (Jang, Wixted, & Huber, 2009). Goodness-of-fit measures are routinely used to assess the adequacy of a model for describing the data, and the model with the best fit is often preferred. There is, however, growing consensus that model fit has to be weighed against model flexibility to attain the goal of selecting the most parsimonious model accounting for the data (e.g., Jang et al. (2009), Myung, Navarro, and Pitt (2006), Roberts and Pashler (2000), Wagenmakers, Ratcliff, Gomez, and Iverson (2004)). In terms of model fit and model flexibility, the goal translates into selecting the model with an optimal trade-off of fit and (lack of) flexibility.

In the present paper, we consider a number of prominent discrete-state, continuous, and hybrid models of recognition memory. The discrete-state models are variants of the so-called threshold models (e.g., Blackwell (1963), Snodgrass and Corwin (1988)), the continuous models are variants of the so-called signal-detection models (e.g., Macmillan and Creelman (2005)), hybrid models combine elements of both classes of models. Subsets of these models have not only been extensively applied in recognition memory, but also in perception, reasoning, working memory, and a host of other fields (Swets, 1986).

In their basic form, the models deal with data from binary decisions on stimuli sampled from two classes of stimuli. They characterize the observed decisions in terms of (a) the respondent's ability to discriminate between the two stimulus classes through the use of different processes, and (b) his or her response biases. One of the two stimulus classes typically has a special role. For example, in recognition memory, one class of stimuli consists of old items that were previously seen; the other class consists of lures that were not previously presented. In perception, one stimulus class frequently consists of signals that are to be detected in the presence of noise, the other class consists of noise stimuli without signal. In reasoning, one stimulus class consists of valid logical forms, the other class consists of invalid forms, and so forth. But there are many other formats such as the two-alternative-forced choice task (Wickens, 2002, chap. 6) that have been addressed by these models. In what follows, we will adopt the terminology of signal and noise stimuli associated with YES and NO responses, respectively, but will keep in mind that stimulus classes and responses will carry appropriately different labels in other fields.

One way to view these models is as measurement models, assessing changes in response bias and in sensitivity in response to experimental manipulations and individual differences (Batchelder & Riefer, 1999; Malmberg, 2008). Of course, for the measurement to be valid, the model assumptions have to be met, and the major cognitive processes involved in generating the data must be adequately incorporated into the model (Batchelder & Riefer, 1999). The different models thereby have different affinities with different classes of underlying process models. Process or computational models describe in a more fine-grained manner the varieties, architecture, and interactions of cognitive processes assumed to produce the observable decisions. They often generate predictions about the effects of experimental manipulations,

* Correspondence to: Institut für Psychologie, Albert–Ludwigs-Universität Freiburg, D-79085 Freiburg, Germany.
E-mail address: christoph.klauer@psychologie.uni-freiburg.de (K.C. Klauer).

0022-2496/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2011.09.002

but do not provide the statistical machinery for model fitting and selection that is available at the measurement level. For example, in recognition memory, signal-detection models should often describe performance well if a global-matching model such as MINERVA2 (Hintzman, 1988) holds at the process level (Brandt, 2007). In this way, the measurement level and the process level are interdependent, and results obtained at the measurement level may imply constraints for theory-building at the process level (see also Malmberg (2008)).

Despite decades of research, the question which of the measurement models provides the best description of the data is still under debate (e.g., Bröder and Schütz (2009, 2011), Dube and Rotello (in press), Dube, Rotello, and Heit (2011), Klauer and Kellen (2011)). One issue that may have contributed to this state of affairs is that model evaluation in this literature has frequently relied on goodness of fit and simple model-selection indices such as AIC and BIC which do not adequately take the relative flexibility of the different models into account as elaborated below.

1. Receiver operating characteristics

The probabilities to respond YES given a signal stimulus and a noise stimulus, respectively, are called the hit and false-alarm rates. One traditional way to contrast the different models considered here is by means of the so-called receiver operating characteristics (ROCs). ROCs plot observed hit rate against false alarm rate across several levels of response bias. Different levels of response bias are traditionally produced via manipulations of baserates of signal trials relative to noise trials or of payoff schedules (Macmillan & Creelman, 2005; McNicol, 1972; Wickens, 2002). Another frequently employed method to generate ROC data is to obtain ratings of the confidence in the YES or NO response, as the case may be, and to interpret the different levels of confidence as emulating different levels of response bias (e.g., Swets (1986)).

Both methods, manipulations of payoff and baserates on the one hand and confidence ratings on the other hands, have different advantages and drawbacks. For example, obtaining ROCs via confidence ratings is much less costly than obtaining them via payoff or baserate manipulations. On the other hand, the models considered here make clearly different predictions for the shape of ROCs based on payoff and baserate manipulations, whereas all models can generate similarly shaped ROCs for rating data when extended to that response format appropriately (e.g., Erdfelder and Buchner (1998), Klauer and Kellen (2010, 2011), Malmberg (2002)). The present paper focuses on data obtained via payoff and baserate manipulations.

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". As already mentioned, there is growing awareness in psychology that goodness of fit and model flexibility should both be weighed in evaluating mathematical models. For example, two special issues of the Journal of Mathematical Psychology (Myung, Balasubramanian, & Pitt, 2000; Wagenmakers & Waldorf, 2006) were recently devoted to this topic.

Model selection is an active research field with many facets and methods (Linhart & Zuccini, 1986). Some of the principled techniques in use are cross-validation methods (e.g., Linhart and Zuccini (1986)), bootstrap simulations to assess model mimicry (e.g., Wagenmakers et al. (2004)), and the use of model-selection indices such as AIC and BIC (e.g., Burnham and Anderson (2005), Liu and Aitkin (2008)). Model-selection indices are used most frequently in applications of the present models. They sum terms quantifying badness of fit, usually minus (two times) the maximum of the model's logarithmized likelihood function, and a penalty for a model's flexibility. Importantly, AIC and BIC quantify flexibility in terms of numbers of parameters and (in the case of BIC) sample size, whereas functional form (i.e., the way in which the parameters are used in the model equations) plays no role. This is both crude and not very helpful in the present context. It is crude, because an additional parameter can have anything between a negligible and a dramatic effect on the model's capability to fit data in general. It is not very helpful because the models considered here often employ the same number of parameters, implying the same flexibility penalty via AIC and BIC.

The purpose of the present study is to bring functional form into consideration using recent developments in the model-selection field based on the minimum-description-length principle (Myung et al., 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 and its estimated parameters as well as the residuals have to be encoded, but not the original data, once the model has been fitted. Inasmuch as the residuals show less variability than the original data, a good model reduces the code needed to describe the data set. The code length is a function of both the model's complexity and its ability to account for the data (Grünwald, 2007). The model with the shortest code is the one striking the optimal balance between fit and parsimony according to this principle.

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 in the so-called 4AFC-2R paradigm (Kellen & Klauer, 2011), 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 y that might in principle occur in the experimental setting. Let f be the model's probability function, x the observed data, and $\hat{\theta}(x)$ the maximum-likelihood estimate of the p model parameters, $\theta = (\theta_1, \ldots, \theta_p)$. NML is given by[1]

$$\mathrm{NML} = -\log f(x|\hat{\theta}(x)) + \log \sum_y f(y|\hat{\theta}(y)).$$

The penalty term, $\log \sum_y f(y|\hat{\theta}(y))$, is a measure of the model's ability to fit data in general, whatever the data. The resulting model-selection index is principled and has a couple of desirable properties (Grünwald, 2007 chap. 7; Myung, Forster, & Brown, 2000; Rissanen, 1996, 2001) such as, in the present context, consistency: Roughly, if one of the models is the correct one, use of NML will select it as sample size increases. Importantly, NML behaves as one would intuitively expect of an index hat takes functional form into account (see e.g., Kellen and Klauer (2011), Su, Myung, and Pitt (2005), Wu et al. (2010a,b)).

This becomes clearer when considering an asymptotic approximation to NML, the Fisher information approximation (FIA). Like NML, it sum minus the maximum of the logarithmized likelihood

[1] Strictly speaking, this expression gives the logarithm of NML, but the logarithm is frequently referred to as the NML index, and we follow this convention.

of the data given the model plus a penalty term FIA_P:

$$\mathrm{FIA} = -\log f(x|\hat{\theta}) + \mathrm{FIA}_P.$$

FIA_P is the sum of two terms:

$$\mathrm{FIA}_P = \frac{p}{2}\log\frac{n}{2\pi} + \mathrm{FIA}_f,$$

where n is the sample size of the data. As can be seen, the first term in FIA_P takes the number of parameters and sample size into account in a fashion almost identical to that involved in BIC. The second term is given by:

$$\mathrm{FIA}_f = \log\int\sqrt{\det I(\theta)}d\theta,$$

where I is the so-called Fisher information matrix of the model for a sample of size one. The Fisher information matrix is the matrix of the expected second partial derivatives of the log-likelihood function. FIA_f is independent of the parameterization of the model and can be seen as a measure of the model's flexibility due to its functional form.

For example, if inequality restrictions are imposed on a model's parameter, then the thus restricted model still has the same number of parameters as the original model, but it is obviously less flexible than the original model. This is reflected both in NML and, via FIA_f, in FIA. It leads to a smaller penalty FIA_f, because the integral over $\sqrt{\det I}$ is now computed over only a subset of the parameter space, namely over those parameter values which satisfy the inequality restrictions, implying a smaller value of the integral. Neither AIC nor BIC would correct for such a change in model flexibility. Many other examples attest to the fact that NML and FIA provide not only principled model-selection indices (Grünwald, 2007; Myung et al., 2000; Rissanen, 1996, 2001), but also reflect differences in different models' flexibility in an intuitively plausible manner (Wu et al., 2010a,b). In simulation studies, use of the model-selection index based on FIA led to more valid results than the use of only goodness-of-fit values, or AIC (Myung et al., 2007), or BIC (Su et al., 2005).

Although more sophisticated than AIC and BIC, more widespread use of FIA and NML has been hampered by the difficulty of computing the indices. Computing NML involves processing all data patterns that can in principle occur in an experiment, which implies prohibitively many computations even for modern computers. FIA_f involves an integration over the p-dimensional parameter space. It is, however, possible to compute FIA_f and NML using Monte Carlo integration for the models considered here. Wu et al. (2010a,b) proposed and developed a method to compute FIA_f for discrete-state models, and we developed similar methods for the models considered here, including the continuous and hybrid models, and the computation of NML.

3. Models

The models considered are shown in Fig. 1. Discrete-state models are the One-High-Threshold model (1HTM) and the Two-High-Threshold model (2HTM). Continuous models are the Equal-Variance-Signal-Detection model (EVSDT) and the Unequal-Variance-Signal-Detection model (UVSDT). Hybrid models are the Dual-Process model (DPSDT), and the Finite-Mixture-Signal-Detection model (MSDT). Additional models arise by imposing a priori inequality restrictions on the parameters as elaborated in the next section.

In the 1HTM (Blackwell, 1963), it is assumed that decisions are based on two discrete states, "detect" or "guess". Given a signal, the item can either be detected with probability D, leading to a YES response, or not, in which case YES is guessed with probability g and NO with probability $1-g$. Response bias is captured by parameter g.

Let $p_{s,i}$ and $p_{n,i}$ be the probabilities of hits and false alarms, respectively, in (baserate or payoff) condition i, $i = 1,\ldots,K$. The 1HTM has parameters $\boldsymbol{\theta} = (g_1, g_2, \ldots, g_K, D)$ and is defined by

$$p_{s,i} = D + (1-D)g_i,$$
$$p_{n,i} = g_i.$$

In the 2HTM (Snodgrass & Corwin, 1988), subjective impressions can exceed a high threshold, that is never exceeded in a noise trial, with probability D_s, leading to a YES response, or they can fall below a low threshold, which is never the case for signal trials, with probability D_n, leading to a NO response. Impressions falling below these two thresholds lead to guessing. A restricted version with $D_n = D_s$ is also often used.

The 2HTM has parameters $\boldsymbol{\theta} = (g_1, g_2, \ldots, g_K, D_s, D_n)$ and is defined by

$$p_{s,i} = D_s + (1-D_s)g_i,$$
$$p_{n,i} = (1-D_n)g_i.$$

In the signal-detection models (Macmillan & Creelman, 2005), it is assumed that each stimulus evokes a value on a strength-of-evidence dimension which is compared to a response criterion c. Values above c give rise to the response YES, values below c to the response NO. Noise and signal stimuli generate distributions on the dimension assumed to be normal with means separated by μ. In the EVSDT, both distributions are assumed to have equal variances, which, without loss of generality, can be set equal to one, whereas the mean of the distribution of noise stimuli can be set to zero. Response bias is captured by parameter c.

Let F be the cumulative distribution function of the standard normal distribution. The EVSDT has parameters $\boldsymbol{\theta} = (c_1, c_2, \ldots, c_K, \mu)$. It is defined by

$$p_{s,i} = F(\mu - c_i),$$
$$p_{n,i} = F(-c_i).$$

In the UVSDT, the variance of the distribution of signal values on the strength-of-evidence dimension, σ_s^2, is allowed to differ from that induced by noise stimuli, σ_n^2. The UVSDT therefore uses a parameter $\sigma = \sigma_s$ for the standard deviation of the distribution induced by signal stimuli, whereas σ_n can be set equal to one as before.

Hence, the UVSDT has parameters $\boldsymbol{\theta} = (c_1, c_2, \ldots, c_K, \mu, \sigma)$ and is defined by

$$p_{s,i} = F\left(\frac{\mu - c_i}{\sigma}\right),$$
$$p_{n,i} = F(-c_i).$$

The DPSDT (Yonelinas, 1997) combines high-threshold and signal-detection assumptions. It is assumed that a certain proportion R of signal stimuli are detected, leading to a YES response. If detection fails, decisions are governed by the EVSDT. The DPSDT has parameters $\boldsymbol{\theta} = (c_1, c_2, \ldots, c_K, \mu, R)$ and is defined by

$$p_{s,i} = R + (1-R)F(\mu - c_i)$$
$$p_{n,i} = F(-c_i).$$

In the MSDT (DeCarlo, 2002), the distribution of signal stimuli is assumed to be a mixture of two equal-variance normal distributions — one corresponding to items that were attended to with mean μ, the other one to signals that were processed during a lapse of attention with mean μ^*, with $\mu \geq \mu^* \geq 0$ and with μ^* often set equal to zero. The proportion of attended items is described by the mixture coefficient λ. Note that the MSDT

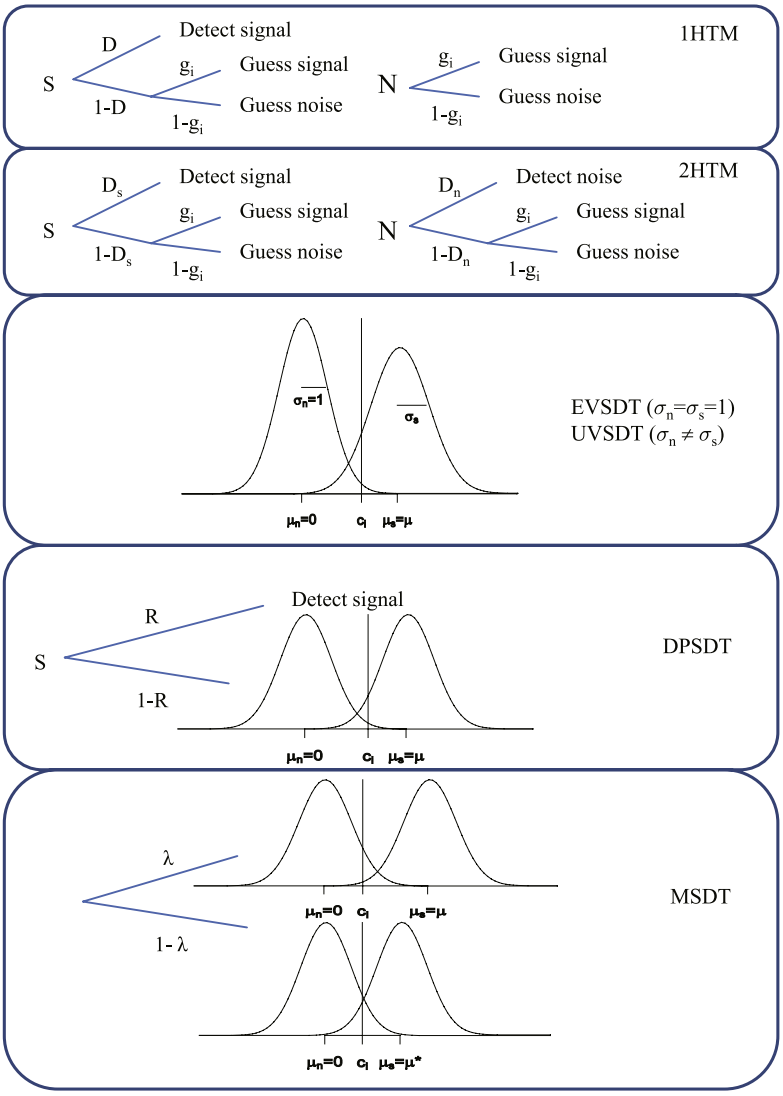

Fig. 1. The discrete-state, continuous, and hybrid models.

is mathematically equivalent to the Variable-Recollection Dual-Process model (VRDP) by Onyper, Zhang, and Howard (2010) in the present context.

The MSDT has parameters $\boldsymbol{\theta} = (c_1, c_2, \ldots, c_K, \mu^*, \mu, \lambda)$, $\mu \geq \mu^* \geq 0$. It is defined by:

$$p_{s,i} = \lambda F(\mu - c_i) + (1 - \lambda)F(\mu^* - c_i)$$
$$p_{n,i} = F(-c_i).$$

In fitting these models to data from experiments with several payoff or baserate conditions, an invariance assumption is made. It is assumed that the manipulation affects only the response-bias parameters. Thus, given K such conditions, there are K different parameters g_i, $i = 1, \ldots, K$, in the discrete-state models and K parameters c_i, $i = 1, \ldots, K$, for the hybrid and continuous models. The remaining, sensitivity-related parameters are assumed to be invariant under payoff or baserate condition and thus, the same parameter values are used for each condition. Some amount of empirical testing of this invariance assumption is implied by the models' goodness-of-fit test via G^2. For example, because $K = 2$ conditions suffice to estimate the sensitivity-related parameters μ and σ of the UVSDT, the goodness-of-fit test can be interpreted as testing whether performance in additional conditions is consistent with the level of performance implied by μ and σ as pinned down by any subset of two conditions among the implemented conditions. Poor model fit can therefore reflect a violation of the invariance assumption. Alternatively, it may mean that the invariance assumption holds, but not for the performance metric provided by UVSDT, so that another metric might be more appropriate (e.g., the one provided by DPSDT). Yet another possibility is, however, that the invariance assumption is truly violated even if another model provides a satisfactory fit. For example, Atkinson (1963) demonstrated that a discrete-state model such as the 2HTM can generate ROCs similar to that of a continuous model such as the UVSDT when principled sensitivity differences are assumed for the baserate or payoff conditions. Poor model fit irrespective of which model is fitted suggests that the invariance assumption is substantially violated for the given data set whatever metric is applied.

EVSDT and 1HTM often perform poorly in fitting ROC data, whereas the other models perform better. We nevertheless include them in the present analyses, because the lack of fit may be offset by greater parsimony.

4. Introducing inequality restrictions and model refinements

One of the advantages of the use of the minimum-description-length framework is that a priori inequality restrictions can be meaningfully taken into account. For example, the flexibility

of the continuous and hybrid models is constrained by the fact that parameters μ (EVSDT, UVSDT, DPSDT, MSDT0, MSDT) and μ^* (MSDT) are theoretically expected to be larger or at best equal to zero. The constraint implies that performance in detecting signals should be bounded from below by chance so that performance should not be reliably below chance.[2] This restriction is accepted a priori by most researchers. Imposing such inequality restrictions (i.e., $\mu \geq 0$, $\mu \geq \mu^* \geq 0$) leads to a bonus in computing FIA_f for these models. This in turn leads to a fairer comparison with the discrete-state models given that these cannot produce below-chance performance and therefore have analogous restrictions implicitly built into their model structure. Of course, if performance is empirically below chance, the bonus in FIA_f may be offset by a drop in goodness of fit, which would punish all models considered here (but see footnote 2). For these reasons, all our analyses include the above inequality restrictions.

Many studies in recognition memory and perception have observed asymmetric ROCs (e.g., Glanzer, Kim, Adams, and Hilford (1999), Ratcliff, Sheu, and Gronlund (1992), Swets, Tanner, Jr, and Birdsall (1961)). To account for this asymmetry, the UVSDT parameter σ_s is often assumed to exceed σ_n (e.g., Swets et al. (1961)), and we therefore considered two variants of the UVSDT model, one without inequality constraint for σ_s and σ_n and one with the constraint that $\sigma_s \geq \sigma_n$, termed UVSDT($\sigma_s \geq \sigma_n$).[3] Again, imposing the constraint implies a benefit in the flexibility term in model selection. The asymmetry in observed ROCs also partly motivated the development of the DPSDT and the MSDT (Jang et al., 2009). In terms of the 2HTM, the asymmetry is captured by a model with the constraint that $D_s \geq D_n$ (see footnote 3), and like for the UVSDT, we accordingly considered two versions of the 2HTM, one without and one with the constraint, the latter was termed 2HTM($D_s \geq D_n$). Imposing these restrictions for UVSDT and 2HTM might be argued to provide a fairer comparison with the hybrid models that have the asymmetry already implicitly built into model structure. Furthermore, a version of the 2HTM with $D_n = D_s$ is often considered, corresponding to the EVSDT assumption that $\sigma_n = \sigma_s$ (see footnote 3), hence a model with $D_n = D_s$ was also considered and termed 2HTM($D_s = D_n$). For the MSDT, we also consider two versions, one with the frequent constraint that $\mu^* = 0$ and one without the constraint. The restricted model was termed MSDT0.

5. Computing FIA and NML via Monte Carlo integration

FIA. Integrating $f = \sqrt{\det I}$ over the parameter space was done by independent importance sampling (Evans & Swartz, 2000, chap. 6). For this purpose, points $\boldsymbol{\theta}_i, i = 1, \ldots, m$, are sampled from a density g defined on the parameter space with $g > 0$ on the parameter space, and the integral $T = \int f$ is approximated by $T_m = \frac{1}{m} \sum_i \frac{f(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}$.

One condition that is sufficient for T_m to converge to the integral in question as m becomes large is that g dominates f, that is that there is a value $M > 0$ so that $f \leq Mg$ for all $\boldsymbol{\theta}$ in the parameter space. In addition, the rate of convergence will largely depend upon the similarity of f and g. Ideally, $\frac{f}{g}$ should be a constant.

Furthermore, if s_m is the standard deviation of the sampled $\frac{f(\boldsymbol{\theta}_i)}{g(\boldsymbol{\theta}_i)}$, then an asymptotically valid $(1 - \alpha)$ confidence interval for T is given by

$$\left(T_m - z_{(1-\frac{\alpha}{2})} \frac{s_m}{\sqrt{m}}, T_m + z_{(1-\frac{\alpha}{2})} \frac{s_m}{\sqrt{m}}\right).$$

This means that for large m, the true value T is contained in the confidence interval with half-lengths $\pm 3 \frac{s_m}{\sqrt{m}}$ with virtual certainty. Note that if g dominates f, s_m will be bounded from above for all m.

Implementing this idea requires, for each model, a numerically stable means to evaluate the determinant of the Fisher information matrix even for extreme parameter values and a suitable proposal density g from which parameter values can be efficiently sampled. We derived simple expressions for the determinants involved that can be evaluated in a numerically stable manner and from which suitable densities g were found as explained in Appendix A.

Four independent streams of Monte Carlo integration via importance sampling were implemented to estimate T. Sampling was continued until given predetermined accuracy criteria were satisfied. Because we need to estimate $\log T$ for the FIA complexity measure, these are formulated in terms of $\log T_m$ (rather than in terms of T_m). The first criterion was that the maximum absolute difference between the streamwise estimates of $\log T$, $\log T_m$, was to be smaller than 0.001. That is, the four independent streams were required to converge on a common estimate.[4]

The second criterion made use of the above confidence interval for each stream. Specifically, we required that for each stream the length of the interval $\left(\log[T_m - 3\frac{s_m}{\sqrt{m}}], \log[T_m + 3\frac{s_m}{\sqrt{m}}]\right)$ was to be smaller than 0.001.

Sampling stopped if both criteria were satisfied simultaneously. The final estimate of T was the grand average $\bar{T}_m$ over the four streams of the streamwise estimates T_m. The resulting estimate of $\log T$ was $\log \bar{T}_m$. It can be expected to be accurate to at least three decimal places.

FORTRAN code of an implementation of the algorithm, calling the IMSL and NAG libraries for numerical computations as well as a routine for computing the logarithm of the distribution function of the normal distribution (Linhart, 2008), can be obtained from the first author.

NML. Similar principles can be used to compute NML. To see this, note that the sum over all data patterns y of the maximum likelihood under a given model, $\sum_y f(y|\hat{\theta}(y))$, can be stated as the integral over $f(y|\hat{\theta}(y))$ with respect to the counting measure μ that assigns measure 1 to each data pattern y that can occur:

$$T = \sum_y f(y|\hat{\theta}(y)) = \int f(y|\hat{\theta}(y)) d\mu(y).$$

Much like in computing FIA_f, we need a density g with respect to μ from which we can (a) sample data patterns efficiently, and that (b) dominates the integrand, so that for some $M > 0$, $f(y|\hat{\theta}(y)) \leq Mg(y)$ for all y, and that (c) is similar to the integrand. For models on finite sample spaces, condition (b) is satisfied if the support

[2] Note, however, that the UVSDT is unique in that it can in principle predict below-chance performance (i.e., lower hit rate than false-alarm rate) even under the restriction $\mu \geq 0$.

[3] The asymmetry can be characterized (in z-coordinates, i.e., in plotting z-transformed hit rates against z-transformed false alarm rates, as well as in the non-transformed coordinates) in terms of the slope of the line that best fits the observed ROC: the slope is often smaller than one. In terms of the UVSDT, the expected slope (in z-coordinates) is $\frac{\sigma_n}{\sigma_s}$, so that the asymmetry implies $\sigma_s \geq \sigma_n$. In terms of the 2HTM, the expected slope (in untransformed coordinates) is $\frac{1-D_s}{1-D_n}$, so that the asymmetry implies $D_s \geq D_n$.

[4] Note that implementing four independent streams and the part of the accuracy criterion based on convergence across streams are superfluous. There is nothing in the theory or practice of integration by independent importance sampling that requires several independent streams of integration. Given that we had no prior experience with Monte Carlo integration, we used several independent streams and monitored convergence across streams (rather than only within streams) as an additional reassurance that the method works as intended.

of (the density proportional to f) is a subset of the support of density g.

Let $y_{r,t,i}$ be the frequency of response r, $r = \text{NO, YES}$, given noise trials ($t = n$) or signal trials ($t = s$), in (baserate or payoff) condition i, $i = 1, \ldots, K$ as stacked in y. We know that the maximum likelihood of a given model is dominated by the maximum likelihood of the saturated model for any given data set. Let $q_{s,i}$ and $q_{n,i}$ be the numbers of signal and noise trials, respectively, in (baserate or payoff) condition i, $i = 1, \ldots, K$. Thus,

$$f(y|\hat{\theta}(y)) \leq \prod_{i=1,\ldots,K} \prod_{t=s,n} \binom{q_{t,i}}{y_{\text{YES},t,i}} \times \left(\frac{y_{\text{NO},t,i}}{q_{t,i}}\right)^{y_{\text{NO},t,i}} \left(\frac{y_{\text{YES},t,i}}{q_{t,i}}\right)^{y_{\text{YES},t,i}},$$

the maximum likelihood of the data pattern y under the saturated model.

Let $h(y)$ be this maximum likelihood of y under the saturated model. Because there is a finite number of data patterns, there exists a number H so that $g(y) = H^{-1}h(y)$ is a density with respect to μ (i.e., so that the values $g(y)$ sum to one). Because of the above, g dominates f. Moreover, sampling from g can be done relatively efficiently. Because of the product structure of g, the frequency counts for each kind of trial and baserate or payoff condition are independent random variables under the probability distribution defined by g. Thus, sampling can proceed for each kind of trial and condition separately. We used rejection sampling for the purpose (Gelman, Carlin, Stern, & Rubin B, 2004, chap. 11).

Consider, for example, noise trials ($t = n$) in (baserate or payoff) condition 1 ($i = 1$). Generate first a candidate frequency count $y_{\text{YES},n,1}$ for YES responses from a uniform distribution on the integers from 0 to $q_{n,1}$ (and set $y_{\text{NO},n,1} = q_{n,1} - y_{\text{YES},n,1}$) along with a random number u from a uniform distribution on the interval [0,1] on the real line. Noting that

$$\xi = \binom{q_{n,1}}{y_{\text{YES},n,1}} \left(\frac{y_{\text{NO},n,1}}{q_{n,1}}\right)^{y_{\text{NO},n,1}} \left(\frac{y_{\text{YES},n,1}}{q_{n,1}}\right)^{y_{\text{YES},n,1}} \leq 1,$$

the frequency count is accepted if $u \leq \xi$. If $u > \xi$, the candidate frequency count is rejected, and a new candidate frequency and uniform variate u are generated.

Sampling m data patterns y^i, $i = 1, \ldots, m$, in this fashion, the quantity $T_m = \frac{1}{m}\sum_i \frac{f(y^i|\hat{\theta}(y^i))}{h(y^i)}$ approximates $H^{-1}T$ as m becomes large. Thus, the desired integral is approximated up to a multiplicative constant H^{-1}. This turns into an additive constant after taking logarithms. The constant can be estimated relatively easily and with high precision by monitoring the acceptance rates in rejection sampling as detailed in Appendix B.

Note that the saturated model is similar to the integrands to the degree to which these specify flexible models. This accounts for the fact that the algorithm converges with a speed that is roughly proportional to the flexibility of the model in question.

In computing NML in this way, it is essential that maximum likelihood estimation is implemented efficiently for each model and data pattern. We used a fast modified Newton algorithm (procedure E04LYF from the NAG FORTRAN library) for the purpose. Because the resulting algorithm converges much slower than that for computing FIA_f, we generated only one random stream of data patterns rather than four as for FIA_f and stopped computation if the length of the above logarithmized confidence interval with $z_{(1-\frac{\alpha}{2})} = 3$ for the integral in question (see subsection "FIA" above) was smaller than 0.1. This ensures that the integral in question is approximated with a precision of one decimal place; a precision that was deemed adequate for the purpose of model selection. The resulting algorithm converges within hours or faster for data sets with less than $K = 5$ baserate or payoff conditions, but it takes days to converge for data sets with $K \geq 5$ and large numbers of trials $q_{t,i}$ per condition (such as $q_{t,i} \geq 1000$) on a fast personal computer.[5] Nevertheless, it is much faster than the computation of NML by enumerating all data patterns, which is entirely infeasible for data sets of any size. A refinement of the algorithm that increases its speed considerably is described in Appendix B. Again, FORTRAN code can be obtained from the first author.

6. The flexibility of the models

The relative flexibility of the models considered here has been a recurrent question in the literature. So far, the question has primarily been addressed by means of simulation studies. For example, simulation studies by Bröder and Schütz (2009) suggested that the UVSDT is more flexible than the 2HTM, whereas Dube's et al. (2011) simulations suggested a draw between UVSDT and 2HTM. The difference between the two conclusions is most likely due to the heuristics of the sampling schemes employed in each study. For example, Bröder and Schütz (2009) directly generated random data sets with above-chance accuracy, whereas Dube et al. (2011) generated random parameter values within prespecified ranges for both models from which data sets were then generated (see also Wixted (2007)). In a similar vein, Jang et al. (2009) explored the flexibility of the EVSDT, UVSDT, DPSDT, and MSDT and the ability of AIC and BIC to recover the true model conditional on given data sets in an extended experimental paradigm. Use of FIA and NML provides a principled way to assess the relative flexibility of the models for ROC data obtained from baserate or payoff manipulations.

FIA. The flexibility of the model due to its functional form is quantified in terms of FIA_f, based on the minimum-description-length framework. FIA_f itself is a function of the model and of the experimental design in terms of the number K of baserate or payoff conditions and in terms of the ratios of signal to noise trials in each such condition. Thus, it has to be computed anew for each model, K, and baserate regime.

Fig. 2 shows the flexibility measure FIA_f for the ten models considered for a design with equal numbers of signal and noise trials in each condition as is typical of experiments with payoff manipulations and for $K = 2$ to $K = 5$ such conditions. The FIA_f values are computed under the restrictions that $\mu \geq 0$ for the continuous and hybrid models and $\mu \geq \mu^* \geq 0$ for the MSDT. Note that all models other than 1HTM, 2HTM($D_s = D_n$), and EVSDT require at least two baserate or payoff conditions to be identified and that MSDT requires three such conditions (therefore, there is no point for MSDT with $K = 2$ in Fig. 2).

Three models, 1HTM, 2HTM($D_s = D_n$), and EVSDT, employ $K + 1$ parameters; six models, 2HTM($D_s \geq D_n$), UVSDT($\sigma_s \geq \sigma_n$), 2HTM, UVSDT, DPSDT, and MSDT0, require $K + 2$ parameters; MSDT has $K + 3$ parameters. The complexity measures can be directly compared between models with the same number of parameters, and we connected the measures for these points by lines to facilitate the comparison in Fig. 2. For models that differ in the number of parameters, the flexibility penalties FIA_P differ not only in FIA_f, but also in the first summand, $\frac{p}{2}\log\frac{n}{2\pi}$, that depends on the number of parameters p and sample size n.

Comparing models that differ only in an inequality restriction (UVSDT($\sigma_s \geq \sigma_n$) and UVSDT as well as 2HTM($D_s \geq D_n$) and 2HTM), it can be seen that the inequality restriction is reflected in a smaller measure of flexibility for the restricted model. Furthermore, among the $K+1$ parameter models, flexibility

[5] The analyses were run on a personal computer equipped with two Intel® Xeon®W5580 processors with clock speed 3.2 GHz each comprising four cores enabling eight threads on a 32 bit operating system.

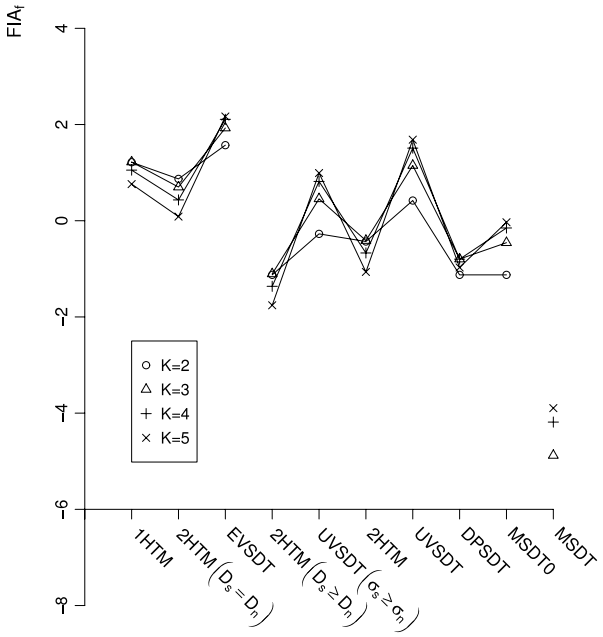

Fig. 2. FIA$_f$ for $K = 2$ to $K = 5$ conditions with equal numbers of signal and noise trials.

increases from 2HTM($D_s = D_n$) over 1HTM to EVSDT. Among the $K + 2$ parameter models, flexibility is consistently least for the 2HTM($D_s \geq D_n$), followed by DPSDT, 2HTM, and MSDT0, the flexibility ordering of which depends on k. Note, however, that MSDT0 is consistently at least as complex as DPSDT, and that both are consistently at least as complex as 2HTM($D_s \geq D_n$). Finally, UVSDT($\sigma_s \geq \sigma_n$) and UVSDT are the most flexible models. Note that these relationships also held when the ratios of signal to noise trials differ from condition to condition as is typical of baserate manipulations. In particular, the continuous models are always most flexible and the discrete-state models least flexible.[6]

NML. FIA is an asymptotic approximation to NML, raising the question of the accuracy of the approximation for data sets of realistic size. NML can be efficiently computed using the above method for small data sets such as data obtained from an individual observed under different baserate or payoff conditions. But FIA can be computed much more efficiently for large data sets as arise in the analysis of aggregate data sets in which response frequencies are summed over participants. Thus, it would be helpful if the FIA approximation were adequate for data sets of sizes as typically seen in analyses of aggregate data (see, e.g., the meta-analysis by Bröder and Schütz (2009, 2011), over 59 aggregate data sets, with trial numbers in the order of 1000 per individual hit or false alarm rate).

In particular, as sample size increases, NML should approach FIA, or equivalently, the expression

$$\mathrm{NML}_f = \mathrm{NML} - \frac{p}{2}\log\frac{n}{2\pi}$$

should approach FIA$_f$, a quantity that is independent of sample size. Fig. 3 shows FIA$_f$ and NML$_f$ values for $K = 2$, 3, 4, and 5 conditions with $n = 10$, 100, or 1000 trials for each individual hit or false alarm rate.

As can be seen, the NML values approach the FIA values from above so that FIA underestimates model flexibility in absolute terms. For model-selection purposes, the vertical positions of the NML and FIA profiles are however irrelevant, and the question of interest is whether FIA approximates the shape of the NML profiles, that is the differences between the models in NML values. To facilitate the evaluation of this question, Fig. 4 shows the NML$_f$ profiles aligned with the FIA$_f$ values for the 1HTM.

A couple of features are worthy of note. First of all, the shape of the NML$_f$ profiles follows that of FIA$_f$, although the NML$_f$ profiles are shallower. This means that the above statements about the relative flexibility of the models are not artifacts of a poor approximation of FIA to NML. Moreover, FIA$_f$ provides a good approximation of the NML$_f$ profiles for all $K + 1$ parameter models (i.e., 1HTM, 2HTM($D_s = D_n$), and EVSDT). For models with more parameters, FIA$_f$ is approached from above by the aligned NML$_f$ values so that FIA underestimates the flexibility of these models relative to the $K + 1$ parameter models. For data sets of sizes as are typical of aggregate data ($n = 1000$), FIA begins to be an adequate approximation also for the $K + 2$ parameter models. FIA consistently underestimates the flexibility of the $K + 3$ parameter model MSDT relative to the other models even for large data sets by a sizable amount (see Navarro (2004), for another example of small sample problems of FIA).

As a tentative conclusion, these analyses suggest that FIA can be used as a proxy for NML for large data sets and $K \geq 3$ with the exception of MSDT, for which the approximation began to be satisfactory only for the largest data set, $K = 5$, $n = 1000$. Thus, if the set of models that are entered into the model-selection exercise is to include MSDT, use of NML must be recommended for smaller data sets. These recommendations will be further refined by means of model recovery studies later.

Example. The size of the differences in flexibility due to functional form is such that it will often make a difference when it is taken into account in model selection. For example, Dube et al. (2011) report one experiment on syllogistic reasoning in which they manipulated perceived baserates in five steps (i.e., $K = 5$) while actually implementing a 50% ratio of signal (valid problems) to noise (invalid problems) trials in each condition. They fitted the 2HTM and the UVSDT and found $G^2(df = 3)$ values for goodness of fit of 3.30 and 0.87, respectively. They concluded that "although both models fit the data well, the SDT [UVSDT] model provided a better description of the data than the MPT [2HTM] model. This indicates that binary ROCs for syllogistic reasoning are not linear" (Dube et al., 2011, p. 158). With $p = 7$ parameters and a total of 3840 trials (there were 60 participants each of whom worked on 64 trials), the penalty term FIA$_P$ in FIA is $\mathrm{FIA}_P = \frac{7}{2}\log(\frac{3840}{2\pi}) + \mathrm{FIA}_f$ with FIA$_f$ given by the model-specific value shown in Fig. 2. Thus, FIA$_P$ equals 21.39 and 24.14 for 2HTM and UVSDT, respectively. Adding one half the G^2 values, FIA amounts to 23.04 and 24.57 for 2HTM and UVSDT, respectively, up to an additive constant.[7] Similarly, NML amounts to 23.3 and 24.7 for 2HTM and UVSDT, respectively.[8] The model with the smaller value is preferred. It turns out that Dube et al.'s conclusion has to be reversed when model complexity is factored in.

[6] For $K = 2$, the $K+2$ parameter models are saturated, they can perfectly account for many, but not all of the data that can arise, and they show overlap in the data that they can account for perfectly. Yet, it can be shown that only DPSDT and MSDT0 can account for exactly the same data for $K = 2$, whereas UVSDT, 2HTM, UVSDT($\sigma \geq 1$), 2HTM($D_s \geq D_n$), and DPSDT = MSDT0 all differ in the set of data that they can account for perfectly (see, e.g., footnote 2).

[7] One half G^2 differs from minus the maximum log-likelihood by only an additive constant.

[8] These analyses assume that UVSDT was fitted with the constraint that $\mu \geq 0$. If this was not the case, the penalty for UVSDT would be increased (by an amount of about log(2)), and the results would favor 2HTM even more strongly.

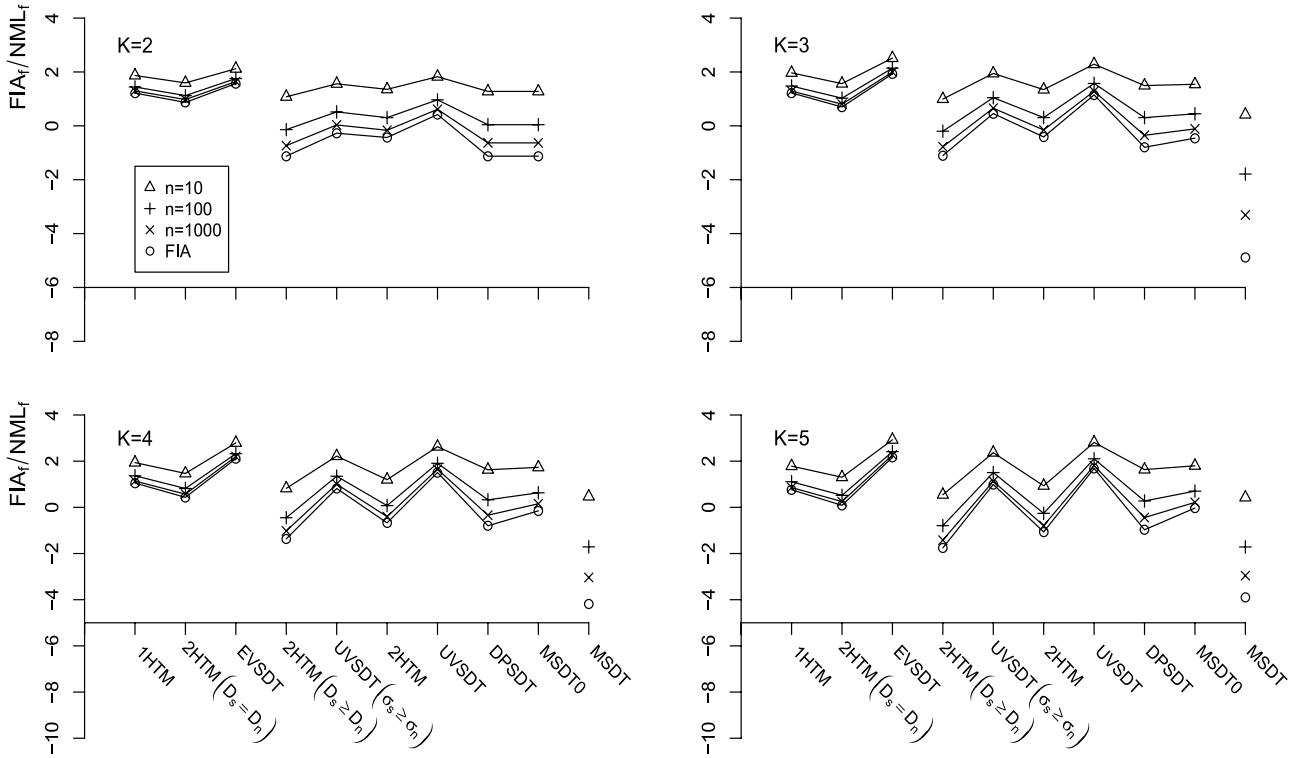

Fig. 3. NML_f and FIA_f for $K = 2$ to $K = 5$ conditions and $n = 10$, 100, and 1000 trials per hit and false alarm rate.

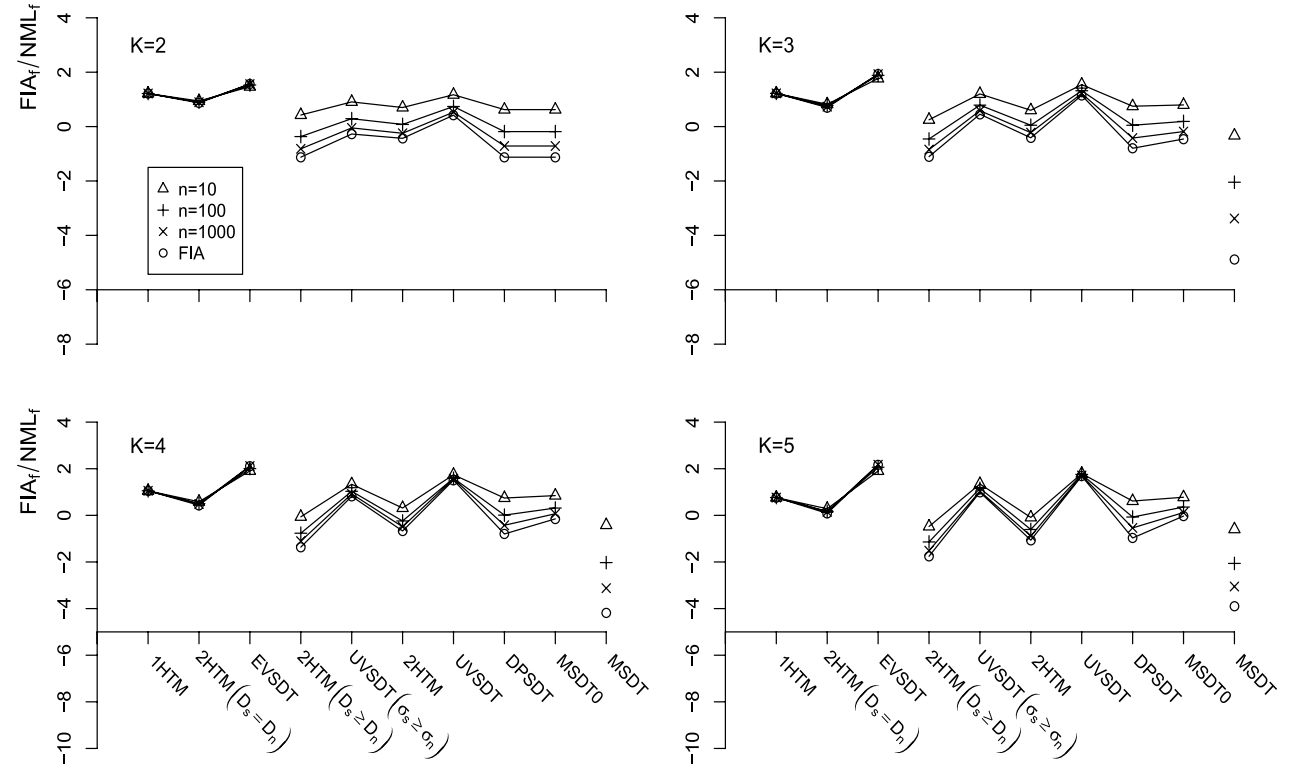

Fig. 4. NML_f and FIA_f for $K = 2$ to $K = 5$ conditions and $n = 10$, 100, and 1000 trials per hit and false alarm rate, aligned on FIA_f(1HTM).

7. Model recovery

The objective of model selection via the minimum-description length principle is to "identify a model family that permits the tightest compression of a data set by effectively filtering out random noise and attending to all of the 'useful' information in the data" (Myung et al., 2006, p. 173), and this is related to identifying the model that is most generalizable, leading to the smallest errors in predicting new data (Myung et al., 2006). A question researchers in recognition memory are often interested in is which model most likely generated a given data set. In this section, we address the issue whether FIA and NML are also useful,

and perhaps even perform better than AIC and BIC, in identifying the underlying model, assuming that one of the models under consideration indeed represents the "truth".

Addressing this question calls for model-recovery studies in which data are generated from a set of models and fitted with these same models. Each subset of models defines a new recovery problem, and we considered a range of these, including comparisons between models with the same number of parameters and between models with different numbers of parameters.

AIC, BIC, NML, and FIA are based on the same evidence, the models' maximum-likelihood values. They differ in how this evidence is used and in particular, how model complexity is quantified and weighed against the evidence. Models with larger penalties due to functional form will thereby be selected less frequently and models with smaller penalties more frequently under NML and FIA than under AIC and BIC, all else being equal. In other words, complex models will tend to be selected less frequently, entailing fewer correct selections if and when the complex model in fact generated the data and more correct selections if and when the simple model in fact generated the data. The question is whether the reduction in correct selections is less in the former case than the increase in correct selections in the latter case so that the overall number of correct decisions is increased when NML and FIA are used.

This is not a trivial question given that NML and FIA were not developed with the goal of optimizing model recovery. For instance, as is clear from the definition of NML, the minimum-description length principle quantifies complexity with respect to all data sets that can occur irrespective of whether they were generated from one of the models under consideration or not, whereas in model recovery models are contrasted with respect to data sets generated from each of these models (Wagenmakers et al., 2004). With respect to this subset of data sets, NML and FIA may or may not appropriately quantify the relative ability of the models to fit these data.[9]

For all analyses, the number of baserate or payoff conditions K was varied, from two to five, and the number of trials per individual hit or false alarm rate was varied in two steps, $n = 50$, as representative for analyses based on individual data, and $n = 1000$ as representative for analyses based on aggregate data. From each model in the set, we generated 10,000 data sets and summarize results in terms of recovery rates, the percentage of simulated data sets in which the model-selection index in question selected the generating model. Overall model-recovery performance is evaluated in terms of the percentage of correct selections across generating models.

A difficult question, requiring judicious choice, is how to sample from each model. Following Myung et al. (2007), we sampled parameter values from Jeffrey's noninformative distribution, with density proportional to $\sqrt{\det I}$ which assigns equal prior probability to every distinguishable probability distribution that is consistent with the model (Balasubramanian, 1997). This makes the sampling independent of the particular parameterization employed for the model, and this transformation-invariance thereby ensures that the sampling scheme is dependent only upon the model itself and not on a particular parameterization thereof. Sampling from Jeffrey's prior was implemented via rejection sampling based on the proposal densities and upper bounds discussed in Appendix A.

[9] Note, however, that Balasubramanian (1997) has shown that model selection by NML and FIA asymptotically approximates model selection by Bayes factors with Jeffrey's prior for each model, providing a theoretical rationale to expect good performance of NML and FIA in the present context (see also Grünwald and Navarro (2009), Karabatsos and Walker (2006)). This is also true of BIC (Wasserman, 2000).

Consider first the pairwise comparisons between non-nested models in the set. For pairwise comparisons, it is relatively simple to assess the optimal level of overall model-recovery performance that can be achieved on the basis of the model's likelihood values as this amounts to finding the optimal value for the penalty difference for the two models in question. For each analysis, we determined the optimal performance level maximizing the overall percentage of correct selections across generating models by means of a fine-grained grid search over a wide range of penalty differences.

Table 1 shows four examples of these analyses, contrasting (a) 1HTM and EVSDT, (b) 2HTM and UVSDT, (c) UVSDT($\sigma_s \geq \sigma_n$) and DPSDT, and (d) DPSDT and MSDT. Note that the first three comparisons contrast models with equal numbers of parameters so that model selection based on AIC and BIC effectively considers only goodness of fit.

Table 1 illustrates several points. As can be seen, NML always performed at least as well as AIC and BIC, and sometimes considerably better (see, e.g., the contrast of 2HTM and UVSDT). In all cases, NML approached the optimal possible recovery performance. FIA generally does as well for larger data sets ($k \geq 3$ or $n = 1000$), with the exception of the comparisons involving MSDT, where the small-sample problem of FIA in approximating NML for MSDT, already noted above, causes its performance in model recovery to decline noticeably, except for the largest data sets with $K \geq 4$ and $n = 1000$. For example, there are several zero cells in FIA performance in contrasting DPSDT and MSDT, reflecting the fact that FIA strongly underestimates the flexibility of the MSDT and thereby strongly favors its selection over that of the DPSDT.

Table 2 shows two examples of recovering from sets of four models, specifically from (a) 2HTM($D_s = D_n$), EVSDT, 2HTM, and UVSDT, and from (b) the four $K + 2$ parameter models that account for the frequently observed asymmetry in ROCs, namely 2HTM($D_s \geq D_n$), UVSDT($\sigma_s \geq \sigma_n$), DPSDT, and MSDT0. These contrasts reveal the same pattern as the pairwise comparisons, with FIA functioning well as an approximation to NML for the larger data sets, and NML improving on AIC and BIC.

A final point that is noteworthy is that model recovery was also well above chance performance in all cases even for $K = 2$, where ROC analyses are not helpful in distinguishing between (the $K + 2$ parameter) models. In fact, it has been repeatedly asserted in the literature that data with $K = 2$ are not diagnostic in discriminating between models for this reason (e.g., Dube and Rotello (in press), Dube et al. (2011)). Although the $K + 2$ parameter models can account perfectly for many data sets with $K = 2$, they do impose differently restrictive inequality restrictions on such data, making the data sometimes diagnostic for discriminating between models.

Nevertheless, even for $K > 2$, many of the generated data sets are relatively non-diagnostic for discriminating between the models. For example, for data sets with performance near chance, or with small differences in sampled response bias, all models frequently make equivalent predictions, so that in line with the parsimony principle the least flexible model is selected irrespective of which model generated the data. This points to the importance of optimizing experimental design in collecting real data so as to maximize the expected differences in model predictions. Progress regarding the maximization of data diagnosticity is currently being made for different classes of models and paradigms (e.g., Cavagnaro, Myung, Pitt, and Kujala (2010)). Such developments have the potential to be of great value if implemented for the models considered here.

Taken together, NML makes better use of the evidence than AIC and BIC in model recovery, and is therefore useful for that purpose although not designed to optimize model recovery. The use of FIA can also be recommended for larger data sets as typically underlie analyses of aggregate data.

Table 1
Model recovery contrasting different models 1 and 2: recovery rates per model (1 vs. 2), overall number of correct recoveries, and optimal recovery rates in percent.

			1HTM vs. EVSDT			2HTM vs. UVSDT			UVSDT($\sigma_s \geq \sigma_n$) vs. DPSDT			DPSDT vs. MSDT		
K	n	Index	1	2	Cor.	1	2	Cor.	1	2	Cor.	1	2	Cor.
2	50	AIC	53	85	69	44	77	60	72	51	62			
		BIC	53	85	69	44	77	60	72	51	62			
		NML	87	60	73	96	38	67	40	96	68			
		FIA	87	59	73	97	35	66	35	98	66	n.a.[a]		
		Opt.	87	58	73	88	49	69	49	89	69			
	1000	AIC	81	93	87	49	78	63	77	51	64			
		BIC	81	93	87	49	78	63	77	51	64			
		NML	93	85	89	99	50	74	51	99	75			
		FIA	93	85	89	99	50	74	50	99	75			
		Opt.	93	85	89	95	56	76	56	95	76			
3	50	AIC	67	90	79	54	90	72	77	57	67	98	9	54
		BIC	67	90	79	54	90	72	77	57	67	99	1	50
		NML	92	75	83	94	64	79	54	94	74	88	30	59
		FIA	92	74	83	96	60	78	47	97	72	0	100	50
		Opt.	92	75	83	90	69	80	56	92	74	82	36	59
	1000	AIC	91	97	94	82	96	89	88	72	80	96	46	71
		BIC	91	97	94	82	96	89	88	72	80	99	21	60
		NML	97	93	95	98	88	93	75	97	86	89	56	73
		FIA	97	93	95	98	88	93	74	98	86	0	99	50
		Opt.	97	93	95	97	89	93	77	95	86	83	63	73
4	50	AIC	74	94	84	63	94	78	84	59	71	97	24	60
		BIC	74	94	84	63	94	78	84	59	71	99	8	53
		NML	94	84	89	95	76	86	62	94	78	87	40	64
		FIA	95	83	89	97	73	85	56	98	77	0	99	50
		Opt.	94	84	89	93	80	86	64	93	79	81	48	64
	1000	AIC	94	98	96	92	98	95	93	77	85	94	65	79
		BIC	94	98	96	92	98	95	93	77	85	99	42	71
		NML	99	96	97	99	95	97	84	98	91	93	66	79
		FIA	99	96	97	99	95	97	83	98	91	77	78	78
		Opt.	98	96	97	99	96	97	85	97	91	90	69	80
5	50	AIC	78	96	87	70	96	83	87	62	74	95	36	65
		BIC	78	96	87	70	96	83	87	62	74	99	14	57
		NML	97	88	92	97	84	91	68	95	82	87	49	68
		FIA	97	87	92	98	82	90	62	98	80	0	99	50
		Opt.	96	88	92	95	87	91	70	94	82	85	52	68
	1000	AIC	95	99	97	95	99	97	95	80	88	93	76	84
		BIC	95	99	97	95	99	97	95	80	88	99	56	78
		NML	99	97	98	99	98	99	89	98	93	95	74	84
		FIA	99	97	98	99	98	99	88	99	93	90	78	84
		Opt.	99	97	98	99	98	99	90	98	94	94	75	85

Note. Cor.=Percentage of correct recoveries across models; Opt.=Penalty difference maximizing correct recoveries across models.
[a] MSDT requires $K \geq 3$ to be identified; for $K = 2$, MSDT0, MSDT, and DPSDT are basically reparameterizations of the same model.

8. General discussion

Use of the minimum-description length principle in model selection, although more sophisticated than model selection via AIC and BIC, has been hampered by the difficulty of computing the model-selection indices FIA and NML flowing from the principle. In this article, we presented methods to compute FIA and NML for the major measurement models of recognition memory currently in use. This allowed us to evaluate the relative flexibility of these models due to functional form. It turns out that the discrete-state models are consistently less flexible than the continuous models with the hybrid models in between. The size of these differences is such that it will frequently lead to different results in model selection compared to the use of the traditional model-selection indices AIC and BIC as exemplified for a real data set and through model-recovery studies.

In evaluating the goodness of the approximation of FIA to NML, FIA provided a good approximation for the relative flexibility of the $K + 1$ parameter models (1HTM, 2HTM($D_s = D_n$), and EVSDT) for data sets of any of the sizes considered, and in addition, good approximations for the $K + 2$ parameter models (2HTM($D_s \geq D_n$), 2HTM, UVSDT($\sigma_s \geq \sigma_n$), UVSDT, DPSDT, and MSDT0) for large data sets as typically arise in the analysis of aggregate data. But FIA begins to approximate NML for the MSDT model only for the largest data sets that were analyzed.

Reassuringly, although not designed for the purpose, use of NML consistently led to better overall model recovery than the use of AIC and BIC, approximating the optimal use of the statistical evidence as could be ascertained for pairwise model comparisons. FIA performed about as well as NML for the larger data sets in line with the results evaluating the goodness of the approximation of FIA to NML.

Model recovery assumes that one of the models under consideration is the "true" model that in fact generated the data. For comparisons between just two models, an alternative tailor-made for the purpose is given by Wagenmaker's et al. (2004) bootstrap approach (see also Myung et al. (2007); Navarro, Pitt, and Myung (2004)), and it would be interesting to pit the two methods, the one based on the minimum-description length principle and Wagenmakers et al.'s, against each other to compare their performance. For recovering from sets of models comprising more than two models, a similar competitor does not exist. What is more, for real data, each model will usually at best provide an approximation to the true model and in this situation, NML and

Table 2
Model recovery contrasting sets of four models: recovery rates per model (1 to 4), and overall number of correct recoveries in percent.

K	n		2HTM($D_s = D_n$) EVSDT, 2HTM, UVSDT					2HTM($D_s \geq D_n$), UVSDT($\sigma_s \geq \sigma_n$), DPSDT, MSDT0				
		Index	1	2	3	4	Cor.	1	2	3	4[a]	Cor.
2	50	AIC	55	78	5	49	47	51	64	36		50
		BIC	59	82	1	38	45	51	64	36		50
		NML	86	56	23	35	50	95	37	22		51
		FIA	72	0	68	33	43	37	34	83		51
	1000	AIC	72	81	30	70	63	54	71	38		55
		BIC	82	91	19	64	64	54	71	38		55
		NML	93	77	57	50	70	98	50	29		59
		FIA	93	76	60	50	70	34	50	93		59
3	50	AIC	57	80	25	63	56	54	67	30	26	44
		BIC	65	89	11	51	54	54	67	30	26	44
		NML	89	64	45	48	61	90	41	32	22	46
		FIA	82	42	62	47	58	87	37	41	19	46
	1000	AIC	75	83	72	89	80	76	79	51	38	61
		BIC	89	97	59	82	82	76	79	51	38	61
		NML	96	89	77	79	85	92	60	60	53	66
		FIA	96	88	78	79	85	91	59	64	51	66
4	50	AIC	56	83	37	71	62	55	72	37	29	48
		BIC	67	93	19	59	59	55	72	37	29	48
		NML	90	74	54	57	69	92	46	40	32	52
		FIA	84	62	66	57	67	90	43	48	26	52
	1000	AIC	76	85	84	93	84	82	83	63	50	70
		BIC	91	98	72	87	87	82	83	63	50	70
		NML	97	93	84	87	90	95	66	73	66	75
		FIA	97	93	84	87	90	95	66	75	64	75
5	50	AIC	57	84	45	76	66	57	76	41	34	52
		BIC	67	95	25	66	63	57	76	41	34	52
		NML	91	78	60	64	73	94	51	45	39	57
		FIA	86	73	68	63	72	92	48	55	32	57
	1000	AIC	78	85	88	95	86	84	88	69	59	75
		BIC	92	99	78	90	90	84	88	69	59	75
		NML	97	96	86	91	93	97	73	80	72	81
		FIA	97	96	87	91	92	96	73	82	71	81

Note. Cor.=Percentage of correct recoveries across models.
[a] For $K = 2$, DPSDT and MSDT0 are basically reparameterizations of the same model.

FIA can best fulfill the purpose for which they were designed, that is, to identify a model that provides the tightest compression of the data.

Note that the results described in this article are paradigm-specific and model-specific, that is they are restricted to the range of models considered here and the paradigm with baserate and payoff manipulations. Although they add to the growing evidence for the usefulness of NML and FIA (e.g., Myung et al. (2007), Su et al. (2005), Wu et al. (2010a,b)), there is no guarantee that NML and FIA will perform as well for other models and/or paradigms such as obtaining ROC data via confidence ratings, and new analyses are required for each new area of application. Note also, however, that the method for computing NML proposed here is potentially useful for any kind of frequency data and in particular may be applicable to computing NML for the paradigm with confidence ratings.

Similarly, the revealed differences in model flexibility are in part paradigm-specific. For example, Kellen and Klauer (2011) analyzed these models in the context of the 4AFC-2R task in which four stimuli are shown in each trial, one signal and three noise stimuli, and the respondent has to pick out the signal stimulus for which he or she has two choices. In this paradigm, the $K + 2$ parameter models were ordered according to flexibility as MSDT0 $=$ DPSDT $<$ 2HTM $<$ UVSDT based on analyses of the space of probability distributions generated by these models as well as on NML analyses.

As a next step, we intend to apply these methods to reanalyzing the large number of data sets available in the recognition-memory literature that have used the paradigm with response-bias manipulations. Previous meta-analyses of these data sets (Bröder & Schütz, 2009, 2011; Dube & Rotello, in press) relied only on goodness of fit and compared only 2HTM and UVSDT. It will be interesting to see whether the use of the modern selection indices applied to a larger set of candidate models will suggest a clear conclusion in the ongoing debate about which of these models provides the best account of the extant data.

The paradigm with baserate or payoff manipulations and the models considered here are widely used not only in recognition memory, but also in perception, and with diminishing frequency in fields such as reasoning, attention, and working memory. The present results should thereby be informative, and the present methods useful, for many researchers working in these different fields.

Acknowledgment

The research reported in this paper was supported by grant Kl 614/32-1 from the Deutsche Forschungsgemeinschaft to the first author.

Appendix A. The Fisher information matrix, its determinant, and sampling densities

Let $p_{s,i}$ and $p_{n,i}$ be the probabilities of hits and false alarms, respectively, in (baserate or payoff) condition i. For any of the models considered here, these are expressed as a function of p parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Let $q_{s,i}$ and $q_{n,i}$ be the numbers

of signal and noise trials, respectively, in condition i, $i = 1, \ldots, K$. Using a result by Su et al. (2005) for models specifying the probability distribution of independent binomial random variables, a generic form of the Fisher information matrix for the present models is a p by p matrix with elements $i_{u,v}$ given by:

$$i_{u,v} = \sum_{i=1}^{K} \left(\frac{q_{n,i}}{p_{n,i}(1-p_{n,i})} \frac{\partial p_{n,i}}{\partial \theta_u} \frac{\partial p_{n,i}}{\partial \theta_v} + \frac{q_{s,i}}{p_{s,i}(1-p_{s,i})} \frac{\partial p_{s,i}}{\partial \theta_u} \frac{\partial p_{s,i}}{\partial \theta_v} \right).$$

The trial numbers $q_{s,i}$ and $q_{n,i}$ were scaled by a multiplicative constant so that they summed to one, $\sum_i (q_{s,i} + q_{n,i}) = 1$, because the complexity term in FIA assumes that the Fisher information matrix is computed for a sample of size one.

1HTM

For the 1HTM, the partial derivatives of $p_{s,i}$ and $p_{n,i}$ are zero except for

$$\frac{\partial}{\partial g_i} p_{s,i} = 1 - D,$$
$$\frac{\partial}{\partial D} p_{s,i} = 1 - g_i, \quad \text{and}$$
$$\frac{\partial}{\partial g_i} p_{n,i} = 1.$$

Let

$$a_i = \frac{(1-D)^2}{p_{s,i}(1-p_{s,i})} q_{s,i},$$
$$b_i = \frac{(1-g_i)^2}{p_{s,i}(1-p_{s,i})} q_{s,i}, \quad \text{and}$$
$$c_i = \frac{1}{p_{n,i}(1-p_{n,i})} q_{n,i}.$$

It follows that the Fisher information matrix I is given by

$$\begin{pmatrix} a_1 + c_1 & 0 & \cdots & 0 & \sqrt{a_1 b_1} \\ 0 & a_2 + c_2 & \cdots & 0 & \sqrt{a_2 b_2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_K + c_K & \sqrt{a_K b_K} \\ \sqrt{a_1 b_1} & \sqrt{a_2 b_2} & \cdots & \sqrt{a_K b_K} & \sum_i b_i \end{pmatrix}.$$

Its determinant is

$$\det I = \sum_i b_i c_i \prod_{j \neq i} (a_j + c_j).$$

To see this, note that

$$I = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

with A being a $K \times K$ diagonal matrix with i-th diagonal element $a_i + c_i$, $C = (\sqrt{a_1 b_1}, \ldots, \sqrt{a_K b_K})$, $B = C'$, and $D = \sum_i b_i$. The above expression of the determinant follows from $\det(I) = \det(A) \det(D - CA^{-1}B)$ (Meyer, 2001, chap. 6). This method can be used to compute the determinants of the Fisher information matrix for all of the models considered here. As shown by Wu et al. (2010b), a suitable density g for the Monte Carlo integration is given by the product of Beta distributions with $\alpha = \beta = \frac{1}{2}$, one for each parameter g_i, $i = 1, \ldots, K$, and D.

2HTM with $D = D_s = D_n$

For the 2HTM($D_s = D_n$), the partial derivatives are zero except for

$$\frac{\partial}{\partial g_i} p_{s,i} = 1 - D,$$
$$\frac{\partial}{\partial D} p_{s,i} = 1 - g_i,$$
$$\frac{\partial}{\partial g_i} p_{n,i} = 1 - D, \quad \text{and}$$
$$\frac{\partial}{\partial D} p_{n,i} = -g_i.$$

Let

$$a_i = \frac{(1-D)^2}{p_{s,i}(1-p_{s,i})} q_{s,i}, \quad \text{and}$$
$$b_i = \frac{(1-D)^2}{p_{n,i}(1-p_{n,i})} q_{n,i}.$$

It follows that the Fisher information matrix I is as given in Box I.

Its determinant is

$$\det I = \frac{1}{(1-D)^2} \sum_{i=1}^{k} a_i b_i \prod_{j \neq i} (a_j + b_j).$$

As shown by Wu et al. (2010b), a suitable density g for the Monte Carlo integration is given by the product of Beta distributions with $\alpha = \beta = \frac{1}{2}$, one for each parameter g_i, $i = 1, \ldots, K$, and D.

2HTM

For the 2HTM, the partial derivatives are zero except for

$$\frac{\partial}{\partial g_i} p_{s,i} = 1 - D_s,$$
$$\frac{\partial}{\partial D_s} p_{s,i} = 1 - g_i,$$
$$\frac{\partial}{\partial g_i} p_{n,i} = 1 - D_n, \quad \text{and}$$
$$\frac{\partial}{\partial D_n} p_{n,i} = -g_i.$$

Let

$$a_i = \frac{(1-D_s)^2}{p_{s,i}(1-p_{s,i})} q_{s,i}, \quad \text{and}$$
$$b_i = \frac{(1-D_n)^2}{p_{n,i}(1-p_{n,i})} q_{n,i}.$$

It follows that the Fisher information matrix I is given by

$$\begin{pmatrix} a_1 + b_1 & 0 & \cdots & 0 & \frac{a_1(1-g_1)}{1-D_s} & \frac{-b_1 g_1}{1-D_n} \\ 0 & a_2 + b_2 & \cdots & 0 & \frac{a_2(1-g_2)}{1-D_s} & \frac{-b_2 g_2}{1-D_n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_K + b_K & \frac{a_K(1-g_K)}{1-D_s} & \frac{-b_K g_K}{1-D_n} \\ \frac{a_1(1-g_1)}{1-D_s} & \frac{a_2(1-g_2)}{1-D_s} & \cdots & \frac{a_K(1-g_K)}{1-D_s} & \frac{\sum_i a_i (1-g_i)^2}{(1-D_s)^2} & 0 \\ \frac{-b_1 g_1}{1-D_n} & \frac{-b_2 g_2}{1-D_n} & \cdots & \frac{-b_K g_K}{1-D_n} & 0 & \frac{\sum_i b_i g_i^2}{(1-D_n)^2} \end{pmatrix}.$$

Its determinant is

$$\det I = \frac{1}{(1-D_s)^2 (1-D_n)^2} \sum_{i<j} a_i b_i a_j b_j (g_i - g_j)^2 \prod_{l \neq i,j} (a_l + b_l).$$

As shown by Wu et al. (2010b), a suitable density g for the Monte Carlo integration is given by the product of Beta distributions with $\alpha = \beta = \frac{1}{2}$, one for each parameter g_i, $i =$

$$\begin{pmatrix} a_1+b_1 & 0 & \cdots & 0 & \frac{(a_1(1-g_1)-b_1g_1)}{1-D} \\ 0 & a_2+b_2 & \cdots & 0 & \frac{(a_2(1-g_2)-b_2g_2)}{1-D} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_K+b_K & \frac{(a_K(1-g_K)-b_Kg_K)}{1-D} \\ \frac{(a_1(1-g_1)-b_1g_1)}{1-D} & \frac{(a_2(1-g_2)-b_2g_2)}{1-D} & \cdots & \frac{(a_K(1-g_K)-b_Kg_K)}{1-D} & \frac{\sum_i a_i(1-g_i)^2+b_ig_i^2}{(1-D)^2} \end{pmatrix}.$$

Box I.

$1, \ldots, K, D_s$, and D_n. To impose the inequality restriction $D_n \leq D_s$ for 2HTM($D_s \geq D_n$), the sampled Beta values for D_n and D_s were ordered and then assigned to D_n and D_s in the appropriate order, whereas the product of Beta densities, g, was multiplied by 2.

EVSDT

Let F and f be the cumulative distribution function and the density function, respectively, of the standard normal distribution. Using parameters $\boldsymbol{\theta} = (c_1, c_2, \ldots, c_K, \mu)$, it is convenient to parameterize the EVSDT as follows:

$$p_{s,i} = F\left(\frac{\mu}{2} - c_i\right),$$
$$p_{n,i} = F\left(-\frac{\mu}{2} - c_i\right).$$

The partial derivatives are zero except for

$$\frac{\partial}{\partial c_i} p_{s,i} = -f\left(\frac{\mu}{2} - c_i\right),$$
$$\frac{\partial}{\partial \mu} p_{s,i} = \frac{1}{2} f\left(\frac{\mu}{2} - c_i\right),$$
$$\frac{\partial}{\partial c_i} p_{n,i} = -f\left(-\frac{\mu}{2} - c_i\right), \quad \text{and}$$
$$\frac{\partial}{\partial \mu} p_{n,i} = -\frac{1}{2} f\left(-\frac{\mu}{2} - c_i\right).$$

Let

$$a_i = \frac{f^2\left(\frac{\mu}{2} - c_i\right)}{p_{s,i}(1-p_{s,i})} q_{s,i}, \quad \text{and}$$
$$b_i = \frac{f^2\left(-\frac{\mu}{2} - c_i\right)}{p_{n,i}(1-p_{n,i})} q_{n,i}.$$

It follows that the Fisher information matrix I is given by

$$\begin{pmatrix} a_1+b_1 & 0 & \cdots & 0 & \frac{1}{2}(b_1-a_1) \\ 0 & a_2+b_2 & \cdots & 0 & \frac{1}{2}(b_2-a_2) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_K+b_K & \frac{1}{2}(b_K-a_K) \\ \frac{1}{2}(b_1-a_1) & \frac{1}{2}(b_2-a_2) & \cdots & \frac{1}{2}(b_K-a_K) & \frac{1}{4}\sum_i a_i+b_i \end{pmatrix}.$$

Its determinant is

$$\det I = \sum_{i=1}^{k} a_i b_i \prod_{j \neq i} (a_j + b_j).$$

Finding an appropriate density g for the Monte Carlo integration with $\sqrt{\det I} \leq Mg$ for some $M > 0$ uses the following result: For $a < 1$ there exists a number $M' > 0$ such that for all x

$$\frac{e^{-x^2}}{F(x)(1-F(x))} \leq M' e^{-\frac{1}{2}ax^2}.$$

It is sufficient to show this for $x \geq 0$. Because $F(x) \geq \frac{1}{2}$ for $x \geq 0$, it is sufficient to show that there is a number M^* with

$$\frac{e^{-x^2}}{1-F(x)} \leq M^* e^{-\frac{1}{2}ax^2}$$

for all $x \geq 0$. Setting $b = 1 - \frac{a}{2}$, this is equivalent to $\frac{\exp(-bx^2)}{1-F(x)} \leq M^*$. Note for later reference that $b > \frac{1}{2}$ (because $a < 1$).

Note furthermore that $1-F(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-\frac{1}{2}t^2)dt$. Hence,

$$\frac{e^{-bx^2}}{1-F(x)} = \sqrt{2\pi} \frac{e^{-\left(b-\frac{1}{2}\right)x^2}}{\int_x^\infty e^{\frac{1}{2}(x^2-t^2)}dt}.$$

Using a variable transformation $t \to t + x$, the term in the denominator, $\int_x^\infty \exp(\frac{1}{2}(x^2-t^2))dt$ is seen to equal $\int_0^\infty \exp(\frac{1}{2}(x^2-(t+x)^2))dt = \int_0^\infty \exp(-\frac{1}{2}(t^2+2xt))dt$, which is trivially greater than $\int_0^1 \exp(-\frac{1}{2}(t^2+2xt))dt$. Because t in the latter integral is always smaller or equal to one, this is in turn greater than $\int_0^1 \exp(-\frac{1}{2}(1+2xt))dt$, which equals $\exp(-\frac{1}{2})\frac{1-\exp(-x)}{x}$. Furthermore, it is not difficult to see, using for example, the Taylor series expression of $\exp(x)$ that

$$\frac{1-e^{-x}}{x} = e^{-x}\frac{e^x-1}{x} \geq e^{-x}.$$

It follows that

$$\frac{e^{-bx^2}}{1-F(x)} \leq \sqrt{2\pi} \frac{e^{-\left(b-\frac{1}{2}\right)x^2}}{e^{-\frac{1}{2}}e^{-x}} = \sqrt{2\pi} e^{\frac{1}{2}} e^{-\left(b-\frac{1}{2}\right)x^2+x}.$$

Because $b - \frac{1}{2} > 0$, $-(b-\frac{1}{2})x^2 + x$ is a quadratic function in x with a finite maximum. This completes the proof.

It follows, absorbing $q_{s,i}$ and $q_{n,i}$ into M', that for some $\bar{M} > 0$ and all i

$$\sqrt{a_i} \leq \bar{M} e^{-\frac{1}{4}a\left(\frac{\mu}{2}-c_i\right)^2}, \quad \text{and}$$
$$\sqrt{b_i} \leq \bar{M} e^{-\frac{1}{4}a\left(-\frac{\mu}{2}-c_i\right)^2}.$$

The determinant of the Fisher information matrix can be expressed as the sum of products of terms a_i and b_i by multiplying out the product to the right in the above equation for the determinant. The sum can be characterized as one over pairs of vectors $(\boldsymbol{k}_a, \boldsymbol{k}_b)$ with $\boldsymbol{k}_a = (k_a(1), \ldots, k_a(K))$ and $\boldsymbol{k}_b = (k_b(1), \ldots, k_b(K))$ containing zeros and ones, and with $k_a(i) = k_b(i) = 1$ for one and only one i (for the term a_ib_i left to the product over $j \neq i$), and $k_a(j) + k_b(j) = 1$ for $j \neq i$. The determinant is the sum over all $K2^{K-1}$ such pairs of vectors of the products $\prod_i a_i^{k_a(i)} b_i^{k_b(i)}$.

Hence, we have (using $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$)

$$\sqrt{\det I} = \sqrt{\sum_{(\boldsymbol{k}_a, \boldsymbol{k}_b)} \prod_i a_i^{k_a(i)} b_i^{k_b(i)}}$$
$$\leq \sum_{(\boldsymbol{k}_a, \boldsymbol{k}_b)} \prod_i \sqrt{a_i^{k_a(i)} b_i^{k_b(i)}}$$

$$\leq \bar{M}^{2K} \sum_{(\boldsymbol{k}_a,\boldsymbol{k}_b)} e^{-\frac{1}{4}a\sum_i\left[k_a(i)\left(\frac{\mu}{2}-c_i\right)^2+k_b(i)\left(-\frac{\mu}{2}-c_i\right)^2\right]}$$
$$\leq \bar{M}^{2K} K 2^{K-1} (2\pi)^{\frac{K+1}{2}} \times \sum_{(\boldsymbol{k}_a,\boldsymbol{k}_b)} \frac{1}{K2^{K-1}} \sqrt{\det \boldsymbol{\Sigma}_{\boldsymbol{k}_a,\boldsymbol{k}_b}} f_{\left(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{k}_a,\boldsymbol{k}_b}\right)}(\boldsymbol{\theta}),$$

where $f_{\left(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{k}_a,\boldsymbol{k}_b}\right)}$ is the probability density function of a $K+1$-variate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{k}_a,\boldsymbol{k}_b}$ defined by its inverse as in Box II.

Exploiting the striking structural analogy between this matrix and the Fisher information of the EVSDT model, I, it follows that the determinant of this matrix (ignoring the factor $\frac{a}{2}$) is given by the formula for I with a_i and b_i replaced by $k_a(i)$ and $k_b(i)$, and hence, that it equals one. This follows from the definition of $(\boldsymbol{k}_a, \boldsymbol{k}_b)$ noting that all $k_a(j)k_b(j)$ equal zero whereas $k_a(j)+k_b(j)=1$ except for one j for which $k_a(j)k_b(j)=1$. Taking the factor $\frac{a}{2}$ into account, reveals that the determinant of the matrix is $\left(\frac{a}{2}\right)^{K+1}$. Hence,

$$\sqrt{\det I} \leq \bar{M}^{2K} K 2^{K-1} (2\pi)^{\frac{K+1}{2}} \left(\frac{2}{a}\right)^{\frac{K+1}{2}} \times \sum_{(\boldsymbol{k}_a,\boldsymbol{k}_b)} \frac{1}{K2^{K-1}} f_{\left(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{k}_a,\boldsymbol{k}_b}\right)}(\boldsymbol{\theta}).$$

The function

$$g(\boldsymbol{\theta}) = \sum_{(\boldsymbol{k}_a,\boldsymbol{k}_b)} \frac{1}{K2^{K-1}} f_{\left(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{k}_a,\boldsymbol{k}_b}\right)}(\boldsymbol{\theta})$$

is the density function of a mixture of multivariate normal distributions, and as just shown, it dominates $\sqrt{\det I}$. We sampled from g with $a=0.9$ for the Monte Carlo integration. Note that the integration scheme based on g integrates over positive and negative values of μ, whereas the parameter space of the EVSDT is restricted to values $\mu \geq 0$. We therefore rejected samples from g with $\mu < 0$ and multiplied g by 2 in correction.

UVSDT

Using parameters $\boldsymbol{\theta} = (c_1, c_2, \ldots, c_K, \mu, s)$, it is convenient to parameterize the UVSDT as follows:

$$p_{s,i} = F\left(\sqrt{s}\left(\frac{\mu}{2}-c_i\right)\right)$$
$$p_{n,i} = F\left(-\frac{\mu}{2}-c_i\right).$$

The partial derivatives are zero except for

$$\frac{\partial}{\partial c_i} p_{s,i} = -\sqrt{s} f\left(\sqrt{s}\left(\frac{\mu}{2}-c_i\right)\right),$$
$$\frac{\partial}{\partial \mu} p_{s,i} = \sqrt{s}\frac{1}{2} f\left(\sqrt{s}\left(\frac{\mu}{2}-c_i\right)\right),$$
$$\frac{\partial}{\partial s} p_{s,i} = \frac{1}{2\sqrt{s}}\left(\frac{\mu}{2}-c_i\right) f\left(\sqrt{s}\left(\frac{\mu}{2}-c_i\right)\right),$$
$$\frac{\partial}{\partial c_i} p_{n,i} = -f\left(-\frac{\mu}{2}-c_i\right), \quad \text{and}$$
$$\frac{\partial}{\partial \mu} p_{n,i} = -\frac{1}{2} f\left(-\frac{\mu}{2}-c_i\right).$$

Let

$$a_i = \frac{f^2\left(\sqrt{s}\left(\frac{\mu}{2}-c_i\right)\right)}{p_{s,i}(1-p_{s,i})} q_{s,i} \quad \text{and}$$
$$b_i = \frac{f^2\left(-\frac{\mu}{2}-c_i\right)}{p_{n,i}(1-p_{n,i})} q_{n,i}.$$

It follows that the Fisher information matrix I is as given in Box III.

Its determinant is

$$\det I = \frac{1}{4}\sum_{i<j} a_i b_i a_j b_j (c_i-c_j)^2 \prod_{l\neq i,j}(a_l s + b_l).$$

For deriving a density dominating $\sqrt{\det I}$, it is again convenient to express the Fisher information as a sum of products, using pairs of vectors $(\boldsymbol{k}_a, \boldsymbol{k}_b)$ with $\boldsymbol{k}_a = (k_a(1), \ldots, k_a(K))$ and $\boldsymbol{k}_b = (k_b(1), \ldots, k_b(K))$ containing zeros and ones, and with $k_a(i) = k_b(i) = 1$ and $k_a(j) = k_b(j) = 1$ for one and only one pair of i and j with $i < j$, and with $k_a(l) + k_b(l) = 1$ for $l \neq i, j$. For a given pair of vectors, denote these special i and j by $i(\boldsymbol{k}_a, \boldsymbol{k}_b)$ and $j(\boldsymbol{k}_a, \boldsymbol{k}_b)$, respectively. Note that there are $L = \frac{K(K-1)}{2}2^{K-2}$ such pairs of vectors.

Like for the EVSDT, this yields for some $M > 0$:

$$\sqrt{\det I} \leq M\frac{1}{2}\sum_{(\boldsymbol{k}_a,\boldsymbol{k}_b)} s^{\frac{\sum_i k_a(i)-2}{2}} |c_{i(\boldsymbol{k}_a,\boldsymbol{k}_b)} - c_{j(\boldsymbol{k}_a,\boldsymbol{k}_b)}| \sqrt{\prod_i a_i^{k_a(i)} b_i^{k_b(i)}}$$
$$\leq M\frac{1}{2s}\sum_{(\boldsymbol{k}_a,\boldsymbol{k}_b)} s^{\frac{\sum_i k_a(i)}{2}} |c_{i(\boldsymbol{k}_a,\boldsymbol{k}_b)} - c_{j(\boldsymbol{k}_a,\boldsymbol{k}_b)}| e^{-\frac{1}{4}a\sum_i\left[k_a(i)\sqrt{s}\left(\frac{\mu}{2}-c_i\right)^2+k_b(i)\left(-\frac{\mu}{2}-c_i\right)^2\right]}.$$

Let

$$g_{\boldsymbol{k}_a,\boldsymbol{k}_b} = \frac{1}{s}\left(\frac{1}{2\pi}\frac{a}{2}\right)^{\frac{K+2}{2}} s^{\frac{\sum_i k_a(i)}{2}} |c_{i(\boldsymbol{k}_a,\boldsymbol{k}_b)} - c_{j(\boldsymbol{k}_a,\boldsymbol{k}_b)}| e^{-\frac{1}{4}a\sum_i\left[k_a(i)\sqrt{s}\left(\frac{\mu}{2}-c_i\right)^2+k_b(i)\left(-\frac{\mu}{2}-c_i\right)^2\right]}.$$

We now have

$$\sqrt{\det I} \leq \frac{ML}{2}(2\pi)^{\frac{K+2}{2}}\left(\frac{2}{a}\right)^{\frac{K+2}{2}} \sum_{(\boldsymbol{k}_a,\boldsymbol{k}_b)} \frac{1}{L} g_{\boldsymbol{k}_a,\boldsymbol{k}_b}.$$

We will show that $g_{\boldsymbol{k}_a,\boldsymbol{k}_b}(\boldsymbol{\theta})$ is a density and how to sample parameters from it. Like for the EVSDT, parameters can then be sampled from a mixture of distributions, in this case L distributions, one for each pair of $(\boldsymbol{k}_a, \boldsymbol{k}_b)$ with densities $g_{\boldsymbol{k}_a,\boldsymbol{k}_b}$ and mixture weights $\frac{1}{L}$. Each component density $g_{\boldsymbol{k}_a,\boldsymbol{k}_b}$ is the product of three densities, (a) the density of the marginal distribution of parameter s, $p(s)$, (b) the density of the conditional distribution of $c_j - c_i$ given s, $p(c_j - c_i|s)$, where i and j are the special values $i = i(\boldsymbol{k}_a, \boldsymbol{k}_b)$ and $j = j(\boldsymbol{k}_a, \boldsymbol{k}_b)$ with $i < j$ associated with $(\boldsymbol{k}_a, \boldsymbol{k}_b)$, and (c) the density of the conditional distribution of $\boldsymbol{\theta}^{j,s} = (c_1, c_2, \ldots, c_i, \ldots, c_{j-1}, c_{j+1}, \ldots, c_K, \mu)$ given s and $c_j - c_i$, $p(\boldsymbol{\theta}^{j,s}|c_j - c_i, s)$.

Sampling proceeded in this order. That is, we first sampled a component distribution from which to sample parameter values with equal probability for each component distribution.

Next, we sampled a value for s from the density

$$p(s) = \pi^{-1}\frac{1}{\sqrt{s}}\frac{1}{1+s}.$$

This can be done by sampling a value r from a Beta distribution with $\alpha = \beta = \frac{1}{2}$ and setting $s = \frac{1-r}{r}$.

Given that, we sampled a value of $\Delta = c_j - c_i$ from the density

$$p(\Delta|s) = \frac{1}{2}\frac{a}{2}\frac{1+s}{2}|\Delta| e^{-\frac{1}{2}\frac{a}{2}\frac{1+s}{2}\Delta^2}.$$

This can be achieved by sampling a value ξ from a χ^2 distribution with two degrees of freedom and setting $|\Delta| = \sqrt{\frac{2}{a}\frac{2}{1+s}\xi}$. The sign

$$\frac{1}{2}a\begin{pmatrix} k_a(1)+k_b(1) & 0 & \cdots & 0 & \frac{1}{2}(k_b(1)-k_a(1)) \\ 0 & k_a(2)+k_b(2) & \cdots & 0 & \frac{1}{2}(k_b(2)-k_a(2)) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & k_a(K)+k_b(K) & \frac{1}{2}(k_b(K)-k_a(K)) \\ \frac{1}{2}(k_b(1)-k_a(1)) & \frac{1}{2}(k_b(2)-k_a(2)) & \cdots & \frac{1}{2}(k_b(K)-k_a(K)) & \frac{1}{4}\sum_i k_a(i)+k_b(i) \end{pmatrix}$$

Box II.

$$\begin{pmatrix} a_1s+b_1 & 0 & \cdots & 0 & \frac{1}{2}(b_1-a_1s) & -\frac{1}{2}\left(\frac{\mu}{2}-c_1\right)a_1 \\ 0 & a_2s+b_2 & \cdots & 0 & \frac{1}{2}(b_2-a_2s) & -\frac{1}{2}\left(\frac{\mu}{2}-c_2\right)a_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_Ks+b_K & \frac{1}{2}(b_K-a_Ks) & -\frac{1}{2}\left(\frac{\mu}{2}-c_K\right)a_K \\ \frac{1}{2}(b_1-a_1s) & \frac{1}{2}(b_2-a_2s) & \cdots & \frac{1}{2}(b_K-a_Ks) & \frac{1}{4}\sum_i a_is+b_i & \frac{1}{4}\sum_i\left(\frac{\mu}{2}-c_i\right)a_i \\ -\frac{1}{2}\left(\frac{\mu}{2}-c_1\right)a_1 & -\frac{1}{2}\left(\frac{\mu}{2}-c_2\right)a_2 & \cdots & -\frac{1}{2}\left(\frac{\mu}{2}-c_K\right)a_K & \frac{1}{4}\sum_i\left(\frac{\mu}{2}-c_i\right)a_i & \frac{1}{4s}\sum_i\left(\frac{\mu}{2}-c_i\right)^2a_i \end{pmatrix}.$$

Box III.

$$\frac{1}{2}a\begin{pmatrix} 2s+2 & 0 & \cdots & 0 & 1-s \\ 0 & k_a(3)s+k_b(3) & \cdots & 0 & \frac{1}{2}(k_b(3)-k_a(3)s) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & k_a(K)s+k_b(K) & \frac{1}{2}(k_b(K)-k_a(K)s) \\ 1-s & \frac{1}{2}(k_b(3)-k_a(3)s) & \cdots & \frac{1}{2}(k_b(K)-k_a(K)s) & \frac{1}{4}\sum_{l=1}^{K}k_a(l)s+k_b(l) \end{pmatrix}.$$

Box IV.

(+ or −) of Δ is then chosen randomly with equal probability for both possibilities.

Finally, given s and $c_j - c_i$, we sampled $\boldsymbol{\theta}^{j,s}$ from a multivariate normal distribution. Without loss of generality assume that $i = 1$ and $j = 2$. Its mean is $\boldsymbol{v} = (-\frac{1}{2}(c_2 - c_1), 0, \ldots, 0)'$, that is, it is a vector of zeros, except for the i-th entry, corresponding to c_i, for which the mean is $-\frac{1}{2}(c_j - c_i)$. Its K by K covariance matrix can be defined by its inverse Σ^{-1} as in Box IV.

Note that there are no columns and rows corresponding to c_2 in this matrix and that the rightmost diagonal element nevertheless sums over all $l = 1, \ldots, K$ and is equal to $[2s + 2 + \sum_{l=3}^{K}(k_a(l)s + k_b(l))]/4$ (because $k_a(1) = k_b(1) = k_a(2) = k_b(2) = 1$). By the structural analogy with the Fisher information matrix of the EVSDT it follows that the determinant of Σ^{-1} is

$$\left(\frac{a}{2}\right)^K [(2s)2]s^{\sum_{l=3}^{K}k_a(l)} = \frac{4}{s}\left(\frac{a}{2}\right)^K s^{\sum_{l=1}^{K}k_a(l)}.$$

Hence, the appropriate density is:

$$p(\boldsymbol{\theta}^{j,s}|c_j - c_i, s) = \frac{2}{\sqrt{s}}(2\pi)^{-\frac{K}{2}} \times \left(\frac{a}{2}\right)^{\frac{K}{2}} s^{\frac{\sum_{l=1}^{K}k_a(l)}{2}} e^{-\frac{1}{2}(\boldsymbol{\theta}^{j,s}-\boldsymbol{v})'\Sigma^{-1}(\boldsymbol{\theta}^{j,s}-\boldsymbol{v})}.$$

The value of c_j is then computed as $c_j = c_i + \Delta$. Note that this involves a variable transformation,

$$\begin{pmatrix} c_i \\ \Delta \end{pmatrix} \to \begin{pmatrix} c_i \\ c_i + \Delta \end{pmatrix},$$

but the (absolute value) of the determinant of the Jacobian matrix of this transformation is one. Hence, the density of $(\boldsymbol{\theta}^{j,s}, \Delta, s)$ is the same as that of $\boldsymbol{\theta}$ when $\Delta = c_j - c_i$. This density is the product of the three densities, $p(s)$, $p(\Delta|s) = p(c_j - c_i|s)$, and $p(\boldsymbol{\theta}^{j,s}|c_j - c_i, s)$. It is left for the reader to verify that $g_{\boldsymbol{k}_a,\boldsymbol{k}_b}(\boldsymbol{\theta}) = p(s)p(c_j - c_i|s)p(\boldsymbol{\theta}^{j,s}|c_j - c_i, s)$. Again, the integration is over positive and negative values of μ, and we therefore rejected samples with $\mu < 0$ and multiplied g by 2 in correction. To impose the inequality restriction $\sigma_s \geq 1$ in UVSDT($\sigma_s \geq \sigma_n$), note that $\sigma_s \geq 1$ is equivalent to $s \leq 1$, which in turn is equivalent to $r \geq 0.5$ in the above sampling scheme for s. We therefore replaced sampled values of r with $r < 0.5$ by $1 - r$ and multiplied g (a second time) by 2 in correction.

DPSDT

Using parameters $\boldsymbol{\theta} = (c_1, c_2, \ldots, c_K, \mu, R)$, it is convenient to parameterize the DPSDT as follows:

$$p_{s,i} = R + (1 - R)F\left(\frac{\mu}{2} - c_i\right)$$

$$p_{n,i} = F\left(-\frac{\mu}{2} - c_i\right).$$

The partial derivatives are zero except for

$$\frac{\partial}{\partial c_i}p_{s,i} = -(1 - R)f\left(\frac{\mu}{2} - c_i\right),$$

$$\frac{\partial}{\partial \mu}p_{s,i} = \frac{1}{2}(1 - R)f\left(\frac{\mu}{2} - c_i\right),$$

$$\frac{\partial}{\partial R}p_{s,i} = 1 - F\left(\frac{\mu}{2} - c_i\right),$$

$$\frac{\partial}{\partial c_i}p_{n,i} = -f\left(-\frac{\mu}{2} - c_i\right), \quad \text{and}$$

$$\frac{\partial}{\partial \mu}p_{n,i} = -\frac{1}{2}f\left(-\frac{\mu}{2} - c_i\right).$$

Let

$$a_i = \frac{(1 - R)f\left(\frac{\mu}{2} - c_i\right)}{(p_{s,i}(1 - p_{s,i}))^{\frac{1}{2}}}\sqrt{q_{s,i}},$$

$$A_i = \frac{1 - F\left(\frac{\mu}{2} - c_i\right)}{(p_{s,i}(1-p_{s,i}))^{\frac{1}{2}}}\sqrt{q_{s,i}}, \quad \text{and}$$

$$b_i = \frac{f\left(-\frac{\mu}{2} - c_i\right)}{(p_{n,i}(1-p_{n,i}))^{\frac{1}{2}}}\sqrt{q_{n,i}}.$$

It follows that the Fisher information matrix I is as given in Box V.

Its determinant is

$$\det I = \sum_{i<j} b_i^2 b_j^2 \,(A_i a_j - A_j a_i)^2 \prod_{l \neq i,j} (a_l^2 + b_l^2).$$

To obtain a dominating density for sampling parameters in the Monte Carlo integration, let $f_i = f(\frac{\mu}{2} - c_i)$ and $F_i = F(\frac{\mu}{2} - c_i)$. Note that (because $R + (1-R)F_i \geq (1-R)F_i$)

$$\begin{aligned}\frac{a_i^2}{q_{s,i}} &= (1-R)^2 \frac{f_i^2}{[R+(1-R)F_i](1-R)(1-F_i)} \\ &= (1-R)\frac{f_i^2}{[R+(1-R)F_i](1-F_i)} \\ &\leq (1-R)\frac{f_i^2}{[(1-R)F_i](1-F_i)} = \frac{f_i^2}{F_i(1-F_i)}\end{aligned}$$

and that (because $R + (1-R)F_j \geq F_j$)

$$\begin{aligned}\frac{(A_i a_j)^2}{q_{s,i}} &= \frac{(1-F_i)^2}{[R+(1-R)F_i](1-R)(1-F_i)}(1-R)^2 \\ &\quad \times \frac{f_j^2}{[R+(1-R)F_j](1-R)(1-F_j)} \\ &= \frac{(1-F_i)^2}{[R+(1-R)F_i](1-F_i)} \frac{f_j^2}{[R+(1-R)F_j](1-F_j)} \\ &\leq \frac{(1-F_i)^2}{[R+(1-R)F_i](1-F_i)} \frac{f_j^2}{F_j(1-F_j)} \\ &= \frac{1}{R}\frac{(1-F_i)}{1+\frac{1-R}{R}F_i}\frac{f_j^2}{F_j(1-F_j)} \\ &\leq \frac{1}{R}(1-F_i)\frac{f_j^2}{F_j(1-F_j)} \leq \frac{1}{R}\frac{f_j^2}{F_j(1-F_j)}.\end{aligned}$$

Using the same pairs of vectors $(\boldsymbol{k}_a, \boldsymbol{k}_b)$ with $\boldsymbol{k}_a = (k_a(1), \ldots, k_a(K))$ and $\boldsymbol{k}_b = (k_b(1), \ldots, k_b(K))$ as for the UVSDT, it follows for some $M > 0$ that

$$\sqrt{\det I} \leq M\frac{1}{\sqrt{R}} \sum_{(\boldsymbol{k}_a, \boldsymbol{k}_b)} \left(e^{\frac{1}{4}a\left(\frac{\mu}{2} - c_{i(\boldsymbol{k}_a,\boldsymbol{k}_b)}\right)^2} + e^{\frac{1}{4}a\left(\frac{\mu}{2} - c_{j(\boldsymbol{k}_a,\boldsymbol{k}_b)}\right)^2} \right) \times e^{-\frac{1}{4}a\sum_i \left[k_a(i)\left(\frac{\mu}{2}-c_i\right)^2 + k_b(i)\left(-\frac{\mu}{2}-c_i\right)^2\right]}.$$

To see this note that $|A_i a_j - A_j a_i| \leq A_i a_j + A_j a_i$ and that $A_i a_j \leq M' \frac{1}{\sqrt{R}} \exp(-\frac{1}{4}a(\frac{\mu}{2} - c_j)^2)$ and $A_j a_i \leq M'' \frac{1}{\sqrt{R}} \exp(-\frac{1}{4}a(\frac{\mu}{2} - c_i)^2)$. The sum in the exponential function in the above equation sums over both of these exponential terms simultaneously (i.e., for both special values i and j, $i(\boldsymbol{k}_a, \boldsymbol{k}_b)$ and $j(\boldsymbol{k}_a, \boldsymbol{k}_b)$), so that we need to multiply by either $e^{\frac{1}{4}a\left(\frac{\mu}{2} - c_{i(\boldsymbol{k}_a,\boldsymbol{k}_b)}\right)^2}$ or $e^{\frac{1}{4}a\left(\frac{\mu}{2} - c_{j(\boldsymbol{k}_a,\boldsymbol{k}_b)}\right)^2}$ to pick out only one of it.

Let

$$g_{1,\boldsymbol{k}_a,\boldsymbol{k}_b} = \frac{1}{\sqrt{R}}\left(\frac{1}{2\pi}\frac{a}{2}\right)^{\frac{K+1}{2}} e^{\frac{1}{4}a\left(\frac{\mu}{2} - c_{i(\boldsymbol{k}_a,\boldsymbol{k}_b)}\right)^2} \times e^{-\frac{1}{4}a\sum_i \left[k_a(i)\left(\frac{\mu}{2}-c_i\right)^2 + k_b(i)\left(-\frac{\mu}{2}-c_i\right)^2\right]} \quad \text{and}$$

$$g_{2,\boldsymbol{k}_a,\boldsymbol{k}_b} = \frac{1}{\sqrt{R}}\left(\frac{1}{2\pi}\frac{a}{2}\right)^{\frac{K+1}{2}} e^{\frac{1}{4}a\left(\frac{\mu}{2} - c_{j(\boldsymbol{k}_a,\boldsymbol{k}_b)}\right)^2} \times e^{-\frac{1}{4}a\sum_i \left[k_a(i)\left(\frac{\mu}{2}-c_i\right)^2 + k_b(i)\left(-\frac{\mu}{2}-c_i\right)^2\right]}.$$

Setting $L = \frac{K(K-1)}{2}2^{K-1}$ (the number of pairs $(\boldsymbol{k}_a, \boldsymbol{k}_b)$ multiplied by two for the two functions g_1 and g_2), we then have

$$\sqrt{\det I} \leq ML(2\pi)^{\frac{K+1}{2}}\left(\frac{2}{a}\right)^{\frac{K+1}{2}} \sum_{(\boldsymbol{k}_a,\boldsymbol{k}_b)}\sum_{t=1}^{2}\frac{1}{L} g_{t,\boldsymbol{k}_a,\boldsymbol{k}_b}.$$

We will show that $g_{t,\boldsymbol{k}_a,\boldsymbol{k}_b}(\boldsymbol{\theta})$ is a density and how to sample parameters from it. Like for the UVSDT, parameters can then be sampled from a mixture of distributions, in this case L distributions, one for each pair $(\boldsymbol{k}_a, \boldsymbol{k}_b)$ and $t = 1, 2$, with densities $g_{t,\boldsymbol{k}_a,\boldsymbol{k}_b}$ and mixture weights $\frac{1}{L}$. Each component density $g_{t,\boldsymbol{k}_a,\boldsymbol{k}_b}$ is the product of three densities, the marginal density of the distribution of parameter R, $p(R)$, the density of the distribution of parameter μ, $p(\mu)$, and the density of the conditional distribution of $\boldsymbol{c} = (c_1, \ldots, c_K)$ given μ, $p_t(\boldsymbol{c}|\mu)$.

The density $p(R)$ is simply $\frac{1}{2\sqrt{R}}$. It is the density of a Beta distribution with $\alpha = \frac{1}{2}$ and $\beta = 1$. The density $p(\mu)$ is that of a truncated normal with mean 0 and variance $\frac{4}{a}$:

$$p(\mu) = \frac{2}{\sqrt{2\pi}}\sqrt{\frac{a}{4}} e^{-\frac{1}{2}\left(\frac{\mu}{\sqrt{\frac{4}{a}}}\right)^2} 1_{\{\mu \geq 0\}}.$$

The density of the conditional distribution of $\boldsymbol{c} = (c_1, \ldots, c_K)$ given μ, $p_t(\boldsymbol{c}|\mu)$, is a multivariate normal distribution with mean $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)$ and $\nu_i = (k_a(i) - k_b(i))\frac{\mu}{2}$ for $i \neq i(\boldsymbol{k}_a, \boldsymbol{k}_b)$, $j(\boldsymbol{k}_a, \boldsymbol{k}_b)$, and $\nu_{i(\boldsymbol{k}_a,\boldsymbol{k}_b)} = -\frac{\mu}{2}$ and $\nu_{j(\boldsymbol{k}_a,\boldsymbol{k}_b)} = 0$ for $t = 1$, and $\nu_{i(\boldsymbol{k}_a,\boldsymbol{k}_b)} = 0$ and $\nu_{j(\boldsymbol{k}_a,\boldsymbol{k}_b)} = -\frac{\mu}{2}$ for $t = 2$. Its covariance matrix is defined by its inverse Σ^{-1}. The inverse is a diagonal matrix with diagonal elements $\frac{a}{2}(k_a(i) + k_b(i)) = \frac{a}{2}$ for $i \neq i(\boldsymbol{k}_a, \boldsymbol{k}_b), j(\boldsymbol{k}_a, \boldsymbol{k}_b)$ and $\frac{a}{2}$ and $2\frac{a}{2}$ for $i = i(\boldsymbol{k}_a, \boldsymbol{k}_b)$ and $i = j(\boldsymbol{k}_a, \boldsymbol{k}_b)$, respectively, for $t = 1$, and $2\frac{a}{2}$ and $\frac{a}{2}$ for $i = i(\boldsymbol{k}_a, \boldsymbol{k}_b)$ and $i = j(\boldsymbol{k}_a, \boldsymbol{k}_b)$, respectively, for $t = 2$. It follows that

$$p_t(\boldsymbol{c}|\mu) = \sqrt{2}(2\pi)^{-\frac{K}{2}}\left(\frac{a}{2}\right)^{\frac{K}{2}} e^{-\frac{1}{2}(\boldsymbol{c}-\boldsymbol{\nu})'\Sigma^{-1}(\boldsymbol{c}-\boldsymbol{\nu})}.$$

It is left for the reader to verify that $g_{t,\boldsymbol{k}_a,\boldsymbol{k}_b}(\boldsymbol{\theta}) = p(R)p(\mu)p_t(\boldsymbol{c}|\mu)$.

MSDT0

The MSDT with $\mu^* = 0$ has parameters $\boldsymbol{\theta} = (c_1, c_2, \ldots, c_K, \mu, \lambda)$. It is convenient to parameterize it as follows:

$$\begin{aligned}p_{s,i} &= \lambda F\left(\frac{\mu}{2} - c_i\right) + (1-\lambda)F\left(-\frac{\mu}{2} - c_i\right) \\ p_{n,i} &= F\left(-\frac{\mu}{2} - c_i\right).\end{aligned}$$

The partial derivatives are zero except for

$$\begin{aligned}\frac{\partial}{\partial c_i}p_{s,i} &= -\lambda f\left(\frac{\mu}{2} - c_i\right) - (1-\lambda)\lambda f\left(-\frac{\mu}{2} - c_i\right), \\ \frac{\partial}{\partial \mu}p_{s,i} &= \frac{1}{2}\lambda f\left(\frac{\mu}{2} - c_i\right) - \frac{1}{2}(1-\lambda)f\left(-\frac{\mu}{2} - c_i\right), \\ \frac{\partial}{\partial \lambda}p_{s,i} &= F\left(\frac{\mu}{2} - c_i\right) - F\left(-\frac{\mu}{2} - c_i\right), \\ \frac{\partial}{\partial c_i}p_{n,i} &= -f\left(-\frac{\mu}{2} - c_i\right), \quad \text{and} \\ \frac{\partial}{\partial \mu}p_{n,i} &= -\frac{1}{2}f\left(-\frac{\mu}{2} - c_i\right).\end{aligned}$$

Let

$$a_i = \frac{\lambda f\left(\frac{\mu}{2} - c_i\right)}{(p_{s,i}(1-p_{s,i}))^{\frac{1}{2}}}\sqrt{q_{s,i}}$$

$$b_i = \frac{(1-\lambda)f\left(-\frac{\mu}{2} - c_i\right)}{(p_{s,i}(1-p_{s,i}))^{\frac{1}{2}}}\sqrt{q_{s,i}}$$

$$\begin{pmatrix} a_1^2+b_1^2 & 0 & \cdots & 0 & \frac{1}{2}(b_1^2-a_1^2) & -A_1a_1 \\ 0 & a_2^2+b_2^2 & \cdots & 0 & \frac{1}{2}(b_2^2-a_2^2) & -A_2a_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & a_K^2+b_K^2 & \frac{1}{2}(b_K^2-a_K^2) & -A_Ka_K \\ \frac{1}{2}(b_1^2-a_1^2) & \frac{1}{2}(b_2^2-a_2^2) & \cdots & \frac{1}{2}(b_K^2-a_K^2) & \frac{1}{4}\sum_i a_i^2+b_i^2 & \frac{1}{2}\sum_i A_ia_i \\ -A_1a_1 & -A_2a_2 & \cdots & -A_Ka_K & \frac{1}{2}\sum_i A_ia_i & \sum_i A_i^2 \end{pmatrix}.$$

Box V.

$$A_i = \frac{F\left(\frac{\mu}{2}-c_i\right)-F\left(-\frac{\mu}{2}-c_i\right)}{(p_{s,i}(1-p_{s,i}))^{\frac{1}{2}}}\sqrt{q_{s,i}}$$

$$d_i = \frac{f\left(-\frac{\mu}{2}-c_i\right)}{(p_{n,i}(1-p_{n,i}))^{\frac{1}{2}}}\sqrt{q_{n,i}}$$

It follows that the Fisher information matrix I is as given in Box VI.

Its determinant is

$$\det I = \sum_{i<j} d_i^2 d_j^2 (A_ia_j - A_ja_i)^2 \prod_{l\neq i,j}((a_l+b_l)^2+d_l^2).$$

To obtain a dominating density, let f_i, F_i, f_i^*, F_i^* be, in order, $f(\frac{\mu}{2}-c_i)$, $F(\frac{\mu}{2}-c_i)$, $f(-\frac{\mu}{2}-c_i)$, and $F(-\frac{\mu}{2}-c_i)$ and note that

$$\begin{aligned}\frac{a_i^2}{q_{s,i}} &= \frac{\lambda^2 f_i^2}{(\lambda F_i+(1-\lambda)F_i^*)(\lambda(1-F_i)+(1-\lambda)(1-F_i^*))} \\ &= \frac{f_i^2}{\left(F_i+\frac{1-\lambda}{\lambda}F_i^*\right)\left(1-F_i+\frac{1-\lambda}{\lambda}(1-F_i^*)\right)} \\ &\leq \frac{f_i^2}{F_i(1-F_i)}.\end{aligned}$$

Similarly, $\frac{b_i^2}{q_{s,i}} \leq \frac{f_i^*}{F_i^*(1-F_i^*)}$. Furthermore,

$$\begin{aligned}\frac{A_i^2}{q_{s,i}} &= \frac{(F_i-F_i^*)^2}{(\lambda F_i+(1-\lambda)F_i^*)(\lambda(1-F_i)+(1-\lambda)(1-F_i^*))} \\ &= \frac{1}{\lambda(1-\lambda)}\frac{(F_i-F_i^*)^2}{\left(F_i+\frac{1-\lambda}{\lambda}F_i^*\right)\left(\frac{\lambda}{1-\lambda}(1-F_i)+1-F_i^*\right)} \\ &\leq \frac{1}{\lambda(1-\lambda)}\frac{(F_i-F_i^*)^2}{F_i(1-F_i^*)} \\ &\leq \frac{1}{\lambda(1-\lambda)}\frac{(F_i-F_i^*)^2}{F_i-F_i^*} \leq \frac{1}{\lambda(1-\lambda)},\end{aligned}$$

taking into account that $1 > F_i > F_i^*$ for $\mu > 0$.

Putting this together and using the same pairs of vectors $(\boldsymbol{k}_a, \boldsymbol{k}_b)$ as for the UVSDT and the DPSDT, it follows for some $M > 0$ that

$$\begin{aligned}\sqrt{\det I} \leq\ & M\frac{1}{\sqrt{\lambda(1-\lambda)}} \\ & \times \sum_{(\boldsymbol{k}_a,\boldsymbol{k}_b)}\left(e^{\frac{1}{4}a\left(\frac{\mu}{2}-c_{i(\boldsymbol{k}_a,\boldsymbol{k}_b)}\right)^2}+e^{\frac{1}{4}a\left(\frac{\mu}{2}-c_{j(\boldsymbol{k}_a,\boldsymbol{k}_b)}\right)^2}\right) \\ & \times 2^{\sum_{l\neq i(\boldsymbol{k}_a,\boldsymbol{k}_b),j(\boldsymbol{k}_a,\boldsymbol{k}_b)} k_b(l)}\, e^{-\frac{1}{4}a\sum_i\left[k_a(i)\left(\frac{\mu}{2}-c_i\right)^2+k_b(i)\left(-\frac{\mu}{2}-c_i\right)^2\right]}.\end{aligned}$$

To see this note that $\sqrt{(a_l+b_l)^2+d_l^2} \leq a_l+b_l+d_l$ and that $b_l^2/q_{s,l}$ and $d_l^2/q_{n,l}$ share the same upper bound $\frac{f_l^*}{F_l^*(1-F_l^*)}$. Hence b_l and d_l themselves share a common upper bound Q_l (i.e., for some $\bar{M} > 0$, $Q_l \leq \bar{M}e^{-\frac{1}{4}a(-\frac{\mu}{2}-c_l)^2}$), so that $a_l+b_l+d_l \leq a_l+2Q_l$.

Let

$$\begin{aligned}g_{1,\boldsymbol{k}_a,\boldsymbol{k}_b} =\ & \frac{2}{\pi}\frac{1}{\sqrt{\lambda(1-\lambda)}}\left(\frac{1}{2\pi}\frac{a}{2}\right)^{\frac{K+1}{2}} e^{\frac{1}{4}a\left(\frac{\mu}{2}-c_{i(\boldsymbol{k}_a,\boldsymbol{k}_b)}\right)^2} \\ & \times e^{-\frac{1}{4}a\sum_i\left[k_a(i)\left(\frac{\mu}{2}-c_i\right)^2+k_b(i)\left(-\frac{\mu}{2}-c_i\right)^2\right]} \quad \text{and}\end{aligned}$$

$$\begin{aligned}g_{2,\boldsymbol{k}_a,\boldsymbol{k}_b} =\ & \frac{2}{\pi}\frac{1}{\sqrt{\lambda(1-\lambda)}}\left(\frac{1}{2\pi}\frac{a}{2}\right)^{\frac{K+1}{2}} \\ & \times e^{\frac{1}{4}a\left(\frac{\mu}{2}-c_{j(\boldsymbol{k}_a,\boldsymbol{k}_b)}\right)^2} e^{-\frac{1}{4}a\sum_i\left[k_a(i)\left(\frac{\mu}{2}-c_i\right)^2+k_b(i)\left(-\frac{\mu}{2}-c_i\right)^2\right]}.\end{aligned}$$

We then have

$$\begin{aligned}\sqrt{\det I} \leq\ & M\left(2\frac{K(K-1)}{2}3^{K-2}\right)\frac{\pi}{2}(2\pi)^{\frac{K+1}{2}} \\ & \times\left(\frac{2}{a}\right)^{\frac{K+1}{2}}\sum_{(\boldsymbol{k}_a,\boldsymbol{k}_b)}\sum_{t=1}^{2}\omega_{\boldsymbol{k}_a,\boldsymbol{k}_b}\, g_{t,\boldsymbol{k}_a,\boldsymbol{k}_b}\end{aligned}$$

with mixture weights

$\omega_{\boldsymbol{k}_a,\boldsymbol{k}_b} = (2^{\frac{K(K-1)}{2}})^{-1}(\frac{1}{3})^{\sum_{l\neq i(\boldsymbol{k}_a,\boldsymbol{k}_b),j(\boldsymbol{k}_a,\boldsymbol{k}_b)} k_a(l)}(\frac{2}{3})^{\sum_{l\neq i(\boldsymbol{k}_a,\boldsymbol{k}_b),j(\boldsymbol{k}_a,\boldsymbol{k}_b)} k_b(l)}$.

Like for the DPSDT, each component density $g_{t,\boldsymbol{k}_a,\boldsymbol{k}_b}$ is the product of three densities, the marginal density of the distribution of parameter λ, $p(\lambda)$, the density of the distribution of parameter μ, $p(\mu)$, and the density of the conditional distribution of $\boldsymbol{c} = (c_1, \ldots, c_K)$ given μ, $p_t(\boldsymbol{c}|\mu)$. The densities for μ and $\boldsymbol{c}$ are exactly the same as for DPSDT. The marginal density of λ is a Beta distribution with $\alpha = \beta = \frac{1}{2}$, hence $p(\lambda) = \frac{1}{\pi\sqrt{\lambda(1-\lambda)}}$.

MSDT

Using parameters $\boldsymbol{\theta} = (c_1, c_2, \ldots, c_K, \mu, \nu, \lambda)$, $\mu \geq 0$, $\nu \geq 0$, it is convenient to parameterize the MSDT as follows:

$$p_{s,i} = \lambda F(\mu-c_i)+(1-\lambda)F(-c_i)$$

$$p_{n,i} = F(-\nu-c_i).$$

The partial derivatives are zero except for

$$\frac{\partial}{\partial c_i}p_{s,i} = -\lambda f(\mu-c_i)-(1-\lambda)\lambda f(-c_i),$$

$$\frac{\partial}{\partial \mu}p_{s,i} = \lambda f(\mu-c_i),$$

$$\frac{\partial}{\partial \lambda}p_{s,i} = F(\mu-c_i)-F(-c_i),$$

$$\frac{\partial}{\partial c_i}p_{n,i} = -f(-\nu-c_i), \quad \text{and}$$

$$\frac{\partial}{\partial \nu}p_{n,i} = -f(-\nu-c_i).$$

Let

$$a_i = \frac{\lambda f(\mu-c_i)}{(p_{s,i}(1-p_{s,i}))^{\frac{1}{2}}}\sqrt{q_{s,i}},$$

$$b_i = \frac{(1-\lambda)f(-c_i)}{(p_{s,i}(1-p_{s,i}))^{\frac{1}{2}}}\sqrt{q_{s,i}},$$

$$\begin{pmatrix} (a_1+b_1)^2+d_1^2 & 0 & \cdots & 0 & \frac{1}{2}(b_1^2-a_1^2+d_1^2) & -A_1(a_1+b_1) \\ 0 & (a_2+b_2)^2+d_2^2 & \cdots & 0 & \frac{1}{2}(b_2^2-a_2^2+d_2^2) & -A_2(a_2+b_2) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & (a_K+b_K)^2+d_K^2 & \frac{1}{2}(b_K^2-a_K^2+d_K^2) & -A_k(a_K+b_K) \\ \frac{1}{2}(b_1^2-a_1^2+d_1^2) & \frac{1}{2}(b_2^2-a_2^2+d_2^2) & \cdots & \frac{1}{2}(b_K^2-a_K^2+d_K^2) & \frac{1}{4}\sum_i[(b_i-a_i)^2+d_i^2] & \frac{1}{2}\sum_i A_i(a_i-b_i) \\ -A_1(a_1+b_1) & -A_2(a_2+b_2) & \cdots & -A_k(a_K+b_K) & \frac{1}{2}\sum_i A_i(a_i-b_i) & \sum_i A_i^2 \end{pmatrix}.$$

Box VI.

$$A_i = \frac{F(\mu-c_i)-F(-c_i)}{(p_{s,i}(1-p_{s,i}))^{\frac{1}{2}}}\sqrt{q_{s,i}}, \quad \text{and}$$
$$d_i = \frac{f(-\nu-c_i)}{(p_{n,i}(1-p_{n,i}))^{\frac{1}{2}}}\sqrt{q_{n,i}}.$$

It follows that the Fisher information matrix I is as given in Box VII.

Its determinant is

$$\det I = \sum_{i<j<l} d_i^2 d_j^2 d_l^2 (A_i a_j b_l - A_j a_i b_l + A_j a_l b_i - A_l a_j b_i + A_l a_i b_j - A_i a_l b_j)^2 \prod_{m\neq i,j,l}((a_m+b_m)^2+d_m^2).$$

Like for the MSDT with $\mu^* = 0$ it can be shown that

$$\frac{a_i^2}{q_{s,i}} \leq \frac{f^2(\mu-c_i)}{F(\mu-c_i)(1-F(\mu-c_i))},$$
$$\frac{b_i^2}{q_{s,i}} \leq \frac{f^2(-c_i)}{F(-c_i)(1-F(-c_i))}, \quad \text{and}$$
$$\frac{A_i^2}{q_{s,i}} \leq \frac{1}{\lambda(1-\lambda)}.$$

For a triple of indices $i < j < m$ define triples of vectors, $(\boldsymbol{k}_a, \boldsymbol{k}_b, \boldsymbol{k}_d)$ with $\boldsymbol{k}_a = (k_a(1), \ldots, k_a(K))$, $\boldsymbol{k}_b = (k_b(1), \ldots, k_b(K))$, and $\boldsymbol{k}_d = (k_d(1), \ldots, k_d(K))$, containing zeros and ones, with $k_d(i) = k_d(j) = k_d(l) = 1$ (for the product terms $d_i^2 d_j^2 d_l^2$ in the determinant), $k_a(j)k_b(l) + k_a(i)k_b(l) + k_a(l)k_b(i) + k_a(j)k_b(i) + k_a(i)k_b(j) + k_a(l)k_b(j) = 1$ (for picking out one of the products $a_p b_q$ in the determinant), and $k_a(m) + k_b(m) + k_d(m) = 1$ for all $m \neq i, j, k$. Given $i < j < l$, let the set of such triples of vectors be $\mathcal{K}_{i,j,l}$. Note that $\mathcal{K}_{i,j,l}$ contains 63^{K-3} such triples. It follows that

$$\sqrt{\det I} \leq M \frac{1}{\sqrt{\lambda(1-\lambda)}} \sum_{i<j<l} \sum_{(\boldsymbol{k}_a,\boldsymbol{k}_b,\boldsymbol{k}_d)\in\mathcal{K}_{i,j,l}} e^{-\frac{1}{4}a\sum_i [k_a(i)(\mu-c_i)^2 + k_b(i)(-c_i)^2 + k_d(i)(-\nu-c_i)^2]}.$$

The exponential term to the right is proportional to the density of a multivariate normal distribution for $(c_1, \ldots, c_K, \mu, \nu)$ with mean vector zero and covariance matrix $\Sigma = E^{-1}$ defined by its inverse E as in Box VIII (ignoring a factor of $\frac{a}{2}$).

We need the marginal distribution of (μ, ν) from this distribution. It is a bivariate normal distribution with mean vector zero and covariance matrix $\Sigma_{(\mu,\nu)}$ defined by partitioning Σ as follows:

$$\Sigma = \begin{pmatrix} \Sigma_{\boldsymbol{c}} & \Sigma_{\boldsymbol{c},(\mu,\nu)} \\ \Sigma_{(\mu,\nu),\boldsymbol{c}} & \Sigma_{(\mu,\nu)} \end{pmatrix}.$$

That is, we need the elements of the inverse of E in the cells corresponding to μ and ν. These can be obtained from the determinant of E and the so-called cofactors of E corresponding to the cells in question. The cofactors themselves are (plus or minus one times) determinants of E with rows and columns $K + 1$ or $K+2$ deleted. All of these determinants can be computed using the method exemplified for the determinant of the Fisher information matrix of the 1HTM. This yields:

$$\Sigma_{(\mu,\nu)} = \frac{2}{a}\begin{pmatrix} 4 & -2 \\ -2 & 2 \end{pmatrix}.$$

On the other hand, the conditional distribution of $\boldsymbol{c} = (c_1, \ldots, c_K)$ given (μ, ν) is also multivariate normal with mean $\boldsymbol{\alpha}$ defined by regressing $\boldsymbol{c}$ on (μ, ν)

$$\boldsymbol{\alpha} = \Sigma_{\boldsymbol{c},(\mu,\nu)} \Sigma_{(\mu,\nu)}^{-1} \begin{pmatrix} \mu \\ \nu \end{pmatrix}$$

and a diagonal covariance matrix Γ with i-th diagonal element equal to $\frac{2}{a}(k_a(i) + k_b(i) + k_d(i))^{-1}$ (because the covariance matrix of the conditional distribution is the inverse of E after deleting columns and rows corresponding to μ and ν; (Kotz, Balakrishnan, & Johnson, 2000, chap. 45)). Note that the determinant of $\Sigma_{(\mu,\nu)}$ is $4\left(\frac{2}{a}\right)^2$, whereas that of Γ is $\frac{1}{4}\left(\frac{2}{a}\right)^K$ (because $k_a(i) + k_b(i) + k_d(i) = 1$ for all i except two, for which the sum is 2, by definition of $(\boldsymbol{k}_a, \boldsymbol{k}_b, \boldsymbol{k}_d)$).

Let $f_{(\boldsymbol{0},\Sigma_{(\mu,\nu)})}$ and $f_{(\boldsymbol{\alpha},\Gamma)}$ be the densities of these multivariate normal distributions of (μ, ν) and $\boldsymbol{c}$ given (μ, ν), respectively. Furthermore, let $p(\lambda)$ be the density of a Beta distribution with $\alpha = \beta = \frac{1}{2}$.

We want to sample values μ and ν greater than zero. Therefore, these parameters were sampled from the above bivariate normal distribution for μ and ν truncated so that $\mu > 0$ and $\nu > 0$ (via rejection sampling). The density of the truncated normal distribution is

$$p(\mu, \nu) = b^{-1} f_{(\boldsymbol{0},\Sigma_{(\mu,\nu)})} 1_{\{\mu>0 \text{ and } \nu>0\}},$$

where b is the probability that two normal random variables with variance one, mean zero, and correlation $-\frac{1}{\sqrt{2}}$ are both greater than zero (the correlation of μ and ν is $-\frac{1}{\sqrt{2}}$ according to $\Sigma_{(\mu,\nu)}$, whereas the variances of the involved normal variates does not affect the probability that both are greater than zero).

Setting $g_{(\boldsymbol{k}_a,\boldsymbol{k}_b,\boldsymbol{k}_d)} = p(\lambda)p(\mu, \nu)f_{(\boldsymbol{\alpha},\Gamma)}$, it follows that

$$\sqrt{\det I} \leq M\left(63^{K-3}\frac{K(K-1)(K-2)}{6}\right)\pi(2\pi)^{\frac{K+2}{2}}\left(\frac{2}{a}\right)^{\frac{K+2}{2}} b \times \sum_{i<j<l}\sum_{(\boldsymbol{k}_a,\boldsymbol{k}_b,\boldsymbol{k}_d)\in\mathcal{K}_{i,j,l}} \omega_{(\boldsymbol{k}_a,\boldsymbol{k}_b,\boldsymbol{k}_d)} g_{(\boldsymbol{k}_a,\boldsymbol{k}_b,\boldsymbol{k}_d)},$$

where

$$\omega_{(\boldsymbol{k}_a,\boldsymbol{k}_b,\boldsymbol{k}_d)} = \frac{6}{K(K-1)(K-2)}\frac{1}{6}\left(\frac{1}{3}\right)^{K-3}.$$

Again, $\sum_{i<j<l}\sum_{(\boldsymbol{k}_a,\boldsymbol{k}_b,\boldsymbol{k}_d)\in\mathcal{K}_{i,j,l}} \omega_{(\boldsymbol{k}_a,\boldsymbol{k}_b,\boldsymbol{k}_d)} g_{(\boldsymbol{k}_a,\boldsymbol{k}_b,\boldsymbol{k}_d)}$ is the density of a mixture of distributions that is relatively easy to sample from.

$$\begin{pmatrix} (a_1+b_1)^2+d_1^2 & 0 & \cdots & 0 & -a_1(a_1+b_1) & d_1^2 & -A_1(a_1+b_1) \\ 0 & (a_2+b_2)^2+d_2^2 & \cdots & 0 & -a_2(a_2+b_2) & d_2^2 & -A_2(a_2+b_2) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & (a_K+b_K)^2+d_K^2 & -a_K(a_K+b_K) & d_K^2 & -A_k(a_K+b_K) \\ -a_1(a_1+b_1) & -a_2(a_2+b_2) & \cdots & -a_K(a_K+b_K) & \sum_i a_i^2 & 0 & \sum_i A_i a_i \\ d_1^2 & d_2^2 & \cdots & d_K^2 & 0 & \sum_i d_i^2 & 0 \\ -A_1(a_1+b_1) & -A_2(a_2+b_2) & \cdots & -A_k(a_K+b_K) & \sum_i A_i a_i & 0 & \sum_i A_i^2 \end{pmatrix}.$$

Box VII.

$$\begin{pmatrix} k_a(1)+k_b(1)+k_d(1) & 0 & \cdots & 0 & -k_a(1) & k_d(1) \\ 0 & k_a(2)+k_b(2)+k_d(2) & \cdots & 0 & -k_a(2) & k_d(2) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \ldots & k_a(K)+k_b(K)+k_d(K) & -k_a(K) & k_d(K) \\ -k_a(1) & -k_a(2) & \cdots & -k_a(K) & \sum_i k_a(i) & 0 \\ k_d(1) & k_d(2) & \cdots & k_d(K) & 0 & \sum_i k_d(i) \end{pmatrix}.$$

Box VIII.

Appendix B. NML: Refined algorithm, normalizing constant, and numerical issues

B.1. A refinement of the algorithm for computing NML

The sampling density for the NML computation via Monte Carlo integration, described in the body of the text, dominates the integrand. The speed of convergence of the algorithm can be increased by measures that make the sampling density more similar to the integrands. The sampling density described is proportional to the maximum likelihood of data patterns under the saturated model. In principle, it is sufficient to sample with probabilities proportional to the maximum likelihood of a more restrictive model as long as that model is a supermodel of the recognition-memory models. The maximum likelihood of a supermodel dominates the maximum likelihood of each submodel.

Thus, sampling can proceed by two-step rejection sampling, in which a proposal data pattern is first generated via rejection sampling from the density described in the body of the text, a supermodel (e.g., a mixture model mixing all ten models considered here) is then fitted, and the proposed data are accepted with probability given by the likelihood ratio of the supermodel and the saturated model.

The resulting algorithm converges much faster, because the maximum-likelihood of the supermodel, which is proportional to the density defined by this sampling scheme, is much more similar to the integrands than the saturated model itself, but there is a cost: The likelihood ratio of the supermodel and the saturated model is likely to be very small for most data sets so that many proposal data patterns have to be generated and the supermodel has to be fitted frequently before a data pattern is accepted for use in the Monte Carlo integration.

We found the trade-off propitious, however, for a supermodel that imposes independent restrictions on the false-alarm and hit rate for each (baserate or payoff) condition separately and for which a closed-form expression of the maximum likelihood exists. This allows us to employ the above rejection sampling for each condition separately and to determine the required maximum likelihoods fast. The algorithm implementing this idea converged reliably faster than the one based on the saturated model alone.

Specifically, we noted that all of the models considered impose restrictions on each single pair of false-alarm and hit rate. Let p_n and p_s be one such pair of false-alarm and hit rate, respectively. Most models imply that $p_s \geq p_n$, but the UVSDT is less restrictive in that it only implies that if $p_n \geq \frac{1}{2}$ then $p_s \geq \frac{1}{2}$ as is easy to see. This is in fact necessary and sufficient for the single pair of probabilities to be consistent with UVSDT. Furthermore, let $y_{r,t}$ be the frequency of response r, $r = \text{NO, YES}$, given noise trials ($t = n$) or signal trials ($t = s$), and let q_t be the number of trials of kind t that were observed. Finally, let l_t be the maximum likelihood of the saturated model for the frequencies of YES and NO responses for trial type t,

$$l_t = \binom{q_t}{y_{\text{YES},t}} \left(\frac{y_{\text{NO},t}}{q_t}\right)^{y_{\text{NO},t}} \left(\frac{y_{\text{YES},t}}{q_t}\right)^{y_{\text{YES},t}}.$$

The maximum likelihood l for the frequencies $y_{r,t}$, $r = \text{NO, YES}$, $t = n, s$, under the above UVSDT restrictions on the underlying probabilities p_s and p_n is that of the saturated model, if

$$\frac{y_{\text{YES},n}}{q_n} < \frac{1}{2} \quad \text{or} \quad \frac{y_{\text{YES},s}}{q_s} \geq \frac{1}{2},$$

and thus $l = l_s l_t$.

Otherwise, the maximum likelihood is given by

$$l = l_s \binom{q_n}{y_{\text{YES},n}} \left(\frac{1}{2}\right)^{q_n}, \quad \text{if } l_s > l_n, \text{ and by}$$

$$l = l_n \binom{q_s}{y_{\text{YES},s}} \left(\frac{1}{2}\right)^{q_s}, \text{ if } l_s \leq l_n.$$

The set of frequencies is accepted with probability given by $\frac{l}{l_s l_n}$ in a second step of rejection sampling.

To summarize, we first generate frequencies from a density proportional to the maximum likelihood of the saturated model for the generated frequencies. This is done for noise and signal trials of each (baserate or payoff) condition separately. Acceptance rates for an individual such step were .42, .13, and .04 for frequencies based on $N = 10$, 100, and 1000 (noise or signal) trials, respectively, indicating that this first rejection-sampling scheme heavily contributes to computation time for large N. Next, we consider the frequencies for noise and signal trials of each condition separately and submit these to a second rejection-sampling step so that the finally accepted frequencies are sampled from the maximum likelihood of the model defined by the above inequality restriction on p_s and p_n. Because this restriction

is implied by all of the models considered here, the resulting sampling density dominates all integrands. Sampling using this scheme is efficient, with acceptance rates of .84, .79, and .76 per (baserate or payoff) condition with $N = 10$, 100, and 1000 noise and signal trials, respectively. The resulting sampling density is more similar to the integrands than the maximum likelihood of the saturated model, leading to a sizable acceleration of the basic algorithm described in the body of the text.

B.2. Estimating the normalizing constant

The NML algorithm as described so far estimates NML up to an additive constant Q. The additive constant that needs to be removed is minus the logarithm of the integral of the probability function of the supermodel characterized in the previous section.

Monitoring the acceptance rates in the rejection sampling steps of the algorithm allows one to estimate that constant. This uses the following result (e.g., Evans and Swartz (2000, chap.3)). If g is a proposal density and f is proportional to a density from which we wish to sample with $f \leq Mg$ for some constant M, then the probability of accepting samples from g in rejection sampling, p_a, equals $\frac{1}{M}\int f$.

It is straightforward to apply this result to the present sampling scheme. Let $q_{s,i}$ and $q_{n,i}$ be the numbers of signal and noise trials, respectively, in condition i, $i = 1, \ldots, K$, as before. Furthermore, let $p_{1,t,i}$ be the estimated acceptance probabilities in the first rejection sampling scheme for sampling from the saturated model for trials of kind $t = s, n$ and condition i, $i = 1, \ldots, K$, and let $p_{2,i}$ be the estimated acceptance probability in the second rejection sampling scheme for sampling subsequently from the model described in the previous section for each condition i. The logarithm of the desired integral, that is minus the additive constant Q, is then estimated by:

$$-Q = \sum_{i=1}^{K} \log(p_{2,i}) + \sum_{t=s,n} \sum_{i=1}^{K} (\log(q_{t,i} + 1) + \log(p_{1,t,i})).$$

If $N_{1,t,i}$ is the number of times proposal frequencies for YES and NO responses were generated from the uniform distribution for trials of type t in condition i, $i = 1, \ldots, K$, in the first rejection sampling scheme, and $N_{2,i}$ the number of times a proposal set of noise and signal responses from the saturated model was entered into the second rejection sampling scheme in condition i, then its asymptotic standard error (SE) is estimated by

$$\mathrm{SE}(Q) = \left(\sum_{i=1}^{K} \frac{1}{N_{2,i}} \frac{1 - p_{2,i}}{p_{2,i}} + \sum_{t=s,n} \sum_{i=1}^{K} \frac{1}{N_{1,t,i}} \frac{1 - p_{1,t,i}}{p_{1,t,i}} \right)^{\frac{1}{2}}.$$

This uses the result that the standard error of an estimate of a binomial probability, $\hat{p}$, based on N trials is $\left(\frac{1}{N}\hat{p}(1-\hat{p})\right)^{\frac{1}{2}}$ and that hence, the standard error of $\log(\hat{p})$ is $\frac{\partial}{\partial p}\log(\hat{p})\left(\frac{1}{N}\hat{p}(1-\hat{p})\right)^{\frac{1}{2}} = \left(\frac{1}{N}\frac{1-\hat{p}}{\hat{p}}\right)^{\frac{1}{2}}$ (Rao, 1973, chap. 6a).

Because $N_{1,t,i}$ and $N_{2,i}$ quickly become very large, this additional source of error is small relative to that implied by the integration via independent importance sampling. In monitoring the accuracy of the NML estimation, we took this additional source of error into account in determining the asymptotic confidence intervals with $z_{(1-\frac{\alpha}{2})} = 3$ so that the final NML estimate can be assumed to have a precision of one decimal place.

An alternative, in the present context less efficient way to estimate the constant is by applying the integration scheme with a known density on the space of data patterns as integrand, such as the product of binomial probability functions with parameter $p = 0.5$, a method also known as inverse importance sampling (Evans & Swartz, 2000, chap. 7.4). We checked that both methods converged on the same estimates of the additive constants.

B.3. Numerical issues

A couple of non-trivial numerical issues require careful attention in the NML computation. These are (a) round-off error in sums over many terms, (b) the problem of extreme parameter values, and (c) local minima in maximum-likelihood estimation.

Computing NML involves computing sums over many maximum-likelihood ratios of widely different absolute sizes. To mitigate possible build-ups of round-off errors, we compute such sums in batches of 1600, ordering values from smallest (in absolute terms) to largest, prior to summing (from smallest to largest).

Many of the data sets for which maximum-likelihood estimates need to be found will be associated with extreme parameter values. It is therefore essential to compute the (logarithms of the) predicted probabilities and their derivatives in a stable manner even for extreme parameter values. To do so, we worked on a log-probability scale in computing predicted probabilities and their first and second derivatives as much as possible. Coupled with efficient algorithms for computing the logarithm of the normal distribution function and for computing $\log(1 + x)$ (Linhart, 2008), this allows us to compute log-likelihood values and their derivatives even for extreme parameter values in a numerically stable fashion.

Another issue is the problem of local maxima in iterative maximum-likelihood estimation. Iterative algorithms of maximum-likelihood estimation do not guarantee that a global maximum will be found. Failure to find the global maximum will lead to an underestimation of the model's flexibility by the estimated NML. To mitigate local-maxima problems, we used rational starting values, based on the method of moments, for the simpler models, and for models that are supermodels of simpler models, we generated rational starting values by using the maximum-likelihood estimates determined for the simpler models. Moreover, whenever the diagnostics of the optimization routine (E04LYF from the NAG library) indicated that a maximum may not have been found, we re-ran the optimization routine with random starting points generated from the proposal densities derived for the FIA computation of the respective model. This was repeated up to three times for the $K + 1$ parameter models, up to five times for the $K + 2$ parameter values, and up to eight times for the $K + 3$ parameter model MSDT.

Prior to fixing these setting, we ensured that the results of the computation were not changed by forcing the algorithm to call the optimization routine more often than implied by these settings. Note finally that pronounced local maxima problems should lead to an underestimation of the flexibility in terms of the computed NML values as already mentioned. As can be seen in Fig. 3, the computed NML values consistently approach FIA from above, however, suggesting that any remaining problems of this kind cannot be grave.

Although a gold standard for assessing the accuracy of our computations of FIA and NML is not available, a couple of findings raise our confidence in the validity of our results. (1) For $K = 1$, the 1HTM is a saturated model, and its FIA_f-value can be determined analytically (Su et al., 2005). Our numerical estimate of FIA_f found that value within the pre-specified accuracy. (2) The FIA values for the discrete-state models also agreed with the values obtained through the use of the independent routine by Wu et al. (2010a,b). (3) The NML estimates approach the FIA estimates as N becomes large (see Fig. 3). Because both estimates depend on very different methods, NML on log-likelihood values at maximum-likelihood estimates aggregated across data sets, FIA on determinants of the Fisher estimates aggregated across samples of

parameter values, this cross-consistency lends credibility to both methods simultaneously.

References

Atkinson, R. C. (1963). A variable sensitivity theory of signal detection. *Psychological Review*, *70*, 91–106.

Balasubramanian, V. (1997). Statistical inference, Occam's razor and statistical mechanics on the space of probability distributions. *Neural Computation*, *9*, 349–368.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial processing tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86.

Blackwell, H. R. (1963). Neural theories of simple visual discriminations. *Journal of the Optical Society of America*, *53*, 129–160.

Brandt, M. (2007). Bridging the gap between measurement models and theories of human memory. *Journal of Psychology*, *215*, 72–85.

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear–or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 587–606.

Bröder, A., & Schütz, J. (2011). Correction to Bröder and Schütz 2009. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1031.

Burnham, K. P., & Anderson, D. R. (2005). *Model selection and multimodel inference* (2nd ed.). New York: Springer.

Cavagnaro, D., Myung, J. I., Pitt, M., & Kujala, J. (2010). Adaptive design optimization: a mutual information based approach to model discrimination in cognitive science. *Neural Computation*, *22*, 887–905.

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: theoretical developments with applications to recognition memory. *Psychological Review*, *109*, 710–721.

Dube, C., & Rotello, C. M. (in press). Binary ROCs in perception and recognition memory are curved. Journal of Experimental Psychology: Learning, Memory, and Cognition.

Dube, C., Rotello, C. M., & Heit, E. (2011). The belief bias effect is aptly named: a reply to Klauer and Kellen (2011). *Psychological Review*, *118*, 155–163.

Erdfelder, E., & Buchner, A. (1998). Process-dissociation measurement models: threshold theory or detection theory? *Journal of Experimental Psychology: General*, *127*, 83–96.

Evans, M., & Swartz, T. (2000). *Approximating integrals via Monte Carlo and deterministic methods*. New York: Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin B, D. (2004). *Bayesian data analyses* (2nd ed.). Boca Raton, Florida: Chapman & Hall/CRC.

Glanzer, M., Kim, K., Adams, J. K., & Hilford, A. (1999). Slope of the receiver operating characteristic in recognition memory. *Journal of Experimental Psychology*, *25*, 500–513.

Grünwald, P. (2007). *The minimum description length principle*. Cambridge, Mass: MIT Press.

Grünwald, P., & Navarro, D. J. (2009). NML, Bayes and true distributions: a comment on Karabatsos and Walker(2006). *Journal of Mathematical Psychology*, *53*, 43–51.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528–551.

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal detection models of yes/no and two-alternative-forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*, 291–306.

Karabatsos, G., & Walker, S. G. (2006). On the normalized maximum likelihood and Bayesian decision theory. *Journal of Mathematical Psychology*, *50*, 517–520.

Kellen, D., & Klauer, K. C. (2011). Evaluating theories of recognition memory using first - and second-choice responses. *Journal of Mathematical Psychology*, *55*, 251–266.

Klauer, K. C., & Kellen, D. (2010). Toward a complete decision model of item and source recognition: a discrete-state approach. *Psychonomic Bulletin & Review*, *17*, 465–478.

Klauer, K. C., & Kellen, D. (2011). Assessing the belief bias effect with ROCs: reply to Dube, Rotello and Heit (2011). *Psychological Review*, *118*, 164–173.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). *Continuous multivariate distributions*: *Vol. 1. Models and applications* (2nd ed.). New York: Wiley.

Linhart, H., & Zuccini, W. (1986). *Model selection*. New York: Wiley.

Linhart, J.-M. (2008). Algorithm 885: computing the logarithm of the normal distribution. *ACM Transactions on Mathematical Software*, *35*, 20:1–20:10.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 275–362.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: a user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Malmberg, K. J. (2002). Observations on the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 380–387.

Malmberg, K. J. (2008). Recognition memory: a review of the critical findings and an integrated theory for relating them. *Cognitive Psychology*, *57*, 335–384.

McNicol, D. (1972). *A primer in signal detection theory*. London: Allen & Unwin.

Meyer, C. D. (2001). *Matrix analysis and applied linear algebra*. Philadelphia, PA: SIAM.

Myung, J. I., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: differential geometry and model selection. *Proceedings of the National Academy of Sciences USA*, *97*, 11170–11175.

Myung, J. I., Forster, M., & Brown, M. W. (2000). A special issue on model selection. *Journal of Mathematical Psychology*, *44*, 1–2.

Myung, J. I., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*, 167–179.

Myung, J. I., Pitt, M. A., & Navarro, D. J. (2007). Does response scaling cause the generalized context model to mimic a prototype model? *Psychonomic Bulletin*, *14*, 1043–1050.

Navarro, D. J. (2004). A note on the applied use of MDL approximations. *Neural Computation*, *16*, 1763–1768.

Navarro, D. J., & Lee, M. D. 2005. An application of minimum description length clustering to partitioning learning curves. In: Proceedings of the 2005 IEEE international symposium on information theory (pp. 587–591). Piscataway, NJ: IEEE.

Navarro, D. J., Pitt, M. A., & Myung, J. I. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*, 47–84.

Onyper, S., Zhang, Y., & Howard, M. W. (2010). Some-or-none recollection: evidence for item and source memory. *Journal of Experimental Psychology: General*, *139*, 341–362.

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, *41*, 227–259.

Rao, C. R. (1973). *Linear statistical inference and its applications*. New York: Wiley.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*, 518–535.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*, 40–47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, *47*, 1712–1717.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34–50.

Su, Y., Myung, J. I., & Pitt, M. A. (2005). Minimum description length and cognitive modeling. In P. Grünwald, J. I. Myung, & M. A. Pitt (Eds.), *Advances in minimum description length: theory and applications* (pp. 411–433). Cambridge, MA: MIT Press.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin*, *99*, 100–117.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, *68*, 301–340.

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28–50.

Wagenmakers, E. J., & Waldorf, L. (2006). Editors introduction. *Journal of Mathematical Psychology*, *50*, 99–100.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*, 92–107.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*, 152–176.

Wu, H., Myung, J. I., & Batchelder, W. H. (2010a). Minimum description length model selection of multinomial processing tree models. *Psychonomic Bulletin & Review*, *17*, 275–286.

Wu, H., Myung, J. I., & Batchelder, W. H. (2010b). On the minimum description length complexity of multinomial processing tree models. *Journal of Mathematical Psychology*, *54*, 291–303.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, *25*, 747–763.