A tutorial on Fisher information

Ly, A.; Marsman, M.; Verhagen, J.; Grasman, R.P.P.P.; Wagenmakers, E.-M.

DOI
10.1016/j.jmp.2017.05.006

Publication date
2017

Document Version
Final published version

Published in
Journal of Mathematical Psychology

Journal of Mathematical Psychology 80 (2017) 40–55

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Tutorial

A Tutorial on Fisher information

Alexander Ly *, Maarten Marsman, Josine Verhagen, Raoul P.P.P. Grasman, Eric-Jan Wagenmakers

University of Amsterdam, Department of Psychological Methods, PO Box 15906, 1001 NK Amsterdam, The Netherlands

HIGHLIGHTS

- We illustrate the use of Fisher information in the three statistical paradigms: frequentist, Bayesian, and MDL.
- Fisher information is used to construct hypothesis tests and confidence intervals.
- Fisher information is used to construct the Jeffreys's prior.
- Fisher information is used to measure model complexity.

ARTICLE INFO

Article history:
Received 22 May 2014
Received in revised form 27 April 2017
Available online 16 August 2017

Keywords:
Confidence intervals
Hypothesis testing
Jeffreys's prior
Minimum description length
Model complexity
Model selection
Statistical modeling

ABSTRACT

In many statistical applications that concern mathematical psychologists, the concept of Fisher information plays an important role. In this tutorial we clarify the concept of Fisher information as it manifests itself across three different statistical paradigms. First, in the frequentist paradigm, Fisher information is used to construct hypothesis tests and confidence intervals using maximum likelihood estimators; second, in the Bayesian paradigm, Fisher information is used to define a default prior; finally, in the minimum description length paradigm, Fisher information is used to measure model complexity.

Mathematical psychologists develop and apply quantitative models in order to describe human behavior and understand latent psychological processes. Examples of such models include Stevens' law of psychophysics that describes the relation between the objective physical intensity of a stimulus and its subjectively experienced intensity (Stevens, 1957); Ratcliff's diffusion model of decision making that measures the various processes that drive behavior in speeded response time tasks (Ratcliff, 1978); and multinomial processing tree models that decompose performance in memory tasks into the contribution of separate latent mechanisms (Batchelder & Riefer, 1980; Chechile, 1973).

When applying their models to data, mathematical psychologists may operate from within different statistical paradigms and focus on different substantive questions. For instance, working within the classical or frequentist paradigm a researcher may wish to test certain hypotheses or decide upon the number of trials to be presented to participants in order to estimate their latent abilities. Working within the Bayesian paradigm a researcher may wish to know how to determine a suitable default prior on the parameters of a model. Working within the minimum description length (MDL) paradigm a researcher may wish to compare rival models and quantify their complexity. Despite the diversity of these paradigms and purposes, they are connected through the concept of Fisher information.

Fisher information plays a pivotal role throughout statistical modeling, but an accessible introduction for mathematical psychologists is lacking. The goal of this tutorial is to fill this gap and illustrate the use of Fisher information in the three statistical paradigms mentioned above: frequentist, Bayesian, and MDL. This work builds directly upon the *Journal of Mathematical Psychology* tutorial article by Myung (2003) on maximum likelihood estimation. The intended target group for this tutorial are graduate students and researchers with an affinity for cognitive modeling and mathematical statistics.

To keep this tutorial self-contained we start by describing our notation and key concepts. We also provide the definition of Fisher

* Corresponding author.
E-mail address: a.ly@uva.nl (A. Ly).

http://dx.doi.org/10.1016/j.jmp.2017.05.006
0022-2496/

information and show how it can be calculated. The ensuing sections exemplify the use of Fisher information for different purposes. Section 2 shows how Fisher information can be used in frequentist statistics to construct confidence intervals and hypothesis tests from maximum likelihood estimators (MLEs). Section 3 shows how Fisher information can be used in Bayesian statistics to define a default prior on model parameters. In Section 4 we clarify how Fisher information can be used to measure model complexity within the MDL framework of inference.

1. Notation and key concepts

Before defining Fisher information it is necessary to discuss a series of fundamental concepts such as the nature of statistical models, probability mass functions, and statistical independence. Readers familiar with these concepts may safely skip to the next section.

A *statistical model* is typically defined through a function $f(x_i \mid \theta)$ that represents how a parameter θ is functionally related to potential outcomes x_i of a random variable X_i. For ease of exposition, we take θ to be one-dimensional throughout this text. The generalization to vector-valued θ can be found in an online Appendix (https://osf.io/hxxsj/), see also Myung and Navarro (2005).

As a concrete example, θ may represent a participant's intelligence, X_i a participant's (future) performance on the ith item of an IQ test, $x_i = 1$ the potential outcome of a correct response, and $x_i = 0$ the potential outcome of an incorrect response on the ith item. Similarly, X_i is the ith trial in a coin flip experiment with two potential outcomes: heads, $x_i = 1$, or tails, $x_i = 0$. Thus, we have the binary outcome space $\mathcal{X} = \{0, 1\}$. The coin flip model is also known as the Bernoulli distribution $f(x_i \mid \theta)$ that relates the coin's propensity $\theta \in (0, 1)$ to land heads to the potential outcomes as

$$f(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}, \quad \text{where } x_i \in \mathcal{X} = \{0, 1\}. \tag{1}$$

Formally, if θ is known, fixing it in the functional relationship f yields a function $p_\theta(x_i) = f(x_i \mid \theta)$ of the potential outcomes x_i. This $p_\theta(x_i)$ is referred to as a *probability density function* (pdf) when X_i has outcomes in a continuous interval, whereas it is known as a *probability mass function* (pmf) when X_i has discrete outcomes. The pmf $p_\theta(x_i) = P(X_i = x_i \mid \theta)$ can be thought of as a data generative device as it specifies how θ defines the chance with which X_i takes on a potential outcome x_i. As this holds for any outcome x_i of X_i, we say that X_i is distributed according to $p_\theta(x_i)$. For brevity, we do not further distinguish the continuous from the discrete case, and refer to $p_\theta(x_i)$ simply as a pmf.

For example, when the coin's true propensity is $\theta^* = 0.3$, replacing θ by θ^* in the Bernoulli distribution yields the pmf $p_{0.3}(x_i) = 0.3^{x_i}0.7^{1-x_i}$, a function of all possible outcomes of X_i. A subsequent replacement $x_i = 0$ in the pmf $p_{0.3}(0) = 0.7$ tells us that this coin generates the outcome 0 with 70% chance.

In general, experiments consist of n trials yielding a potential set of outcomes $x^n = (x_1, \ldots, x_n)$ of the random vector $X^n = (X_1, \ldots, X_n)$. These n random variables are typically assumed to be *independent and identically distributed* (i.i.d.). Identically distributed implies that each of these n random variables is governed by one and the same θ, while independence implies that the joint distribution of all these n random variables simultaneously is given by a product, that is,

$$f(x^n \mid \theta) = f(x_1 \mid \theta) \times \cdots \times f(x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta). \tag{2}$$

As before, when θ is known, fixing it in this relationship $f(x^n \mid \theta)$ yields the (joint) pmf of X^n as $p_\theta(x^n) = p_\theta(x_1) \times \cdots \times p_\theta(x_n) = \prod_{i=1}^{n} p_\theta(x_i)$.

In psychology the i.i.d. assumption is typically evoked when experimental data are analyzed in which participants have been confronted with a sequence of n items of roughly equal difficulty. When the participant can be either correct or incorrect on each trial, the participant's performance X^n can then be related to an n-trial coin flip experiment governed by one single θ over all n trials. The random vector X^n has 2^n potential outcomes x^n. For instance, when $n = 10$, we have $2^n = 1024$ possible outcomes and we write $\mathcal{X}^n$ for the collection of all these potential outcomes. The chance of observing a potential outcome x^n is determined by the coin's propensity θ as follows

$$\begin{aligned} f(x^n \mid \theta) &= f(x_1 \mid \theta) \times \cdots \times f(x_n \mid \theta) \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}, \quad \text{where } x^n \in \mathcal{X}^n. \end{aligned} \tag{3}$$

When the coin's true propensity θ is $\theta^* = 0.6$, replacing θ by θ^* in Eq. (3) yields the joint pmf $p_{0.6}(x^n) = f(x^n \mid \theta = 0.6) = 0.6^{\sum_{i=1}^n x_i}0.4^{n-\sum_{i=1}^n x_i}$. The pmf with a particular outcome entered, say, $x^n = (1, 1, 1, 1, 1, 1, 1, 0, 0, 0)$ reveals that the coin with $\theta^* = 0.6$ generates this particular outcome with 0.18% chance.

1.1. Definition of Fisher information

In practice, the true value of θ is not known and has to be inferred from the observed data. The first step typically entails the creation of a data summary. For example, suppose once more that X^n refers to an n-trial coin flip experiment and suppose that we observed $x^n_{\text{obs}} = (1, 0, 0, 1, 1, 1, 1, 0, 1, 1)$. To simplify matters, we only record the number of heads as $Y = \sum_{i=1}^n X_i$, which is a function of the data. Applying our function to the specific observations yields the realization $y_{\text{obs}} = Y(x^n_{\text{obs}}) = 7$. Since the coin flips X^n are governed by θ, so is a function of X^n; indeed, θ relates to the potential outcomes y of Y as

$$f(y \mid \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad \text{where } y \in \mathcal{Y} = \{0, 1, \ldots, n\}, \tag{4}$$

where $\binom{n}{y} = \frac{n!}{y!(n-y)!}$ enumerates the possible sequences of length n that consist of y heads and $n-y$ tails. For instance, when flipping a coin $n = 10$ times, there are 120 possible sequences of zeros and ones that contain $y = 7$ heads and $n - y = 3$ tails. The distribution $f(y \mid \theta)$ is known as the binomial distribution.

The summary statistic Y has $n + 1$ possible outcomes, whereas X^n has 2^n. For instance, when $n = 10$ the statistic Y has only 11 possible outcomes, whereas X^n has 1024. This reduction results from the fact that the statistic Y ignores the order with which the data are collected. Observe that the conditional probability of the raw data given $Y = y$ is equal to $P(X^n \mid Y = y, \theta) = 1/\binom{n}{y}$ and that it does not depend on θ. This means that after we observe $Y = y$ the conditional probability of X^n is independent of θ, even though each of the distributions of X^n and Y separately do depend on θ. We, therefore, conclude that there is no information about θ left in X^n after observing $Y = y$ (Fisher, 1920; Stigler, 1973).

More generally, we call a function of the data, say, $T = t(X^n)$ a *statistic*. A statistic is referred to as *sufficient* for the parameter θ, if the expression $P(X^n \mid T = t, \theta)$ does not depend on θ itself. To quantify the amount of information about the parameter θ in a sufficient statistic T and the raw data, Fisher introduced the following measure.

Definition 1 (*Fisher Information*)**.** The *Fisher information* $I_X(\theta)$ of a random variable X about θ is defined as[1]

$$I_X(\theta) = \begin{cases} \sum_{x \in \mathcal{X}} \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(x \mid \theta) \right)^2 p_\theta(x) & \text{if } X \text{ is discrete,} \\ \int_{\mathcal{X}} \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(x \mid \theta) \right)^2 p_\theta(x) \mathrm{d}x & \text{if } X \text{ is continuous} \end{cases} \tag{6}$$

The derivative $\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(x \mid \theta)$ is known as the *score function*, a function of x, and describes how sensitive the model (i.e., the functional form f) is to changes in θ at a particular θ. The Fisher information measures the overall sensitivity of the functional relationship f to changes of θ by weighting the sensitivity at each potential outcome x with respect to the chance defined by $p_\theta(x) = f(x \mid \theta)$. The weighting with respect to $p_\theta(x)$ implies that the Fisher information about θ is an expectation.

Similarly, Fisher information $I_{X^n}(\theta)$ within the random vector X^n about θ is calculated by replacing $f(x \mid \theta)$ with $f(x^n \mid \theta)$, thus, $p_\theta(x)$ with $p_\theta(x^n)$ in the definition. Moreover, under the assumption that the random vector X^n consists of n i.i.d. trials of X it can be shown that $I_{X^n}(\theta) = nI_X(\theta)$, which is why $I_X(\theta)$ is also known as the unit Fisher information.[2] Intuitively, an experiment consisting of $n = 10$ trials is expected to be twice as informative about θ compared to an experiment consisting of only $n = 5$ trials. ◇

Intuitively, we cannot expect an arbitrary summary statistic T to extract more information about θ than what is already provided by the raw data. Fisher information adheres to this rule, as it can be shown that

$$I_{X^n}(\theta) \geq I_T(\theta), \tag{7}$$

with equality if and only if T is a sufficient statistic for θ.

Example 1 (*The Information About* θ *Within the Raw Data and A Summary Statistic*)**.** A direct calculation with a Bernoulli distributed random vector X^n shows that the Fisher information about θ within an n-trial coin flip experiment is given by

$$I_{X^n}(\theta) = nI_X(\theta) = n \frac{1}{\theta(1-\theta)}, \tag{8}$$

where $I_X(\theta) = \frac{1}{\theta(1-\theta)}$ is the Fisher information of θ within a single trial. As shown in Fig. 1, the unit Fisher information $I_X(\theta)$ depends on θ. Similarly, we can calculate the Fisher information about θ within the summary statistic Y by using the binomial model instead. This yields $I_Y(\theta) = \frac{n}{\theta(1-\theta)}$. Hence, $I_{X^n}(\theta) = I_Y(\theta)$ for any value of θ. In other words, the expected information in Y about θ is the same as the expected information about θ in X^n, regardless of the value of θ. ◇

Observe that the information in the raw data X^n and the statistic Y are equal for every θ, and specifically also for its unknown true value θ^*. That is, there is no statistical information about θ lost when we use a sufficient statistic Y instead of the raw data X^n.

[1] Under mild regularity conditions Fisher information is equivalently defined as

$$I_X(\theta) = -E\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(X \mid \theta) \right) = \begin{cases} -\sum_{x \in \mathcal{X}} \left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(x \mid \theta) \right) p_\theta(x) & \text{if } X \text{ is discrete,} \\ -\int_{\mathcal{X}} \left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(x \mid \theta) \right) p_\theta(x) \mathrm{d}x & \text{if } X \text{ is continuous.} \end{cases} \tag{5}$$

where $\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(x \mid \theta)$ denotes the second derivate of the logarithm of f with respect to θ.

[2] Note the abuse of notation—we dropped the subscript i for the ith random variable X_i and denote it simply by X instead.

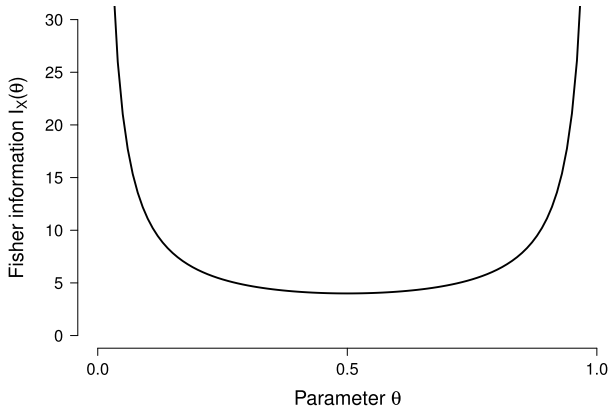

Fig. 1. The unit Fisher information $I_X(\theta) = \frac{1}{\theta(1-\theta)}$ as a function of θ within the Bernoulli model. As θ reaches zero or one the expected information goes to infinity.

This is particular useful when the data set X^n is large and can be replaced by single number Y.

2. The role of Fisher information in frequentist statistics

Recall that θ is unknown in practice and to infer its value we might: (1) provide a best guess in terms of a point estimate; (2) postulate its value and test whether this value aligns with the data, or (3) derive a confidence interval. In the frequentist framework, each of these inferential tools is related to the Fisher information and exploits the data generative interpretation of a pmf. Recall that given a model $f(x^n \mid \theta)$ and a known θ, we can view the resulting pmf $p_\theta(x^n)$ as a recipe that reveals how θ defines the chances with which X^n takes on the potential outcomes x^n.

This data generative view is central to Fisher's conceptualization of the *maximum likelihood estimator* (MLE; Fisher, 1912, 1922, 1925; LeCam, 1990; Myung, 2003). For instance, the binomial model implies that a coin with a hypothetical propensity $\theta = 0.5$ will generate the outcome $y = 7$ heads out of $n = 10$ trials with 11.7% chance, whereas a hypothetical propensity of $\theta = 0.7$ will generate the same outcome $y = 7$ with 26.7% chance. Fisher concluded that an actual observation $y_{\text{obs}} = 7$ out of $n = 10$ is therefore more likely to be generated from a coin with a hypothetical propensity of $\theta = 0.7$ than from a coin with a hypothetical propensity of $\theta = 0.5$. Fig. 2 shows that for this specific observation $y_{\text{obs}} = 7$, the hypothetical value $\theta = 0.7$ is the maximum likelihood *estimate*; the number $\hat{\theta}_{\text{obs}} = 0.7$. This estimate is a realization of the maximum likelihood *estimator* (MLE); in this case, the MLE is the function $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i = \frac{1}{n}Y$, i.e., the sample mean. Note that the MLE is a statistic, that is, a function of the data.

2.1. Using Fisher information to design an experiment

Since X^n depends on θ so will a function of X^n, in particular, the MLE $\hat{\theta}$. The distribution of the potential outcomes of the MLE $\hat{\theta}$ is known as the *sampling distribution* of the estimator and denoted as $f(\hat{\theta}_{\text{obs}} \mid \theta)$. As before, when θ^* is assumed to be known, fixing it in $f(\hat{\theta}_{\text{obs}} \mid \theta)$ yields the pmf $p_{\theta^*}(\hat{\theta}_{\text{obs}})$, a function of the potential outcomes of $\hat{\theta}$. This function f between the parameter θ and the potential outcomes of the MLE $\hat{\theta}$ is typically hard to describe, but for n large enough it can be characterized by the Fisher information.

For i.i.d. data and under general conditions,[3] the difference between the true θ^* and the MLE converges in distribution to a

[3] Basically, when the Fisher information exists for all parameter values. For details see the advanced accounts provided by Bickel, Klaassen, Ritov, and Wellner (1993), Hájek (1970), Inagaki (1970), LeCam (1970) and the online appendix.

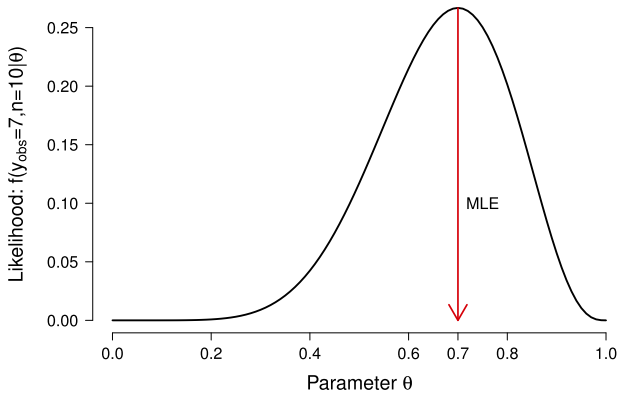

Fig. 2. The likelihood function based on observing $y_{\text{obs}} = 7$ heads in $n = 10$ trials. For these data, the MLE is equal to $\hat{\theta}_{\text{obs}} = 0.7$, see the main text for the interpretation of this function.

normal distribution, that is,

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{D} \mathcal{N}(0, I_X^{-1}(\theta^*)), \quad \text{as } n \to \infty. \tag{9}$$

Hence, for large enough n, the "error" is approximately normally distributed[4]

$$(\hat{\theta} - \theta^*) \overset{D}{\approx} \mathcal{N}\Big(0, 1/\big(nI_X(\theta^*)\big)\Big). \tag{10}$$

This means that the MLE $\hat{\theta}$ generates potential estimates $\hat{\theta}_{\text{obs}}$ around the true value θ^* with a standard error given by the inverse of the square root of the Fisher information at the true value θ^*, i.e., $1/\sqrt{nI_X(\theta^*)}$, whenever n is large enough. Note that the chances with which the estimates of $\hat{\theta}$ are generated depend on the true value θ^* and the sample size n. Observe that the standard error decreases when the unit information $I_X(\theta^*)$ is high or when n is large. As experimenters we do not have control over the true value θ^*, but we can affect the data generating process by choosing the number of trials n. Larger values of n increase the amount of information in X^n, heightening the chances of the MLE producing an estimate $\hat{\theta}_{\text{obs}}$ that is close to the true value θ^*. The following example shows how this can be made precise.

Example 2 (*Designing a Binomial Experiment with the Fisher Information*)**.** Recall that the potential outcomes of a normal distribution fall within one standard error of the population mean with 68% chance. Hence, when we choose n such that $1/\sqrt{nI_X(\theta^*)} = 0.1$ we design an experiment that allows the MLE to generate estimates within 0.1 distance of the true value with 68% chance. To overcome the problem that θ^* is not known, we solve the problem for the worst case scenario. For the Bernoulli model this is given by $\theta = 1/2$, the least informative case, see Fig. 1. As such, we have $1/\sqrt{nI_X(\theta^*)} \leq 1/\sqrt{nI_X(1/2)} = 1/(2\sqrt{n}) = 0.1$, where the last equality is the target requirement and is solved by $n = 25$.

This leads to the following interpretation. After simulating $k = 100$ data sets $x^n_{\text{obs},1}, \ldots, x^n_{\text{obs},k}$ each with $n = 25$ trials, we can apply to each of these data sets the MLE yielding k estimates $\hat{\theta}_{\text{obs},1}, \ldots, \hat{\theta}_{\text{obs},k}$. The sampling distribution implies that at least 68 of these $k = 100$ estimate are expected to be at most 0.1 distance away from the true θ^*. ◇

[4] Note that $\hat{\theta}$ is random, while the true value θ^* is fixed. As such, the errors $\hat{\theta} - \theta^*$ and $\sqrt{n}(\hat{\theta} - \theta^*)$ are also random. We used $\xrightarrow{D}$ to convey that the distribution of the left-hand side goes to the distribution on the right-hand side in Eq. (9). Similarly, $\overset{D}{\approx}$ implies that the distribution of the left-hand side is *approximately* equal to the distribution given on the right hand-side in Eq. (10). Hence, for finite n there will be an error due to using the normal distribution as an approximation to the true sampling distribution. This approximation error is ignored in the constructions given below, see the appendix for a more thorough discussion.

2.2. Using Fisher information to construct a null hypothesis test

The (asymptotic) normal approximation to the sampling distribution of the MLE can also be used to construct a null hypothesis test. When we postulate that the true value equals some hypothesized value of interest, say, $\theta^* = \theta_0$, a simple plugin then allows us to construct a prediction interval based on our knowledge of the normal distribution. More precisely, the potential outcomes x^n with n large enough and generated according to $p_{\theta^*}(x^n)$ leads to potential estimates $\hat{\theta}_{\text{obs}}$ that fall within the range

$$\left(\theta^* - 1.96\sqrt{\tfrac{1}{n}I_X^{-1}(\theta^*)}, \theta^* + 1.96\sqrt{\tfrac{1}{n}I_X^{-1}(\theta^*)}\right), \tag{11}$$

with (approximately) 95% chance. This 95%-prediction interval Eq. (11) allows us to construct a point null hypothesis test based on a pre-experimental postulate $\theta^* = \theta_0$.

Example 3 (*A Null Hypothesis Test for a Binomial Experiment*)**.** Under the null hypothesis $H_0 : \theta^* = \theta_0 = 0.5$, we predict that an outcome of the MLE based on $n = 10$ trials will lie between $(0.19, 0.81)$ with 95% chance. This interval follows from replacing θ^* by θ_0 in the 95%-prediction interval Eq. (11). The data generative view implies that if we simulate $k = 100$ data sets each with the same $\theta^* = 0.5$ and $n = 10$, we would then have k estimates $\hat{\theta}_{\text{obs},1}, \ldots, \hat{\theta}_{\text{obs},k}$ of which five are expected to be outside this 95% interval $(0.19, 0.81)$. Fisher, therefore, classified an outcome of the MLE that is smaller than 0.19 or larger than 0.81 as extreme under the null and would then reject the postulate $H_0 : \theta_0 = 0.5$ at a significance level of 0.05. ◇

The normal approximation to the sampling distribution of the MLE and the resulting null hypothesis test is particularly useful when the exact sampling distribution of the MLE is unavailable or hard to compute.

Example 4 (*An MLE Null Hypothesis Test for the Laplace Model*)**.** Suppose that we have n i.i.d. samples from the Laplace distribution

$$f(x_i \mid \theta) = \tfrac{1}{2b}\exp\Big(-\tfrac{|x_i - \theta|}{b}\Big), \tag{12}$$

where θ denotes the population mean and the population variance is given by $2b^2$. It can be shown that the MLE for this model is the sample median, $\hat{\theta} = \hat{M}$, and the unit Fisher information is $I_X(\theta) = b^{-2}$. The exact sampling distribution of the MLE is unwieldy (Kotz, Kozubowski, & Podgorski, 2001) and not presented here. Asymptotic normality of the MLE is practical, as it allows us to discard the unwieldy exact sampling distribution and, instead, base our inference on a more tractable (approximate) normal distribution with a mean equal to the true value θ^* and a variance equal to b^2/n. For $n = 100$, $b = 1$ and repeated sampling under the hypothesis $H_0 : \theta^* = \theta_0$, approximately 95% of the estimates (the observed sample medians) are expected to fall in the range $(\theta_0 - 0.196, \theta_0 + 0.196)$. ◇

2.3. Using Fisher information to compute confidence intervals

An alternative to both point estimation and null hypothesis testing is interval estimation. In particular, a 95%-confidence interval can be obtained by replacing in the prediction interval Eq. (11) the unknown true value θ^* by an estimate $\hat{\theta}_{\text{obs}}$. Recall that a simulation with $k = 100$ data sets each with n trials leads to $\hat{\theta}_{\text{obs},1}, \ldots, \hat{\theta}_{\text{obs},k}$ estimates, and each estimate leads to a different 95%-confidence interval. It is then expected that 95 of these $k = 100$ intervals encapsulate the true value θ^*.[5] Note

[5] But see Brown, Cai, and DasGupta (2001).

that these intervals are centered around different points whenever the estimates differ and that their lengths differ, as the Fisher information depends on θ.

Example 5 (*An MLE Confidence Interval for the Bernoulli Model*)**.** When we observe $y_{\text{obs},1} = 7$ heads in $n = 10$ trials, the MLE then produces the estimate $\hat{\theta}_{\text{obs},1} = 0.7$. Replacing θ^* in the prediction interval Eq. (11) with $\theta^* = \hat{\theta}_{\text{obs},1}$ yields an approximate 95%-confidence interval (0.42, 0.98) of length 0.57. On the other hand, had we instead observed $y_{\text{obs},2} = 6$ heads, the MLE would then yield $\hat{\theta}_{\text{obs},2} = 0.6$ resulting in the interval (0.29, 0.90) of length 0.61. ◇

In sum, Fisher information can be used to approximate the sampling distribution of the MLE when n is large enough. Knowledge of the Fisher information can be used to choose n such that the MLE produces an estimate close to the true value, construct a null hypothesis test, and compute confidence intervals.

3. The role of Fisher information in Bayesian statistics

This section outlines how Fisher information can be used to define the Jeffreys's prior, a default prior commonly used for estimation problems and for nuisance parameters in a Bayesian hypothesis test (e.g., Bayarri, Berger, Forte, & García-Donato, 2012; Dawid, 2011; Gronau, Ly, & Wagenmakers, 2017; Jeffreys, 1961; Li & Clyde, 2015; Liang, Paulo, Molina, Clyde, & Berger, 2008; Ly, Raj, Marsman, Etz, & Wagenmakers, 2017; Ly, Marsman, & Wagenmakers, in press; Ly, Verhagen, & Wagenmakers, 2016a, b; Robert, 2016). To illustrate the desirability of the Jeffreys's prior we first show how the naive use of a uniform prior may have undesirable consequences, as the uniform prior depends on the representation of the inference problem, that is, on how the model is parameterized. This dependence is commonly referred to as lack of invariance: different parameterizations of the same model result in different posteriors and, hence, different conclusions. We visualize the representation problem using simple geometry and show how the geometrical interpretation of Fisher information leads to the Jeffreys's prior that is parameterization-invariant.

3.1. Bayesian updating

Bayesian analysis centers on the observations x^n_{obs} for which a generative model f is proposed that functionally relates the observed data to an unobserved parameter θ. Given the observations x^n_{obs}, the functional relationship f is inverted using Bayes' rule to infer the relative plausibility of the values of θ. This is done by replacing the potential outcome part x^n in f by the actual observations yielding a *likelihood function* $f(x^n_{\text{obs}} \mid \theta)$, which is a function of θ. In other words, x^n_{obs} is known, thus, fixed, and the true θ is unknown, therefore, free to vary. The candidate set of possible values for the true θ is denoted by Θ and referred to as the parameter space. Our knowledge about θ is formalized by a distribution $g(\theta)$ over the parameter space Θ. This distribution is known as the prior on θ, as it is set before any datum is observed. We can use Bayes' theorem to calculate the posterior distribution over the parameter space Θ given the data that were actually observed as follows

$$g(\theta \mid X^n = x^n_{\text{obs}}) = \frac{f(x^n_{\text{obs}} \mid \theta) g(\theta)}{\int_\Theta f(x^n_{\text{obs}} \mid \theta) g(\theta) \, \mathrm{d}\theta}. \tag{13}$$

This expression is often verbalized as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}. \tag{14}$$

The posterior distribution is a combination of what we knew before we saw the data (i.e., the information in the prior), and what we have learned from the observations in terms of the likelihood (e.g., Lee & Wagenmakers, 2013). Note that the integral is now over θ and not over the potential outcomes.

3.2. Failure of the uniform distribution on the parameter as a non-informative prior

When little is known about the parameter θ that governs the outcomes of X^n, it may seem reasonable to express this ignorance with a uniform prior distribution $g(\theta)$, as no parameter value of θ is then favored over another. This leads to the following type of inference:

Example 6 (*Uniform Prior on* θ)**.** Before data collection, θ is assigned a uniform prior, that is, $g(\theta) = 1/V_\Theta$ with a normalizing constant of $V_\Theta = 1$ as shown in the left panel of Fig. 3. Suppose that we observe coin flip data x^n_{obs} with $y_{\text{obs}} = 7$ heads out of $n = 10$ trials. To relate these observations to the coin's propensity θ we use the Bernoulli distribution as our $f(x^n \mid \theta)$. A replacement of x^n by the data actually observed yields the likelihood function $f(x^n_{\text{obs}} \mid \theta) = \theta^7(1-\theta)^3$, which is a function of θ. Bayes' theorem now allows us to update our prior to the posterior that is plotted in the right panel of Fig. 3. ◇

Note that a uniform prior on θ has the length, more generally, volume, of the parameter space as the normalizing constant; in this case, $V_\Theta = 1$, which equals the length of the interval $\Theta = (0, 1)$. Furthermore, a uniform prior can be characterized as the prior that gives equal probability to all sub-intervals of equal length. Thus, the a priori probability of finding the true value θ^* within a sub-interval $J_\theta = (\theta_a, \theta_b) \subset \Theta = (0, 1)$ is given by the relative length of J_θ with respect to the length of the parameter space, that is,

$$P\left(\theta^* \in J_\theta\right) = \int_{J_\theta} g(\theta) \mathrm{d}\theta = \frac{1}{V_\Theta} \int_{\theta_a}^{\theta_b} 1 \mathrm{d}\theta = \frac{\theta_b - \theta_a}{V_\Theta}. \tag{15}$$

Hence, before any datum is observed, the uniform prior expresses the belief $P(\theta^* \in J_\theta) = 0.20$ of finding the true value θ^* within the interval $J_\theta = (0.6, 0.8)$. After observing x^n_{obs} with $y_{\text{obs}} = 7$ out of $n = 10$, this prior is updated to the posterior belief of $P(\theta^* \in J_\theta \mid x^n_{\text{obs}}) = 0.54$, see the shaded areas in Fig. 3.

Although intuitively appealing, it can be unwise to choose the uniform distribution by default, as the results are highly dependent on how the model is parameterized. In what follows, we show how a different parameterization leads to different posteriors and, consequently, different conclusions.

Example 7 (*Different Representations, Different Conclusions*)**.** The propensity of a coin landing heads up is related to the angle ϕ with which that coin is bent. Suppose that the relation between the angle ϕ and the propensity θ is given by the function $\theta = h(\phi) = \frac{1}{2} + \frac{1}{2}\left(\frac{\phi}{\pi}\right)^3$, chosen here for mathematical convenience.[6] When ϕ is positive the tail side of the coin is bent inwards, which increases the coin's chances to land heads. As the function $\theta = h(\phi)$ also admits an inverse function $h^{-1}(\theta) = \phi$, we have an equivalent formulation of the problem in Example 6, but now described in terms of the angle ϕ instead of the propensity θ.

As before, in order to obtain a posterior distribution, Bayes' theorem requires that we specify a prior distribution. As the problem is formulated in terms of ϕ, one may believe that a non-informative choice is to assign a uniform prior $\tilde{g}(\phi)$ on ϕ, as

[6] Another example involves the logit formulation of the Bernoulli model, that is, in terms of $\phi = \log(\frac{\theta}{1-\theta})$, where $\Phi = \mathbb{R}$. This logit formulation is the basic building block in item response theory. We did not discuss this example as the uniform prior on the logit cannot be normalized and, therefore, not easily represented in the plots.

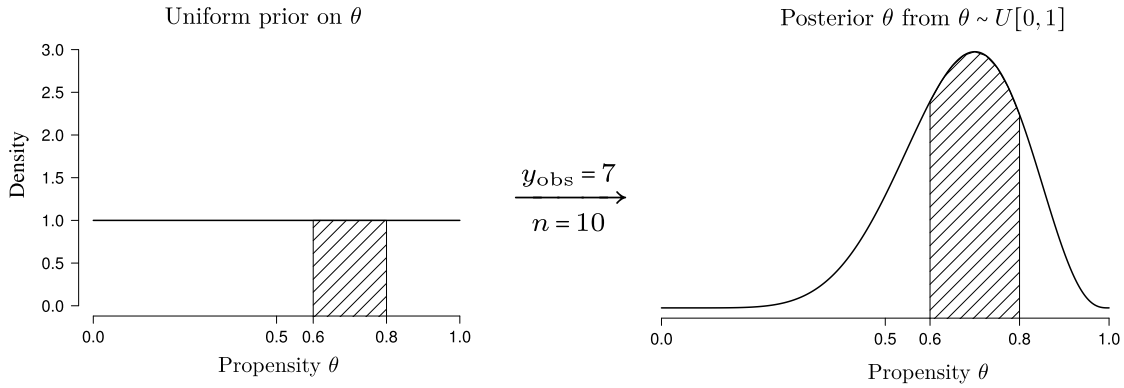

Fig. 3. Bayesian updating based on observations x^n_{obs} with $y_{\text{obs}} = 7$ heads out of $n = 10$ tosses. In the left panel, the uniform prior distribution assigns equal probability to every possible value of the coin's propensity θ. In the right panel, the posterior distribution is a compromise between the prior and the observed data.

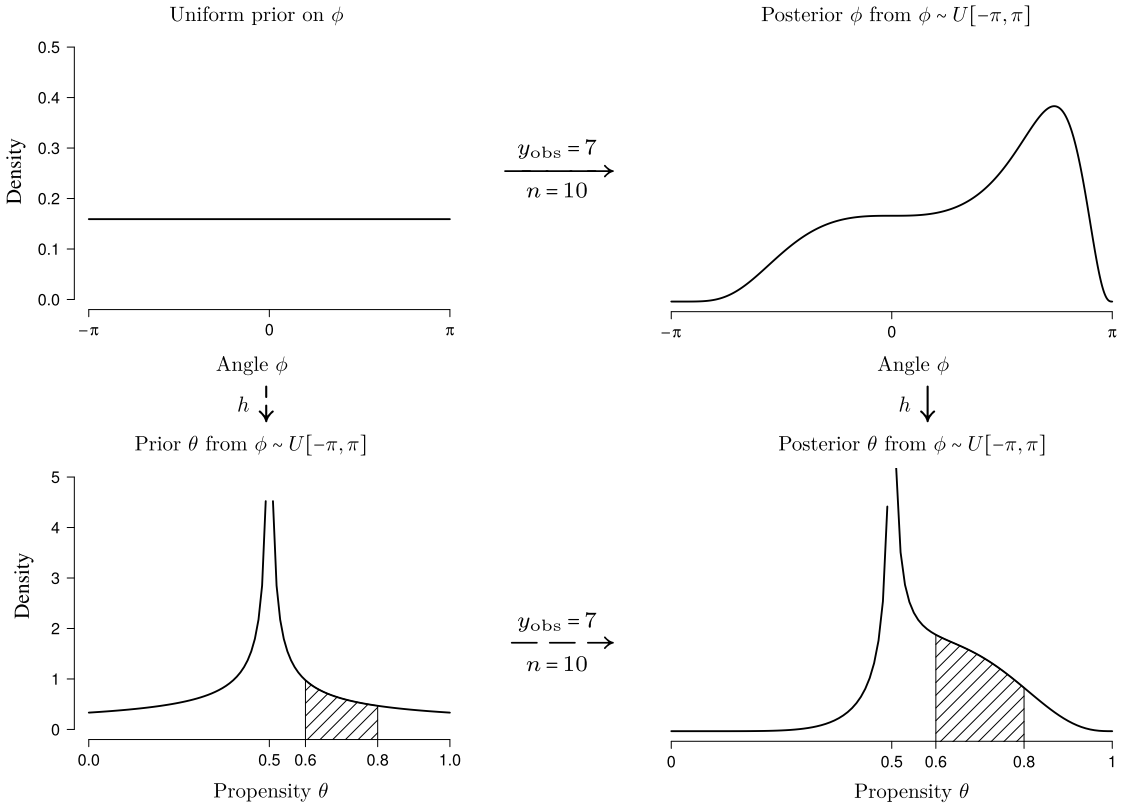

Fig. 4. Bayesian updating based on observations x^n_{obs} with $y_{\text{obs}} = 7$ heads out of $n = 10$ tosses when a uniform prior distribution is assigned to the coin's angle ϕ. The uniform distribution is shown in the top-left panel. Bayes' theorem results in a posterior distribution for ϕ that is shown in the top-right panel. This posterior $\tilde{g}(\phi \mid x^n_{\text{obs}})$ is transformed into a posterior on θ (bottom-right panel) using $\theta = h(\phi)$. The same posterior on θ is obtained if we proceed via an alternative route in which we first transform the uniform prior on ϕ to the corresponding prior on θ and then apply Bayes' theorem with the induced prior on θ. A comparison to the results from Fig. 3 reveals that posterior inference differs notably depending on whether a uniform distribution is assigned to the angle ϕ or to the propensity θ.

this means that no value of ϕ is favored over another. A uniform prior on ϕ is in this case given by $\tilde{g}(\phi) = 1/V_\Phi$ with a normalizing constant $V_\Phi = 2\pi$, because the parameter ϕ takes on values in the interval $\Phi = (-\pi, \pi)$. This uniform distribution expresses the belief that the true ϕ^* can be found in any of the intervals $(-1.0\pi, -0.8\pi), (-0.8\pi, -0.6\pi), \ldots, (0.8\pi, 1.0\pi)$ with 10% probability, because each of these intervals is 10% of the total length, see the top-left panel of Fig. 4. For the same data as before, the posterior calculated from Bayes' theorem is given in top-right panel of Fig. 4. As the problem in terms of the angle ϕ is equivalent to that of $\theta = h(\phi)$ we can use the function h to translate the posterior in terms of ϕ to a posterior on θ, see the bottom-right panel of Fig. 4. This posterior on θ is noticeably different from the posterior on θ shown in Fig. 3.

Specifically, the uniform prior on ϕ corresponds to the prior belief $\tilde{P}(\theta^* \in J_\theta) = 0.13$ of finding the true value θ^* within the interval $J_\theta = (0.6, 0.8)$. After observing x^n_{obs} with $y_{\text{obs}} = 7$ out of $n = 10$, this prior is updated to the posterior belief of $\tilde{P}(\theta^* \in J_\theta \mid x^n_{\text{obs}}) = 0.29$,[7] see the shaded areas in Fig. 4. Crucially, the earlier analysis that assigned a uniform prior to the propensity θ yielded a posterior probability $P(\theta^* \in J_\theta \mid x^n_{\text{obs}}) = 0.54$, which is markedly different from the current analysis that assigns a uniform prior to the angle ϕ.

The same posterior on θ is obtained when the prior on ϕ is first translated into a prior on θ (bottom-left panel) and then updated to

[7] The tilde makes explicit that the prior and posterior are derived from the uniform prior $\tilde{g}(\phi)$ on ϕ.

a posterior with Bayes' theorem. Regardless of the stage at which the transformation is applied, the resulting posterior on θ differs substantially from the result plotted in the right panel of Fig. 3. ◇

Thus, the uniform prior distribution is not a panacea for the quantification of prior ignorance, as the conclusions depend on how the problem is parameterized. In particular, a uniform prior on the coin's angle $\tilde{g}(\phi) = 1/V_\Phi$ yields a highly informative prior in terms of the coin's propensity θ. This lack of invariance caused Karl Pearson, Ronald Fisher and Jerzy Neyman to reject 19th century Bayesian statistics that was based on the uniform prior championed by Pierre-Simon Laplace. This rejection resulted in, what is now known as, frequentist statistics, see also Hald (2008), Lehmann (2011), and Stigler (1986).

3.3. A default prior by Jeffreys's rule

Unlike the other fathers of modern statistical thoughts, Harold Jeffreys continued to study Bayesian statistics based on formal logic and his philosophical convictions of scientific inference (see, e.g., Aldrich, 2005; Etz and Wagenmakers, 2017; Jeffreys, 1961; Ly et al., 2016a, b; Robert, Chopin, and Rousseau, 2009; Wrinch and Jeffreys, 1919, 1921, 1923). Jeffreys concluded that the uniform prior is unsuitable as a default prior due to its dependence on the parameterization. As an alternative, Jeffreys (1946) proposed the following prior based on Fisher information

$$g_J(\theta) = \frac{1}{V}\sqrt{I_X(\theta)}, \quad \text{where } V = \int_\Theta \sqrt{I_X(\theta)}\mathrm{d}\theta, \tag{16}$$

which is known as the prior derived from Jeffreys's rule or the *Jeffreys's prior* in short. The Jeffreys's prior is parameterization-invariant, which implies that it leads to the same posteriors regardless of how the model is represented.

Example 8 (*Jeffreys's Prior*)**.** The Jeffreys's prior of the Bernoulli model in terms of ϕ is

$$g_J(\phi) = \frac{3\phi^2}{V\sqrt{\pi^6 - \phi^6}}, \quad \text{where } V = \pi, \tag{17}$$

which is plotted in the top-left panel of Fig. 5. The corresponding posterior is plotted in the top-right panel, which we transformed into a posterior in terms of θ using the function $\theta = h(\phi)$ shown in the bottom-right panel.[8]

Similarly, we could have started with the Jeffreys's prior in terms of θ instead, that is,

$$g_J(\theta) = \frac{1}{V\sqrt{\theta(1-\theta)}}, \quad \text{where } V = \pi. \tag{18}$$

The Jeffreys's prior and posterior on θ are plotted in the bottom-left and the bottom-right panel of Fig. 5, respectively. The Jeffreys's prior on θ corresponds to the prior belief $P_J(\theta^* \in J_\theta) = 0.14$ of finding the true value θ^* within the interval $J_\theta = (0.6, 0.8)$. After observing x^n_{obs} with $y_{\text{obs}} = 7$ out of $n = 10$, this prior is updated to the posterior belief of $P_J(\theta^* \in J_\theta \mid x^n_{\text{obs}}) = 0.53$, see the shaded areas in Fig. 5. The posterior is identical to the one obtained from the previously described updating procedure that starts with the Jeffreys's prior on ϕ instead of on θ. ◇

This example shows that the Jeffreys's prior leads to the same posterior knowledge regardless of how we as researcher represent the problem. Hence, the same conclusions about θ are drawn regardless of whether we (1) use Jeffreys's rule to construct a prior on θ and update with the observed data, or (2) use Jeffreys's rule to construct a prior on ϕ, update to a posterior distribution on ϕ, which is then transformed to a posterior on θ.

[8] The subscript J makes explicit that the prior and posterior are based on the prior derived from Jeffreys's rule, i.e., $g_J(\theta)$ on θ, or equivalently, $g_J(\phi)$ on ϕ.

3.4. Geometrical properties of Fisher information

In the remainder of this section we make intuitive that the Jeffreys's prior is in fact uniform in the model space. We elaborate on what is meant by model space and how this can be viewed geometrically. This geometric approach illustrates (1) the role of Fisher information in the definition of the Jeffreys's prior, (2) the interpretation of the shaded area, and (3) why the normalizing constant is $V = \pi$, regardless of the chosen parameterization.

3.4.1. The model space $\mathcal{M}$

Before we describe the geometry of statistical models, recall that a pmf can be thought of as a data generating device of X, as the pmf specifies the chances with which X takes on the potential outcomes 0 and 1. Each such pmf has to fulfill two conditions: (i) the chances have to be non-negative, that is, $0 \leq p(x) = P(X = x)$ for every possible outcome x of X, and (ii) to explicitly convey that there are $w = 2$ outcomes, and none more, the chances have to sum to one, that is, $p(0) + p(1) = 1$. We call the largest set of functions that adhere to conditions (i) and (ii) the complete set of pmfs $\mathcal{P}$.

As any pmf from $\mathcal{P}$ defines $w = 2$ chances, we can represent such a pmf as a vector in w dimensions. To simplify notation, we write $p(X)$ for all w chances simultaneously, hence, $p(X)$ is the vector $p(X) = [p(0), p(1)]$ when $w = 2$. The two chances with which a pmf $p(X)$ generates outcomes of X can be simultaneously represented in the plane with $p(0) = P(X = 0)$ on the horizontal axis and $p(1) = P(X = 1)$ on the vertical axis. In the most extreme case, we have the pmfs $p(X) = [1, 0]$ or $p(X) = [0, 1]$. These two extremes are linked by a straight line in the left panel of Fig. 6. *Any* pmf – and the true pmf $p^*(X)$ of X in particular – can be uniquely identified with a vector on the line and vice versa. For instance, the pmf $p_e(X) = [1/2, 1/2]$ (i.e., the two outcomes are generated with the same chance) is depicted as the dot on the line.

This vector representation allows us to associate to each pmf of X a norm, that is, a length. Our intuitive notion of length is based on the *Euclidean norm* and entails taking the root of the sums of squares. For instance, we can associate to the pmf $p_e(X)$ the length $\|p_e(X)\|_2 = \sqrt{(1/2)^2 + (1/2)^2} = 1/\sqrt{2} \approx 0.71$. On the other hand, the length of the pmf that states that $X = 1$ is generated with 100% chance has length one. Note that by eye, we conclude that $p_e(X)$, the arrow pointing to the dot in the left panel in Fig. 6 is indeed much shorter than the arrow pointing to extreme pmf $p(X) = [0, 1]$.

This mismatch in lengths can be avoided when we represent each pmf $p(X)$ by two times its square root instead (Kass, 1989), that is, by $m(X) = 2\sqrt{p(X)} = [2\sqrt{p(0)}, 2\sqrt{p(1)}]$.[9] A pmf that is identified as the vector $m(X)$ is now two units away from the origin, that is, $\|m(X)\|_2 = \sqrt{m(0)^2 + m(1)^2} = \sqrt{4(p(0) + p(1))} = 2$. For instance, the pmf $p_e(X)$ is now represented as $m_e(X) \approx [1.41, 1.41]$. The model space $\mathcal{M}$ is the collection of all transformed pmfs and represented as the surface of (the positive part of) a circle, see the right panel of Fig. 6.[10] By representing the set of all possible pmfs of X as vectors $m(X) = 2\sqrt{p(X)}$ that reside on the sphere $\mathcal{M}$, we adopted our intuitive notion of distance. As a result, we can now, by simply looking at the figures, clarify that a uniform prior on the parameter space may lead to a very informative prior in the model space $\mathcal{M}$.

[9] The factor two is used to avoid a scaling of a quarter, though, its precise value is not essential for the ideas conveyed here. To simplify matters, we also call $m(X)$ a pmf.

[10] Hence, the model space $\mathcal{M}$ is the collection of all functions on $\mathcal{X}$ such that (i) $m(x) \geq 0$ for every outcome x of X, and (ii) $\sqrt{m(0)^2 + m(1)^2} = 2$. This vector representation of all the pmfs on X has the advantage that it also induces an inner product, which allows one to project one vector onto another, see Rudin (1991, p. 4), van der Vaart (1998, p. 94) and the online appendix.

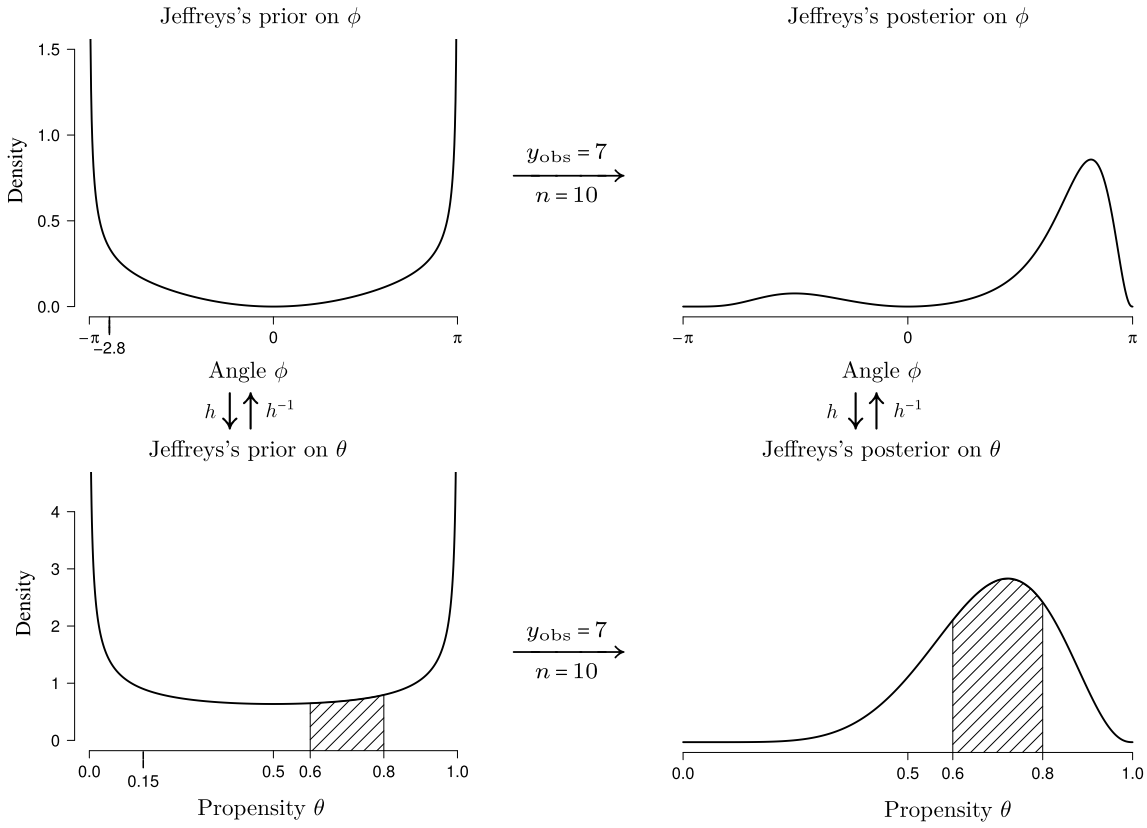

Fig. 5. For priors constructed through Jeffreys's rule it does not matter whether the problem is represented in terms of the angles ϕ or its propensity θ. Thus, not only is the problem equivalent due to the transformations $\theta = h(\phi)$ and its backwards transformation $\phi = h^{-1}(\theta)$, the prior information is the same in both representations. This also holds for the posteriors.

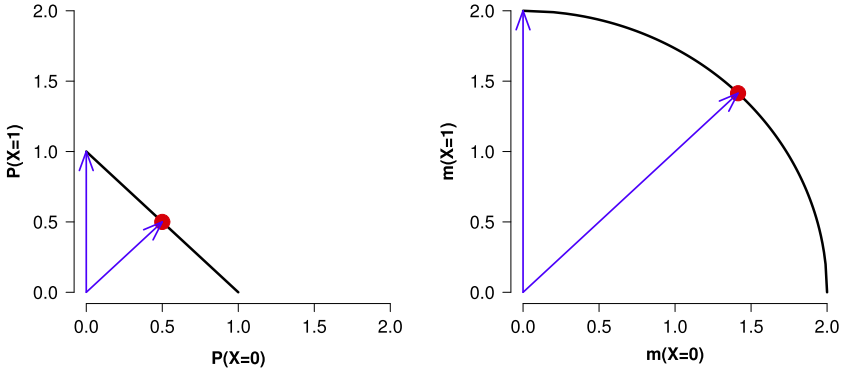

Fig. 6. The true pmf of X with the two outcomes $\{0, 1\}$ has to lie on the line (left panel) or more naturally on the positive part of the circle (right panel). The dot represents the pmf $p_e(X)$.

3.4.2. Uniform on the parameter space versus uniform on the model space

As $\mathcal{M}$ represents the largest set of pmfs, any model defines a subset of $\mathcal{M}$. Recall that the function $f(x \mid \theta)$ represents how we believe a parameter θ is functionally related to an outcome x of X. For each θ this parameterization yields a pmf $p_\theta(X)$ and, thus, also the vector $m_\theta(X) = 2\sqrt{p_\theta(X)}$. We denote the resulting set of vectors $m_\theta(X)$ so created by $\mathcal{M}_\Theta$. For instance, the Bernoulli model $f(x \mid \theta) = \theta^x(1-\theta)^{1-x}$ consists of pmfs given by $p_\theta(X) = [f(0 \mid \theta), f(1 \mid \theta)] = [1-\theta, \theta]$, which we represent as the vectors $m_\theta(X) = [2\sqrt{1-\theta}, 2\sqrt{\theta}]$. Doing this for every θ in the parameter space Θ yields the candidate set of pmfs $\mathcal{M}_\Theta$. In this case, we obtain a saturated model, since $\mathcal{M}_\Theta = \mathcal{M}$, see the left panel in Fig. 7, where the right most square on the curve corresponds to $m_0(X) = [2, 0]$. By following the curve in an anti-clockwise manner we encounter squares that represent the pmfs $m_\theta(X)$ corresponding to $\theta = 0.1, 0.2, \ldots, 1.0$ respectively. In the right panel of Fig. 7 the same procedure is repeated, but this time in terms of ϕ at $\phi = -1.0\pi, -0.8\pi, \ldots, 1.0\pi$. Indeed, filling in the gaps shows that the Bernoulli model in terms of θ and ϕ fully overlap with the largest set of possible pmfs, thus, $\mathcal{M}_\Theta = \mathcal{M} = \mathcal{M}_\Phi$. Fig. 7 makes precise what is meant when we say that the models $\mathcal{M}_\Theta$ and $\mathcal{M}_\Phi$ are equivalent; the two models define the same candidate set of pmfs that we believe to be viable data generating devices for X.

However, θ and ϕ represent $\mathcal{M}$ in a substantially different manner. As the representation $m(X) = 2\sqrt{p(X)}$ respects our natural notion of distance, we conclude, by eye, that a uniform division of θs with distance, say, $d\theta = 0.1$ does not lead to a uniform

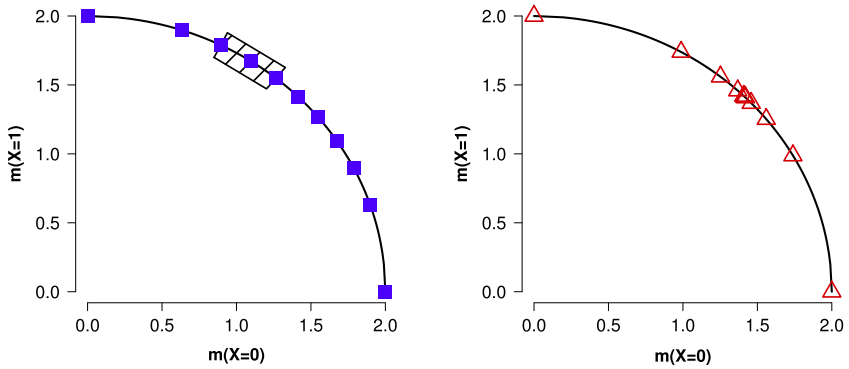

Fig. 7. The parameterization in terms of propensity θ (left panel) and angle ϕ (right panel) differ from each other substantially, and from a uniform prior in the model space. Left panel: The eleven squares (starting from the right bottom going anti-clockwise) represent pmfs that correspond to $\theta = 0.0, 0.1, 0.2, \ldots, 0.9, 1.0$. The shaded area corresponds to the shaded area in the bottom-left panel of Fig. 5 and accounts for 14% of the model's length. Right panel: Similarly, the eleven triangles (starting from the right bottom going anti-clockwise) represent pmfs that correspond to $\phi = -1.0\pi, -0.8\pi, -\cdots 0.8\pi, 1.0\pi$.

partition of the model. More extremely, a uniform division of ϕ with distance $\mathrm{d}\phi = 0.2\pi$ (10% of the length of the parameter space) also does not lead to a uniform partition of the model. In particular, even though the intervals $(-\pi, -0.8\pi)$ and $(-0.2\pi, 0)$ are of equal length in the parameter space Φ, they do not have an equal displacement in the model $\mathcal{M}_\Phi$. In effect, the right panel of Fig. 7 shows that the 10% probability that the uniform prior on ϕ assigns to $\phi^* \in (-\pi, -0.8\pi)$ in parameter space is redistributed over a larger arc length of the model $\mathcal{M}_\Phi$ compared to the 10% assigned to $\phi^* \in (-0.2\pi, 0)$. Thus, a uniform distribution on ϕ favors the pmfs $m_\phi(X)$ with ϕ close to zero. Note that this effect is canceled by the Jeffreys's prior, as it puts more mass near the end points compared to $\phi = 0$, see the top-left panel of Fig. 5. Similarly, the left panel of Fig. 7 shows that the uniform prior $g(\theta)$ also fails to yield an equiprobable assessment of the pmfs in model space. Again, the Jeffreys's prior in terms of θ compensates for the fact that the interval $(0, 0.1)$ as compared to $(0.5, 0.6)$ in Θ is more spread out in model space. However, it does so less severely compared to the Jeffreys's prior on ϕ. To illustrate, we added additional tick marks on the horizontal axis of the priors in the left panels of Fig. 5. The tick mark at $\phi = -2.8$ and $\theta = 0.15$ both indicate the 25% quantiles of their respective Jeffreys's priors. Hence, the Jeffreys's prior allocates more mass to the boundaries of ϕ than to the boundaries of θ to compensate for the difference in geometry, see Fig. 7. More generally, the Jeffreys's prior uses Fisher information to convert the geometry of the model to the parameter space.

Note that because the Jeffreys's prior is specified using the Fisher information, it takes the functional relationship $f(x \mid \theta)$ into account. The functional relationship makes precise how the parameter is linked to the data and, thus, gives meaning and context to the parameter. On the other hand, a prior on ϕ specified without taking the functional relationship $f(x \mid \phi)$ into account is a prior that neglects the context of the problem. For instance, the right panel of Fig. 7 shows that this neglect with a uniform prior on ϕ results in having the geometry of $\Phi = (-\pi, \pi)$ forced onto the model $\mathcal{M}_\Phi$.

3.5. Uniform prior on the model

Fig. 7 shows that neither a uniform prior on θ, nor a uniform prior on ϕ yields a uniform prior on the model. Alternatively, we can begin with a uniform prior on the model $\mathcal{M}$ and convert this into priors on the parameter spaces Θ and Φ. This uniform prior on the model translated to the parameters is exactly the Jeffreys's prior.

Recall that a prior on a space S is uniform, if it has the following two defining features: (i) the prior is proportional to one, and (ii) a normalizing constant given by $V_S = \int_S 1\mathrm{d}s$ that equals the length, more generally, volume of S. For instance, a replacement of s by ϕ and S by $\Phi = (-\pi, \pi)$ yields the uniform prior on the angles with the normalizing constant $V_\Phi = \int_\Phi 1\mathrm{d}\phi = 2\pi$. Similarly, a replacement of s by the pmf $m_\theta(X)$ and S by the function space $\mathcal{M}_\Theta$ yields a uniform prior on the model $\mathcal{M}_\Theta$. The normalizing constant then becomes a daunting looking integral in terms of displacements $\mathrm{d}m_\theta(X)$ between vectors in model space $\mathcal{M}_\Theta$. Fortunately, it can be shown, see the online appendix, that V simplifies to

$$V = \int_{\mathcal{M}_\Theta} 1\mathrm{d}m_\theta(X) = \int_\Theta \sqrt{I_X(\theta)}\mathrm{d}\theta. \tag{19}$$

Thus, V can be computed in terms of θ by multiplying the distances $\mathrm{d}\theta$ in Θ by the root of the Fisher information. Heuristically, this means that the root of the Fisher information translates displacements $\mathrm{d}m_\theta(X)$ in the model $\mathcal{M}_\Theta$ to distances $\sqrt{I_X(\theta)}\mathrm{d}\theta$ in the parameter space Θ.

Recall from Example 8 that regardless of the parameterization, the normalizing constant of the Jeffreys's prior was π. To verify that this is indeed the length of the model, we use the fact that the circumference of a quarter circle with radius $r = 2$ can also be calculated as $V = (2\pi r)/4 = \pi$.

Given that the Jeffreys's prior corresponds to a uniform prior on the model, we deduce that the shaded area in the bottom-left panel of Fig. 5 with $P_J(\theta^* \in J_\theta) = 0.14$, implies that the model interval $J_m = \big(m_{0.6}(X), m_{0.8}(X)\big)$, the shaded area in the left panel of Fig. 7, accounts for 14% of the model's length. After updating the Jeffreys's prior with the observations x^n_{obs} consisting of $y_{\text{obs}} = 7$ out of $n = 10$ the probability of finding the true data generating pmf $m^*(X)$ in this interval of pmfs J_m is increased to 53%.

In conclusion, we verified that the Jeffreys's prior is a prior that leads to the same conclusion regardless of how we parameterize the problem. This parameterization-invariance property is a direct result of shifting our focus from finding the true parameter value within the parameter space to the proper formulation of the estimation problem—as discovering the true data generating pmf $m_{\theta^*}(X) = 2\sqrt{p_{\theta^*}(X)}$ in $\mathcal{M}_\Theta$ and by expressing our prior ignorance as a uniform prior on the model $\mathcal{M}_\Theta$.

4. The role of Fisher information in minimum description length

In this section we graphically show how Fisher information is used as a measure of model complexity and its role in model selection within the minimum description length framework (MDL; de Rooij & Grünwald, 2011; Grünwald, 2007; Grünwald, Myung, & Pitt, 2005; Myung, Forster, & Browne, 2000; Myung, Navarro, & Pitt, 2006; Pitt, Myung, & Zhang, 2002).

The primary aim of a model selection procedure is to select a single model from a set of competing models, say, models $\mathcal{M}_1$ and $\mathcal{M}_2$, that best suits the observed data x^n_{obs}. Many model selection procedures have been proposed in the literature, but the most popular methods are those based on penalized maximum likelihood criteria, such as the Akaike information criterion (AIC; Akaike, 1974; Burnham & Anderson, 2002), the Bayesian information criterion (BIC; Raftery, 1995; Schwarz, 1978), and the Fisher information approximation (FIA; Grünwald, 2007; Rissanen, 1996). These criteria are defined as follows

$$\text{AIC} = -2\log f_j\big(x^n_{\text{obs}} \mid \hat{\theta}_j(x^n_{\text{obs}})\big) + 2d_j, \tag{20}$$

$$\text{BIC} = -2\log f_j\big(x^n_{\text{obs}} \mid \hat{\theta}_j(x^n_{\text{obs}})\big) + d_j\log(n), \tag{21}$$

$$\text{FIA} = \underbrace{-\log f_j\big(x^n_{\text{obs}} \mid \hat{\theta}_j(x^n_{\text{obs}})\big)}_{\text{Goodness-of-fit}} + \underbrace{\frac{d_j}{2}\log\frac{n}{2\pi}}_{\text{Dimensionality}} + \underbrace{\log\left(\int_{\Theta}\sqrt{\det I_{\mathcal{M}_j}(\theta_j)}\,\mathrm{d}\theta_j\right)}_{\text{Geometric complexity}}, \tag{22}$$

where n denotes the sample size, d_j the number of free parameters, $\hat{\theta}_j$ the MLE, $I_{\mathcal{M}_j}(\theta_j)$ the unit Fisher information, and f_j the functional relationship between the potential outcome x^n and the parameters θ_j within model $\mathcal{M}_j$.[11] Hence, except for the observations x^n_{obs}, all quantities in the formulas depend on the model $\mathcal{M}_j$. We made this explicit using a subscript j to indicate that the quantity, say, θ_j belongs to model $\mathcal{M}_j$.[12] For all three criteria, the model yielding the lowest criterion value is perceived as the model that generalizes best (Myung & Pitt , in press).

Each of the three model selection criteria tries to strike a balance between model fit and model complexity. Model fit is expressed by the goodness-of-fit terms, which involves replacing the potential outcomes x^n and the unknown parameter θ_j of the functional relationships f_j by the actually observed data x^n_{obs}, as in the Bayesian setting, and the maximum likelihood estimate $\hat{\theta}_j(x^n_{\text{obs}})$, as in the frequentist setting.

The positive terms in the criteria account for model complexity. A penalization of model complexity is necessary, because the support in the data cannot be assessed by solely considering goodness-of-fit, as the ability to fit observations increases with model complexity (e.g., Roberts & Pashler, 2000). As a result, the more complex model necessarily leads to better fits but may in fact overfit the data. The overly complex model then captures idiosyncratic noise rather than general structure, resulting in poor model generalizability (Myung, Forster et al., 2000; Wagenmakers & Waldorp, 2006).

The focus in this section is to make intuitive how FIA acknowledges the trade-off between goodness-of-fit and model complexity in a principled manner by graphically illustrating this model selection procedure, see also Balasubramanian (1996), Kass (1989), Klaassen and Lenstra (2003), Myung, Balasubramanian, and Pitt (2000), and Rissanen (1996). We exemplify the concepts with simple multinomial processing tree (MPT) models (e.g., Batchelder and Riefer, 1999; Klauer and Kellen, 2011; Wu, Myung, and Batchelder, 2010). For a more detailed treatment of the subject we refer to the online appendix, de Rooij and Grünwald (2011), Grünwald (2007), Myung et al. (2006), and the references therein.

[11] For vector-valued parameters θ_j, we have a Fisher information matrix and $\det I_{\mathcal{M}_j}(\theta_j)$ refers to the determinant of this matrix. This determinant is always non-negative, because the Fisher information matrix is always a positive semidefinite symmetric matrix. Intuitively, volumes and areas cannot be negative, see also the online appendix.

[12] For the sake of clarity, we will use different notations for the parameters within the different models. We introduce two models in this section: the model $\mathcal{M}_1$ with parameter $\theta_1 = \vartheta$ which we pit against the model $\mathcal{M}_2$ with parameter $\theta_2 = \alpha$.

4.0.1. The description length of a model

Recall that each model specifies a functional relationship f_j between the potential outcomes of X and the parameters θ_j. This f_j is used to define a so-called *normalized maximum likelihood* (NML) code. For the jth model its NML code is defined as

$$p_{\text{NML}}(x^n_{\text{obs}} \mid \mathcal{M}_j) = \frac{f_j(x^n_{\text{obs}} \mid \hat{\theta}_j(x^n_{\text{obs}}))}{\sum_{x^n\in\mathcal{X}^n} f_j(x^n \mid \hat{\theta}_j(x^n))}, \tag{23}$$

where the sum in the denominator is over all possible outcomes x^n in $\mathcal{X}^n$, and where $\hat{\theta}_j$ refers to the MLE within model $\mathcal{M}_j$. The NML code is a relative goodness-of-fit measure, as it compares the observed goodness-of-fit term against the sum of all possible goodness-of-fit terms. Note that the actual observations x^n_{obs} only affect the numerator, by a plugin of x^n_{obs} and its associated maximum likelihood estimate $\hat{\theta}(x^n_{\text{obs}})$ into the functional relationship f_j belonging to model $\mathcal{M}_j$. The sum in the denominator consists of the same plugins, but for every possible realization of X^n.[13] Hence, the denominator can be interpreted as a measure of the model's collective goodness-of-fit or the model's fit capacity. Consequently, for every set of observations x^n_{obs}, the NML code outputs a number between zero and one that can be transformed into a non-negative number by taking the negative logarithm as[14]

$$-\log p_{\text{NML}}(x^n_{\text{obs}} \mid \mathcal{M}_j) = -\log f_j\big(x^n_{\text{obs}} \mid \hat{\theta}_j(x^n_{\text{obs}})\big) + \underbrace{\log\sum f_j(x^n \mid \hat{\theta}_j(x^n))}_{\text{Model complexity}}, \tag{24}$$

which is called the description length of model $\mathcal{M}_j$. Within the MDL framework, the model with the shortest description length is the model that best describes the observed data x^n_{obs}.

The model complexity term is typically hard to compute, but Rissanen (1996) showed that it can be well-approximated by the dimensionality and the geometrical complexity terms. That is,

$$\text{FIA} = -\log f_j\big(x^n_{\text{obs}} \mid \hat{\theta}_j(x^n_{\text{obs}})\big) + \frac{d_j}{2}\log\frac{n}{2\pi} + \log\left(\int_{\Theta}\sqrt{\det I_{\mathcal{M}_j}(\theta_j)}\,\mathrm{d}\theta_j\right),$$

is an approximation of the description length of model $\mathcal{M}_j$. The determinant is simply the absolute value when the number of free parameters d_j is equal to one. Furthermore, the integral in the geometrical complexity term coincides with the normalizing constant of the Jeffreys's prior, which represented the volume of the model. In other words, a model's fit capacity is proportional to its volume in model space as one would expect.

In sum, within the MDL philosophy, a model is selected if it yields the shortest description length, as this model uses the functional relationship f_j that best extracts the regularities from x^n_{obs}. As the description length is often hard to compute, we approximate it with FIA instead (Heck, Moshagen, & Erdfelder, 2014). To do so, we have to characterize (1) all possible outcomes of X, (2) propose at least two models which will be pitted against each other, (3) identify the model characteristics: the MLE $\hat{\theta}_j$ corresponding to $\mathcal{M}_j$, and its volume $V_{\mathcal{M}_j}$. In the remainder of this section we show that FIA selects the model that is closest to the data with an additional penalty for model complexity.

[13] As before, for continuous data, the sum is replaced by an integral.

[14] Quite deceivingly the minus sign actually makes this measure positive, as $-\log(y) = \log(1/y) \geq 0$ if $0 \leq y \leq 1$.

4.1. A new running example and the geometry of a random variable with $w = 3$ outcomes

To graphically illustrate the model selection procedure underlying MDL we introduce a random variable X that has $w = 3$ number of potential outcomes.

Example 9 (*A Psychological Task with Three Outcomes*)**.** In the training phase of a source-memory task, the participant is presented with two lists of words on a computer screen. List $\mathcal{L}$ is projected on the left-hand side and list $\mathcal{R}$ is projected on the right-hand side. In the test phase, the participant is presented with two words, side by side, that can stem from either list, thus, *ll*, *lr*, *rl*, *rr*. At each trial, the participant is asked to categorize these pairs as either:

- L meaning both words come from the left list, i.e., "*ll*" ,
- M meaning the words are mixed, i.e., "*lr*" or "*rl*",
- R meaning both words come from the right list, i.e., "*rr*".

For simplicity we assume that the participant will be presented with n test pairs X^n of equal difficulty. ◇

For the graphical illustration of this new running example, we generalize the ideas presented in Section 3.4.1 from $w = 2$ to $w = 3$ dimensions. Recall that a pmf of X with w number of outcomes can be written as a w-dimensional vector. For the task described above we know that a data generating pmf defines the three chances $p(X) = [p(L), p(M), p(R)]$ with which X generates the outcomes $[L, M, R]$ respectively.[15] As chances cannot be negative, (i) we require that $0 \leq p(x) = P(X = x)$ for every outcome x in $\mathcal{X}$, and (ii) to explicitly convey that there are $w = 3$ outcomes, and none more, these $w = 3$ chances have to sum to one, that is, $\sum_{x\in\mathcal{X}} p(x) = 1$. We call the largest set of functions that adhere to conditions (i) and (ii) the complete set of pmfs $\mathcal{P}$. The three chances with which a pmf $p(X)$ generates outcomes of X can be simultaneously represented in three-dimensional space with $p(L) = P(X = L)$ on the left most axis, $p(M) = P(X = M)$ on the right most axis and $p(R) = P(X = R)$ on the vertical axis as shown in the left panel of Fig. 8.[16] In the most extreme case, we have the pmfs $p(X) = [1, 0, 0]$, $p(X) = [0, 1, 0]$ or $p(X) = [0, 0, 1]$, which correspond to the corners of the triangle indicated by *pL*, *pM* and *pR*, respectively. These three extremes are linked by a triangular plane in the left panel of Fig. 8. *Any* pmf – and the true pmf $p^*(X)$ in particular – can be uniquely identified with a vector on the triangular plane and vice versa. For instance, a possible true pmf of X is $p_e(X) = [1/3, 1/3, 1/3]$ (i.e., the outcomes L, M and R are generated with the same chance) and depicted as a (red) dot on the simplex.

This vector representation allows us to associate to each pmf of X the Euclidean norm. For instance, the representation in the left panel of Fig. 8 leads to an extreme pmf $p(X) = [1, 0, 0]$ that is one unit long, while $p_e(X) = [1/3, 1/3, 1/3]$ is only $\sqrt{(1/3)^2 + (1/3)^2 + (1/3)^2} \approx 0.58$ units away from the origin. As before, we can avoid this mismatch in lengths by considering the vectors $m(X) = 2\sqrt{p(X)}$, instead. Any pmf that is identified as $m(X)$ is now two units away from the origin. The model space $\mathcal{M}$ is the collection of all transformed pmfs and represented as the surface of (the positive part of) the sphere in the right panel of Fig. 8. By representing the set of possible pmfs of X as $m(X) = 2\sqrt{p(X)}$, we adopted our intuitive notion of distance. As a result, the selection mechanism underlying MDL can be made intuitive by simply looking at the forthcoming plots.

[15] As before we write $p(X) = [p(L), p(M), p(R)]$ with a capital X to denote all the w number of chances simultaneously and we used the shorthand notation $p(L) = p(X = L)$, $p(M) = p(X = M)$ and $p(R) = p(X = R)$.

[16] This is the three-dimensional generalization of Fig. 6.

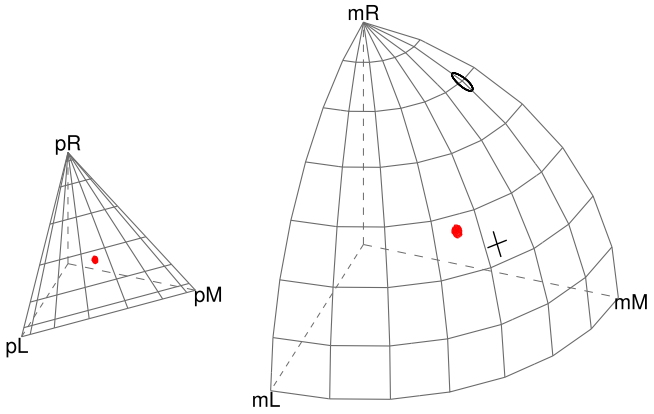

Fig. 8. Every point on the sphere corresponds to a pmf of a categorical distribution with $w = 3$ categories. In particular, the dot refers to the pmf $p_e(x) = [1/3, 1/3, 1/3]$, the circle represents the pmf given by $p(X) = [0.01, 0.18, 0.81]$, while the cross represents the pmf $p(X) = [0.25, 0.5, 0.25]$.

4.2. The individual-word and the only-mixed strategy

To ease the exposition, we assume that both words presented to the participant come from the right list $\mathcal{R}$, thus, "*rr*" for the two models introduced below. As model $\mathcal{M}_1$ we take the so-called individual-word strategy. Within this model $\mathcal{M}_1$, the parameter is $\theta_1 = \vartheta$, which we interpret as the participant's "right-list recognition ability". With chance ϑ the participant then correctly recognizes that the first word originates from the right list and repeats this procedure for the second word, after which the participant categorizes the word pair as L, M, or R, see the left panel of Fig. 9 for a schematic description of this strategy as a processing tree. Fixing the participant's "right-list recognition ability" ϑ yields the following pmf

$$f_1(X \mid \vartheta) = \left[(1-\vartheta)^2, 2\vartheta(1-\vartheta), \vartheta^2\right]. \tag{25}$$

For instance, when the participant's true ability is $\vartheta^* = 0.9$, the three outcomes $[L, M, R]$ are then generated with the following three chances $f_1(X \mid 0.9) = [0.01, 0.18, 0.81]$, which is plotted as a circle in Fig. 8. On the other hand, when $\vartheta^* = 0.5$ the participant's generating pmf is then $f_1(X \mid \vartheta = 0.5) = [0.25, 0.5, 0.25]$, which is depicted as the cross in model space $\mathcal{M}$. The set of pmfs so defined forms a curve that goes through both the cross and the circle, see the left panel of Fig. 10.

As a competing model $\mathcal{M}_2$, we take the so-called only-mixed strategy. For the task described in Example 9, we might pose that participants from a certain clinical group are only capable of recognizing mixed word pairs and that they are unable to distinguish the pairs "*rr*" from "*ll*" resulting in a random guess between the responses L and R, see the right panel of Fig. 9 for the processing tree. Within this model $\mathcal{M}_2$ the parameter is $\theta_2 = \alpha$, which is interpreted as the participant's "mixed-list differentiability skill" and fixing it yields the following pmf

$$f_2(X \mid \alpha) = \left[(1-\alpha)/2, \alpha, (1-\alpha)/2\right]. \tag{26}$$

For instance, when the participant's true differentiability is $\alpha^* = 1/3$, the three outcomes $[L, M, R]$ are then generated with the same chance $f_2(X \mid 1/3) = [1/3, 1/3, 1/3]$, which, as before, is plotted as the dot in Fig. 10. On the other hand, when $\alpha^* = 0.5$ the participant's generating pmf is then given by $f_2(X \mid \alpha = 0.5) = [0.25, 0.5, 0.25]$, i.e., the cross. The set of pmfs so defined forms a curve that goes through both the dot and the cross, see the left panel of Fig. 10.

The plots show that the models $\mathcal{M}_1$ and $\mathcal{M}_2$ are neither saturated nor nested, as the two models define proper subsets of $\mathcal{M}$ and

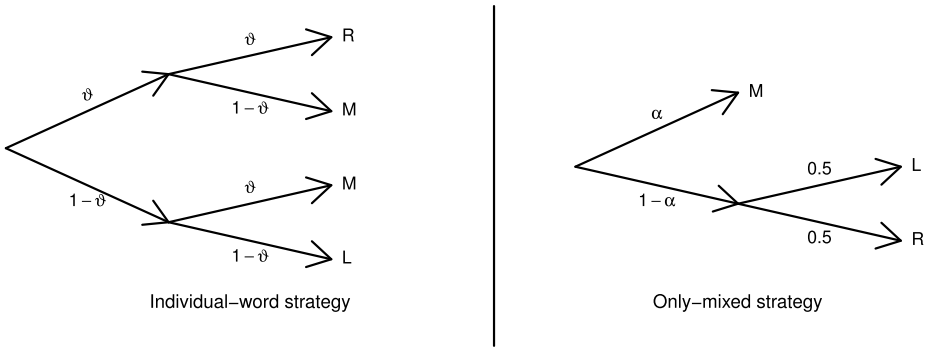

Fig. 9. Two MPT models that theorize how a participant chooses the outcomes L, M, or R in the source-memory task described in the main text. The left panel schematically describes the individual-word strategy, while the right model schematically describes the only-mixed strategy.

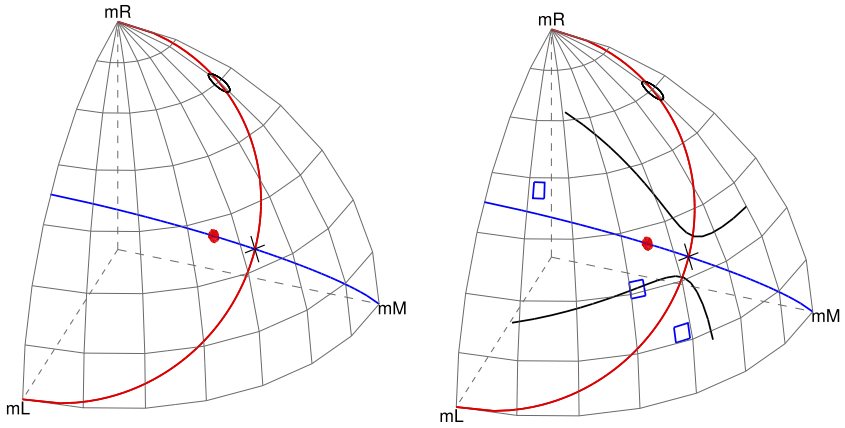

Fig. 10. Left panel: The set of pmfs that are defined by the individual-list strategy $\mathcal{M}_1$ form a curve that goes through both the cross and the circle, while the pmfs of the only-mixed strategy $\mathcal{M}_2$ correspond to the curve that goes through both the cross and the dot. Right panel: The model selected by FIA can be thought of as the model closest to the empirical pmf with an additional penalty for model complexity. The selection between the individual-list and the only-mixed strategy by FIA based on $n = 30$ trials is formalized by the additional curves—the only-mixed strategy is preferred over the individual-list strategy, when the observations yield an empirical pmf that lies between the two non-decision curves. The top, middle and bottom squares corresponding to the data sets $x^n_{\text{obs},1}$, $x^n_{\text{obs},2}$ and $x^n_{\text{obs},3}$ in Table 1, which are best suited to $\mathcal{M}_2$, either, and $\mathcal{M}_1$, respectively. The additional penalty is most noticeable at the cross, where the two models share a pmf. Observations with $n = 30$ yielding an empirical pmf in this area are automatically assigned to the simpler model, i.e., the only-mixed strategy $\mathcal{M}_2$.

only overlap at the cross. Furthermore, the plots also show that $\mathcal{M}_1$ and $\mathcal{M}_2$ are both one-dimensional, as each model is represented as a line in model space. Hence, the dimensionality terms in all three information criteria are the same. Consequently, AIC and BIC will only discriminate these two models based on goodness-of-fit alone. This particular model comparison, thus, allows us to highlight the role Fisher information plays in the MDL model selection philosophy.

4.3. Model characteristics

4.3.1. The maximum likelihood estimators

For FIA we need to compute the goodness-of-fit terms, thus, we need to identify the MLEs for the parameters within each model. For the models at hand, the MLEs are

$$\begin{aligned} \hat{\theta}_1 &= \hat{\vartheta} = (Y_M + 2Y_R)/(2n) \quad \text{for } \mathcal{M}_1, \quad \text{and} \\ \hat{\theta}_2 &= \hat{\alpha} = Y_M/n \quad \text{for } \mathcal{M}_2, \end{aligned} \tag{27}$$

where Y_L, Y_M and $Y_R = n - Y_L - Y_M$ are the number of L, M and R responses in the data consisting of n trials.

Estimation is a within model operation and it can be viewed as projecting the so-called *empirical (i.e., observed) pmf* corresponding to the data onto the model. For i.i.d. data with $w = 3$ outcomes the empirical pmf corresponding to x^n_{obs} is defined as $\hat{p}_{\text{obs}}(X) = [y_L/n, y_M/n, y_R/n]$. Hence, the empirical pmf gives the relative occurrence of each outcome in the sample. For instance, the observations x^n_{obs} consisting of $[y_L = 3, y_M = 3, y_R = 3]$ responses correspond to the observed pmf $\hat{p}_{\text{obs}}(X) = [1/3, 1/3, 1/3]$, i.e., the dot in Fig. 10. Note that this observed pmf $\hat{p}_{\text{obs}}(X)$ does not reside on the curve of $\mathcal{M}_1$.

Nonetheless, when we use the MLE $\hat{\vartheta}$ of $\mathcal{M}_1$, we as researchers bestow the participant with a "right-list recognition ability" ϑ and implicitly assume that she used the individual-word strategy to generate the observations. In other words, we only consider the pmfs on the curve of $\mathcal{M}_1$ as viable explanations of how the participant generated her responses. For the data at hand, we have the estimate $\hat{\vartheta}_{\text{obs}} = 0.5$. If we were to generalize the observations x^n_{obs} under $\mathcal{M}_1$, we would then plug this estimate into the functional relationship f_1 resulting in the predictive pmf $f_1(X \mid \hat{\vartheta}_{\text{obs}}) = [0.25, 0.5, 0.25]$. Hence, even though the number of L, M and R responses were equal in the observations x^n_{obs}, under $\mathcal{M}_1$ we expect that this participant will answer with twice as many M responses compared to the L and R responses in a next set of test items. Thus, for predictions, part of the data is ignored and considered as noise.

Geometrically, the generalization $f_1(X \mid \hat{\vartheta}_{\text{obs}})$ is a result of projecting the observed pmf $\hat{p}_{\text{obs}}(X)$, i.e., the dot, onto the cross that does reside on the curve of $\mathcal{M}_1$.[17] Observe that amongst all pmfs on $\mathcal{M}_1$, the projected pmf is closest to the empirical pmf $\hat{p}_{\text{obs}}(X)$. Under $\mathcal{M}_1$ the projected pmf $f_1(X \mid \hat{\vartheta}_{\text{obs}})$, i.e., the cross,

[17] This resulting pmf $f_1(X \mid \hat{\vartheta}_{\text{obs}})$ is also known as the Kullback–Leibler projection of the empirical pmf $\hat{p}_{\text{obs}}(X)$ onto the model $\mathcal{M}_1$. White (1982) used this projection to study the behavior of the MLE under model misspecification.

Table 1
The description lengths for three observations $x^n_{\text{obs}} = [y_L, y_M, y_R]$, where y_L, y_M, y_R are the number of observed responses L, M and R respectively.

$x^n_{\text{obs}} = [y_L, y_M, y_R]$	$\text{FIA}_{\mathcal{M}_1}(x^n_{\text{obs}})$	$\text{FIA}_{\mathcal{M}_2}(x^n_{\text{obs}})$	Preferred model
$x^n_{\text{obs},1} = [12, 1, 17]$	42	26	$\mathcal{M}_2$
$x^n_{\text{obs},2} = [14, 10, 6]$	34	34	Tie
$x^n_{\text{obs},3} = [12, 16, 2]$	29	32	$\mathcal{M}_1$

is perceived as structural, while any deviations from the curve of $\mathcal{M}_1$ is labeled as noise. When generalizing the observations, we ignore noise. Hence, by estimating the parameter ϑ, we implicitly restrict our predictions to only those pmfs that are defined by $\mathcal{M}_1$. Moreover, evaluating the prediction at x^n_{obs} and, subsequently, taking the negative logarithm yields the goodness-of-fit term; in this case, $-\log f_1(x^n_{\text{obs}} \mid \hat{\vartheta}_{\text{obs}} = 0.5) = 10.4$.

Which part of the data is perceived as structural or as noise depends on the model. For instance, when we use the MLE $\hat{\alpha}$, we restrict our predictions to the pmfs of $\mathcal{M}_2$. For the data at hand, we get $\hat{\alpha}_{\text{obs}} = 1/3$ and the plugin yields $f_2(X \mid \hat{\alpha}_{\text{obs}}) = [1/3, 1/3, 1/3]$. Again, amongst all pmfs on $\mathcal{M}_2$, the projected pmf is closest to the empirical pmf $\hat{p}_{\text{obs}}(X)$. In this case, the generalization under $\mathcal{M}_2$ coincides with the observed pmf $\hat{p}_{\text{obs}}(X)$. Hence, under $\mathcal{M}_2$ there is no noise, as the empirical pmf $\hat{p}_{\text{obs}}(X)$ was already on the model. Geometrically, this means that $\mathcal{M}_2$ is closer to the empirical pmf than $\mathcal{M}_1$, which results in a lower goodness-of-fit term $-\log f_2(x^n_{\text{obs}} \mid \hat{\alpha}_{\text{obs}} = 1/3) = 9.9$.

This geometric interpretation allows us to make intuitive that data sets with the same goodness-of-fit terms will be as far from $\mathcal{M}_1$ as from $\mathcal{M}_2$. Equivalently, $\mathcal{M}_1$ and $\mathcal{M}_2$ identify the same amount of noise within x^n_{obs}, when the two models fit the observations equally well. For instance, Fig. 10 shows that observations x^n_{obs} with an empirical pmf $\hat{p}_{\text{obs}}(X) = [0.25, 0.5, 0.25]$ are equally far from $\mathcal{M}_1$ as from $\mathcal{M}_2$. Note that the closest pmf on $\mathcal{M}_1$ and $\mathcal{M}_2$ are both equal to the empirical pmf, as $f_1(X \mid \hat{\vartheta}_{\text{obs}} = 0.5) = \hat{p}_{\text{obs}}(X) = f_2(X \mid \hat{\alpha}_{\text{obs}} = 1/2)$. As a result, the two goodness-of-fit terms will be equal to each other.

In sum, goodness-of-fit measures a model's proximity to the observed data. Consequently, models that take up more volume in model space will be able to be closer to a larger number of data sets. In particular, when, say, $\mathcal{M}_3$ is nested within $\mathcal{M}_4$, this means that the distance between $\hat{p}_{\text{obs}}(X)$ and $\mathcal{M}_3$ (noise) is at least the distance between $\hat{p}_{\text{obs}}(X)$ and $\mathcal{M}_4$. Equivalently, for any data set, $\mathcal{M}_4$ will automatically label more of the observations as structural. Models that excessively identify parts of the observations as structural are known to overfit the data. Overfitting has an adverse effect on generalizability, especially when n is small, as $\hat{p}_{\text{obs}}(X)$ is then dominated by sampling error. In effect, the more voluminous model will then use this sampling error, rather than the structure, for its predictions. To guard ourselves from overfitting, thus, bad generalizability, the information criteria AIC, BIC and FIA all penalize for model complexity. AIC and BIC only do this via the dimensionality terms, while FIA also take the models' volumes into account.

4.3.2. Geometrical complexity

For both models the dimensionality term is given by $\frac{1}{2}\log(\frac{n}{2\pi})$. Recall that the geometrical complexity term is the logarithm of the model's volume, which for the individual-word and the only-mixed strategy are given by

$$V_{\mathcal{M}_1} = \int_0^1 \sqrt{I_{\mathcal{M}_1}(\theta)}\mathrm{d}\theta = \sqrt{2}\pi \quad \text{and} \qquad V_{\mathcal{M}_2} = \int_0^1 \sqrt{I_{\mathcal{M}_2}(\alpha)}\mathrm{d}\alpha = \pi, \tag{28}$$

respectively. Hence, the individual-word strategy is a more complex model, because it has a larger volume, thus, capacity to fit data compared to the only-mixed strategy. After taking logs, we see that the individual-word strategy incurs an additional penalty of $1/2\log(2)$ compared to the only-mixed strategy.

4.4. Model selection based on the minimum description length principle

With all model characteristics at hand, we only need observations to illustrate that MDL model selection boils down to selecting the model that is closest to the observations with an additional penalty for model complexity. Table 1 shows three data sets $x^n_{\text{obs},1}, x^n_{\text{obs},2}, x^n_{\text{obs},3}$ with $n = 30$ observations. The three associated empirical pmfs are plotted as the top, middle and lower rectangles in the right panel of Fig. 10, respectively. Table 1 also shows the approximation of each model's description length using FIA. Note that the first observed pmf, the top rectangle in Fig. 10, is closer to $\mathcal{M}_2$ than to $\mathcal{M}_1$, while the third empirical pmf, the lower rectangle, is closer to $\mathcal{M}_1$. Of particular interest is the middle rectangle, which lies on an additional black curve that we refer to as a non-decision curve; observations that correspond to an empirical pmf that lies on this curve are described equally well by $\mathcal{M}_1$ and $\mathcal{M}_2$. For this specific comparison, we have the following decision rule: FIA selects $\mathcal{M}_2$ as the preferred model whenever the observations correspond to an empirical pmf between the two non-decision curves, otherwise, FIA selects $\mathcal{M}_1$. Fig. 10 shows that FIA, indeed, selects the model that is closest to the data except in the area where the two models overlap—observations consisting of $n = 30$ trials with an empirical pmf near the cross are considered better described by the simpler model $\mathcal{M}_2$. Hence, this yields an incorrect decision even when the empirical pmf is exactly equal to the true data generating pmf that is given by, say, $f_1(X \mid \vartheta = 0.51)$. This automatic preference for the simpler model, however, decreases as n increases. The left and right panel of Fig. 11 show the non-decision curves when $n = 120$ and n (extremely) large, respectively. As a result of moving non-decision bounds, the data set $x^n_{\text{obs},4} = [56, 40, 24]$ that has the same observed pmf as $x^n_{\text{obs},2}$, i.e., the middle rectangle, will now be better described by model $\mathcal{M}_1$.

For (extremely) large n, the additional penalty due to $\mathcal{M}_1$ being more voluptuous than $\mathcal{M}_2$ becomes irrelevant and the sphere is then separated into quadrants: observations corresponding to an empirical pmf in the top-left or bottom-right quadrant are better suited to the only-mixed strategy, while the top-right and bottom-left quadrants indicate a preference for the individual-word strategy $\mathcal{M}_1$. Note that pmfs on the non-decision curves in the right panel of Fig. 11 are as far apart from $\mathcal{M}_1$ as from $\mathcal{M}_2$, which agrees with our geometric interpretation of goodness-of-fit as a measure of the model's proximity to the data. This quadrant division is only based on the two models' goodness-of-fit terms and yields the same selection as one would get from BIC (e.g., Rissanen, 1996). For large n, FIA, thus, selects the model that is closest to the empirical pmf. This behavior is desirable, because asymptotically the empirical pmf is not distinguishable from the true data generating pmf. As such, the model that is closest to the empirical pmf will then also be closest to the true pmf. Hence, FIA asymptotically selects the model that is closest to the true pmf. As a result, the projected pmf within the closest model is then expected to yield the best predictions amongst the competing models.

4.5. Fisher information and generalizability

Model selection by MDL is sometimes perceived as a formalization of Occam's razor (e.g., Balasubramanian, 1996; Grünwald, 1998), a principle that states that the most parsimonious model should be chosen when the models under consideration fit the observed data equally well. This preference for the parsimonious

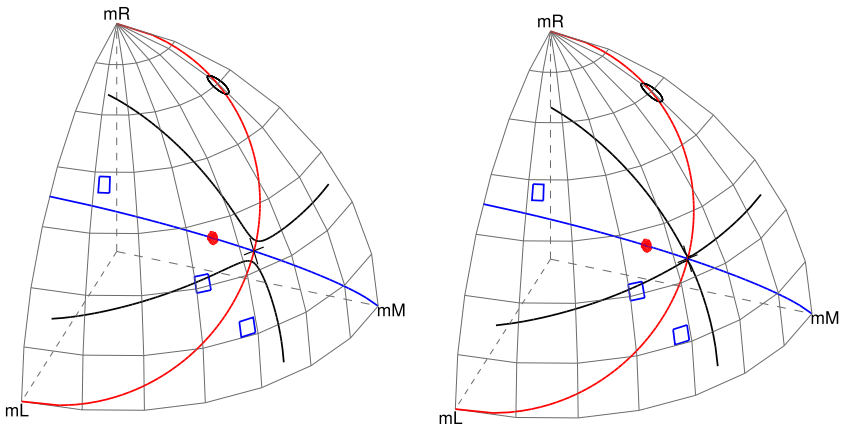

Fig. 11. For n large the additional penalty for model complexity becomes irrelevant. The plotted non-decision curves are based on $n = 120$ and $n = 10{,}000$ trials in the left and right panel respectively. In the right panel only the goodness-of-fit matters in the model comparison. The model selected is then the model that is closest to the observations.

model is based on the belief that the simpler model is better at predicting new (as yet unseen) data coming from the same source, as was shown by Pitt et al. (2002) with simulated data.

To make intuitive why the more parsimonious model, on average, leads to better predictions, we assume, for simplicity, that the true data generating pmf is given by $f(X \mid \theta^*)$, thus, the existence of a true parameter value θ^*. As the observations are expected to be contaminated with sampling error, we also expect an estimation error, i.e., a distance $\mathrm{d}\theta$ between the maximum likelihood estimate $\hat{\theta}_{\text{obs}}$ and the true θ^*. Recall that in the construction of Jeffreys's prior Fisher information was used to convert displacement in model space to distances on parameter space. Conversely, Fisher information transforms the estimation error in parameter space to a generalization error in model space. Moreover, the larger the Fisher information at θ^* is, the more it will expand the estimation error into a displacement between the prediction $f(X \mid \hat{\theta}_{\text{obs}})$ and the true pmf $f(X \mid \theta^*)$. Thus, a larger Fisher information at θ^* will push the prediction further from the true pmf resulting in a bad generalization. Smaller models have, on average, a smaller Fisher information at θ^* and will therefore lead to more stable predictions that are closer to the true data generating pmf. Note that the generalization scheme based on the MLE plugin $f(X \mid \hat{\theta}_{\text{obs}})$ ignores the error at each generalization step. The Bayesian counterpart, on the other hand, does take these errors into account, see Dawid (2011), Ly, Etz, Marsman, and Wagenmakers (2017), Marsman, Ly, and Wagenmakers (2016) and see Erven, Grünwald, and De Rooij (2012), Grünwald and Mehta (2016), van der Pas and Grünwald (2014), and Wagenmakers, Grünwald, and Steyvers (2006) for a prequential view of generalizability.

5. Concluding comments

Fisher information is a central statistical concept that is of considerable relevance for mathematical psychologists. We illustrated the use of Fisher information in three different statistical paradigms: in the frequentist paradigm, Fisher information was used to construct hypothesis tests and confidence intervals; in the Bayesian paradigm, Fisher information was used to specify a default, parameterization-invariant prior distribution; finally, in the paradigm of information theory, data compression, and minimum description length, Fisher information was used to measure model complexity. Note that these three paradigms highlight three uses of the functional relationship f between potential observations x^n and the parameters θ. Firstly, in the frequentist setting, the second argument was fixed at a supposedly known parameter value θ_0 or $\hat{\theta}_{\text{obs}}$ resulting in a probability mass function, a function of the potential outcomes $f(\cdot \mid \theta_0)$. Secondly, in the Bayesian setting, the first argument was fixed at the observed data resulting in a likelihood function, a function of the parameters $f(x_{\text{obs}} \mid \cdot)$. Finally, in the information geometric setting both arguments were free to vary, i.e., $f(\cdot \mid \cdot)$ and plugged in by the observed data and the maximum likelihood estimate.

To ease the exposition we only considered Fisher information of one-dimensional parameters. The generalization of the concepts introduced here to vector valued θ can be found in the online appendix (https://osf.io/hxxsj/). A complete treatment of all the uses of Fisher information throughout statistics would require a book (e.g., Frieden, 2004) rather than a tutorial article. Due to the vastness of the subject, the present account is by no means comprehensive. Our goal was to use concrete examples to provide more insight about Fisher information, something that may benefit psychologists who propose, develop, and compare mathematical models for psychological processes. Other uses of Fisher information are in the detection of model misspecification (Golden, 1995, 2000; Waldorp, 2009; Waldorp, Christoffels, and van de Ven, 2011; Waldorp, Huizenga, and Grasman, 2005; White, 1982), in the reconciliation of frequentist and Bayesian estimation methods through the Bernstein–von Mises theorem (Bickel and Kleijn, 2012; Rivoirard and Rousseau, 2012; van der Vaart, 1998; Yang and Le Cam, 2000), in statistical decision theory (e.g., Berger, 1985; Hájek, 1972; Korostelev and Korosteleva, 2011; Ray and Schmidt-Hieber, 2016; Wald, 1949), in the specification of objective priors for more complex models (e.g., Ghosal, Ghosh, and Ramamoorthi, 1997; Grazian and Robert, 2015; Kleijn and Zhao, 2013), and computational statistics and generalized MCMC sampling in particular (e.g., Banterle, Grazian, Lee, and Robert, 2015; Girolami and Calderhead, 2011; Grazian and Liseo, 2014; Gronau, Sarafoglou et al., 2017).

In sum, Fisher information is a key concept in statistical modeling. We hope to have provided an accessible and concrete tutorial article that explains the concept and some of its uses for applications that are of particular interest to mathematical psychologists.

Acknowledgments

This work was supported by the starting grant "Bayes or Bust" awarded by the European Research Council (283876). The authors would like to thank Jay Myung, Trisha Van Zandt, and three anonymous reviewers for their comments on an earlier version of this paper. The discussions with Helen Steingroever, Jean-Bernard Salomond, Fabian Dablander, Nishant Mehta, Alexander Etz, Quentin Gronau and Sacha Epskamp led to great improvements of the manuscript. Moreover, the first author is grateful to Chris Klaassen, Bas Kleijn and Henk Pijls for their patience and enthusiasm with which they taught, and answered questions from a not very docile student.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Aldrich, J. (2005). The statistical education of Harold Jeffreys. *International Statistical Review*, *73*(3), 289–307.

Balasubramanian, V. (1996). A geometric formulation of Occam's razor for inference of parametric distributions, arXiv Preprint Adap-Org/9601001.

Banterle, M., Grazian, C., Lee, A., & Robert, C. P. (2015). Accelerating Metropolis-Hastings algorithms by delayed acceptance, arXiv Preprint arXiv:1503.00996.

Batchelder, W. H., & Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review*, *87*, 375–397.

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

Bayarri, M., Berger, J., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, *40*(3), 1550–1577.

Berger, J. O. (1985). *Statistical decision theory and bayesian analysis*. Springer Verlag.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., & Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore.

Bickel, P. J., & Kleijn, B. J. K. (2012). The semiparametric Bernstein–von Mises Theorem. *The Annals of Statistics*, *40*(1), 206–237.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 101–117.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information–theoretic approach*. (second ed.). New York: Springer Verlag.

Chechile, R. A. (1973). The Relative Storage and Retrieval Losses in Short–Term Memory as a Function of the Similarity and Amount of Information Processing in the Interpolated Task. Ph.D. thesis University of Pittsburgh.

Dawid, A. P. (2011). Posterior model probabilities. In D. M. Gabbay, P. S. Bandyopadhyay, M. R. Forster, P. Thagard, & J. Woods (Eds.), *Handbook of the Philosophy of Science, Vol. 7* (pp. 607–630). Elsevier, North-Holland.

de Rooij, S., & Grünwald, P. D. (2011). Luckiness and regret in minimum description length inference. In D. M. Gabbay, P. S. Bandyopadhyay, M. R. Forster, P. Thagard, & J. Woods (Eds.), *Handbook of the philosophy of science, Vol. 7* (pp. 865–900). Elsevier, North-Holland.

Erven, T. v., Grünwald, P., & De Rooij, S. (2012). Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC–BIC dilemma. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *74*(3), 361–417.

Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*(2), 313–329.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, *41*, 155–160.

Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error.. *Monthly Notices of the Royal Astronomical Society*, *80*, 758–770.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical Or Physical Character*, *222*, 309–368.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*(5), 700–725.

Frieden, B. R. (2004). *Science from Fisher information: A unification*. Cambridge University Press.

Ghosal, S., Ghosh, J., & Ramamoorthi, R. (1997). Non-informative priors via sieves and packing numbers. In *Advances in statistical decision theory and applications* (pp. 119–132). Springer.

Girolami, M., & Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *73*(2), 123–214.

Golden, R. M. (1995). Making correct statistical inferences using the wrong probability model. *Journal of Mathematical Psychology*, *39*, 3–20.

Golden, R. M. (2000). Statistical tests for comparing possibly misspecified and nonnested models. *Journal of Mathematical Psychology*, *44*(1), 153–170.

Grazian, C., & Liseo, B. (2014). Approximate integrated likelihood via ABC methods, arXiv Preprint arXiv:1403.0387.

Grazian, C., & Robert, C. P. (2015). Jeffreys' priors for mixture estimation. In *Bayesian statistics from methods to models and applications* (pp. 37–48). Springer.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2017). Informed Bayesian *t*-Tests, arXiv Preprint arXiv:1704.02479.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., & Marsman, M. (2017). A tutorial on bridge sampling, arXiv Preprint arXiv:1703.05984.

Grünwald, P. D. (1998). The Minimum Description Length Principle and Reasoning under Uncertainty, Ph.D. thesis, ILLC and University of Amsterdam.

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

Grünwald, P. D., & Mehta, N. A. (2016). Fast rates with unbounded losses, arXiv Preprint arXiv:1605.00252.

Grünwald, P. D., Myung, I. J., & Pitt, M. A. (Eds.). (2005). *Advances in minimum description length: theory and Applications*. Cambridge, MA: MIT Press.

Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Zeitschrift FÜR Wahrscheinlichkeitstheorie Und Verwandte Gebiete*, *14*(4), 323–330.

Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* Vol. 1, (pp. 175–194).

Hald, A. (2008). *A history of parametric statistical inference from Bernoulli to Fisher, 1713-1935*. Springer Science & Business Media.

Heck, D. W., Moshagen, M., & Erdfelder, E. (2014). Model selection by minimum description length: lower-bound sample sizes for the fisher information approximation. *Journal of Mathematical Psychology*, *60*, 29–34.

Inagaki, N. (1970). On the limiting distribution of a sequence of estimators with uniformity property. *Annals of the Institute of Statistical Mathematics*, *22*(1), 1–13.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *186*(1007), 453–461.

Jeffreys, H. (1961). *Theory of probability*. (third ed). Oxford, UK: Oxford University Press.

Kass, R. E. (1989). The geometry of asymptotic inference. *Statistical Science*, *4*(3), 188–234.

Klaassen, C. A., & Lenstra, A. J. (2003). Vanishing Fisher information. *Acta Applicandae Mathematicae*, *78*(1), 193–200.

Klauer, K. C., & Kellen, D. (2011). The flexibility of models of recognition memory: an analysis by the minimum-description length principle. *Journal of Mathematical Psychology*, *55*(6), 430–450.

Kleijn, B. J. K., & Zhao, Y. (2013). Criteria for posterior consistency, arXiv Preprint arXiv:1308.1263.

Korostelev, A. P., & Korosteleva, O. (2011). *Mathematical statistics: asymptotic minimax theory, Vol. 119*. American Mathematical Society.

Kotz, S., Kozubowski, T. J., & Podgorski, K. (2001). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. New York: Springer.

LeCam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, *41*(3), 802–828.

LeCam, L. (1990). Maximum likelihood: an introduction. *International Statistical Review/Revue Internationale De Statistique*, *58*(2), 153–171.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Lehmann, E. L. (2011). *Fisher, Neyman, and the creation of classical statistics*. Springer Science & Business Media.

Li, Y., & Clyde, M. A. (2015). Mixtures of g-priors in generalized linear models, arXiv Preprint arXiv:1503.06913.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, *103*(481).

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2017). Replication Bayes factors. *Manuscript in Preparation*.

Ly, A., Marsman, M., & Wagenmakers, E.-J. (in press) Analytic posteriors for Pearson's correlation coefficient. Statistica Neerlandica.

Ly, A., Raj, A., Marsman, M., Etz, A., & Wagenmakers, E.-J. (2017). Bayesian reanalyses from summary statistics and the strength of statistical evidence. *Manuscript Submitted for Publication*.

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016a). An evaluation of alternative methods for testing hypotheses, from the perspective of harold Jeffreys. *Journal of Mathematical Psychology*, *72*, 43–55.

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016b). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32.

Marsman, M., Ly, A., & Wagenmakers, E.-J. (2016). Four requirements for an acceptable research program. *Basic and Applied Social Psychology*, *38*(6), 308–312.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90–100.

Myung, I. J., Balasubramanian, V., & Pitt, M. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, *97*(21), 11170–11175.

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

Myung, J. I., & Navarro, D. J. (2005). Information matrix. *Encyclopedia of Statistics in Behavioral Science*.

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

Myung, J., & Pitt, M. A. (in press). Model comparison in psychology. In *Wixted, J. and Wagenmakers, E.-J., editors, The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience (Fourth Edition)*, volume 5: Methodology. John Wiley & Sons, New York, NY.

Pitt, M., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition.. *Psychological Review*, *109*(3), 472–491.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological Methodology* (pp. 111–196). Cambridge: Blackwells.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ray, K., & Schmidt-Hieber, J. (2016). Minimax theory for a class of nonlinear statistical inverse problems. *Inverse Problems*, *32*(6), 065003.

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

Rivoirard, V., & Rousseau, J. (2012). Bernstein–von Mises theorem for linear functionals of the density. *The Annals of Statistics*, *40*(3), 1489–1523.

Robert, C. P. (2016). The expected demise of the Bayes factor. *Journal of Mathematical Psychology*, *72*, 33–37.

Robert, C. P., Chopin, N., & Rousseau, J. (2009). Harold Jeffreys's Theory of Probability revisited. *Statistical Science*, 141–172.

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

Rudin, W. (1991). Functional analysis. *International series in pure and applied mathematics*. (second ed). (p. xviii+424). New York: McGraw-Hill, Inc..

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181.

Stigler, S. (1973). Studies in the history of probability and statistics. XXXII Laplace, Fisher, and the discovery of the concept of sufficiency. *Biometrika*, *60*(3), 439–445.

Stigler, S. (1986). *The history of statistics: the measurement of uncertainty before 1900*. Belknap Press.

van der Pas, S., & Grünwald, P. D. (2014). Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in single parameter model selection, arXiv Preprint arXiv:1408.5724.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

Wagenmakers, E.-J., Grünwald, P. D., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, *50*, 149–166.

Wagenmakers, E.-J., & Waldorp, L. (2006). Model selection: Theoretical Developments and Applications [Special Issue]. *Journal of Mathematical Psychology*, *50*(2).

Wald, A. (1949). Statistical decision functions. *The Annals of Mathematical Statistics*, 165–205.

Waldorp, L. (2009). Robust and unbiased variance of GLM coefficients for misspecified autocorrelation and hemodynamic response models in fMRI. *International Journal of Biomedical Imaging*, *2009*, 723912.

Waldorp, L., Christoffels, I., & van de Ven, V. (2011). Effective connectivity of fMRI data using ancestral graph theory: Dealing with missing regions. *NeuroImage*, *54*(4), 2695–2705.

Waldorp, L., Huizenga, H., & Grasman, R. (2005). The Wald test and Cramér–Rao bound for misspecified models in electromagnetic source analysis. *IEEE Transactions on Signal Processing*, *53*(9), 3427–3435.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25.

Wrinch, D., & Jeffreys, H. (1919). On some aspects of the theory of probability. *Philosophical Magazine*, *38*, 715–731.

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*, 369–390.

Wrinch, D., & Jeffreys, H. (1923). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *45*, 368–375.

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

Yang, G. L., & Le Cam, L. (2000). *Asymptotics in statistics: Some basic concepts*. Berlin: Springer-Verlag.