International Journal of Mathematics for Industry
Vol. 11, No. 1 (2020) 1930001 (29 pages)

DOI: 10.1142/S2661335219300018

World Scientific
www.worldscientific.com

Minimum description length revisited

Peter Grünwald[*,†,¶] and Teemu Roos[‡,§]
[*]*National Research Institute for Mathematics and Computer Science in the Netherlands (CWI)*
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
[†]*Leiden University, P.O. Box 9500*
2300 RA Leiden, The Netherlands
[‡]*Department of Computer Science, University of Helsinki*
P.O. Box 68, FI-00014, Helsinki, Finland
[§]*Helsinki Insitute of Information Technology (HIIT)*
P.O. Box 68, FI-00014, Helsinki, Finland
[¶]*pdg@cwi.nl*

Received 20 August 2019
Accepted 16 November 2019
Published 12 March 2020

This is an up-to-date introduction to and overview of the Minimum Description Length (MDL) Principle, a theory of inductive inference that can be applied to general problems in statistics, machine learning and pattern recognition. While MDL was originally based on data compression ideas, this introduction can be read without any knowledge thereof. It takes into account all major developments since 2007, the last time an extensive overview was written. These include new methods for model selection and averaging and hypothesis testing, as well as the first completely general definition of *MDL estimators*. Incorporating these developments, MDL can be seen as a powerful extension of both penalized likelihood and Bayesian approaches, in which penalization functions and prior distributions are replaced by more general luckiness functions, average-case methodology is replaced by a more robust worst-case approach, and in which methods classically viewed as highly distinct, such as AIC versus BIC and cross-validation versus Bayes can, to a large extent, be viewed from a unified perspective.

Keywords: MDL Principle; model selection; penalized estimation; universal prediction.

[¶]Corresponding author.

1. Introduction

The *Minimum Description Length* (*MDL*) *Principle*[1–4] is a theory of inductive inference that can be applied to general problems in statistics, machine learning and pattern recognition. Broadly speaking, it states that the best explanation for a given set of data is provided by the shortest description of that data. In 2007, one of us published the book *The Minimum Description Length Principle* (Ref. 4 referred to as G07 from now on), giving a detailed account of most works in the MDL area that had been done until then. During the last 10 years, several new practical MDL methods have been designed, and there have been exciting theoretical developments as well. It therefore seemed time to present an up-to-date combined introduction and review.

Why read this overview?

While the MDL idea has been shown to be very powerful in theory, and there have been a fair number of successful practical implementations, massive deployment has been hindered by two issues: first, in order to apply MDL, one needs to have basic knowledge of both statistics and information theory. To remedy this situation, here we present, for the first time, the MDL Principle *without resorting to information theory*: all the material can be understood without any knowledge of data compression, which should make it a much easier read for statisticians and machine learning researchers novel to MDL. A second issue is that many classical MDL procedures are either computationally highly intensive (for example, MDL variable selection as in Example 4 below) and hence less suited for our big data age, or they seem to require somewhat arbitrary restrictions of parameter spaces (e.g., NML with $v \equiv 1$ as in Sec. 2). Yet, over the last 10 years, there have been exciting developments — some of them very recent — which mostly resolve these issues. Incorporating these developments, MDL can be seen as a powerful extension of both penalized likelihood and Bayesian approaches, in which penalization functions and prior distributions are replaced by more general luckiness functions, average-case methodology is replaced by a more robust worst-case approach, and in which methods classically viewed as highly distinct, such as AIC versus BIC and cross-validation versus Bayes can, to some extent, be viewed from a unified perspective; as such, this paper should also be of interest to researchers working on the foundations of statistics and machine learning.

History of the field, recent advances and overview of this paper

MDL was introduced in 1978 by Jorma Rissanen in his paper *Modeling by the Shortest Data Description*. The paper coined the term *MDL* and introduced and analyzed the two-part code for parametric models. The two-part code is the simplest instance of a *universal code* or, equivalently, *universal probability distribution*, the cornerstone concept of MDL theory. MDL theory was greatly extended in the 1980s, when Rissanen published a sequence of ground-breaking papers at a remarkable pace, several of which introduced new types of universal distributions. It came to full blossom in the 1990s, with further major contributions from, primarily, Jorma Rissanen, Andrew Barron and Bin Yu, culminating in their overview paper[3] and the collection[5] with additional chapters by other essential contributors such as Kenji Yamanishi. The book G07 provides a more exhaustive treatment of this early work, including discussion of important precursors/alternatives to MDL such as MML,[6] "ideal", Kolmogorov complexity-based MDL[7] and Solomonoff's theory of induction.[8] Universal distributions are still central to MDL. We introduce them in a concise yet self-contained way, including substantial underlying motivation, in Sec. 2, incorporating the extensions to and new insights into these basic building blocks that have been gathered over the last 10 years. These include more general formulations of arguably the most fundamental universal code, the *Normalized Maximum Likelihood* (*NML*) *Distribution*, including faster ways to calculate it as well. We devote a separate section to new universal codes, with quite pleasant properties for practical use, most notably the *switch distribution* (Sec. 3.1), which can be used for model selection combining almost the best of AIC and BIC; and the Reverse Information Projection (*RIPr*)-universal code (Sec. 3.3) specially geared to hypothesis testing with composite null hypotheses, leading to several advantages over classical Neyman–Pearson tests. In Sec. 4 we review recent developments on fast calculation of NML-type distributions for model selection for *graphical models* (Bayesian networks and the like), leading to methods which appear to be more robust in practice than the standard Bayesian ones.

Recent extensions of MDL theory and practical implementations to latent variable and irregular models are treated in Sec. 5. Then, in Sec. 6 we review developments relating to consistency and convergence properties of MDL methods. First, while originally MDL estimation was formulated solely in terms of discretized estimators (reflecting the fact that coding always requires discretization), it has gradually become clear that a much larger class of estimators (including maximum likelihood for "simple" models, and, in some circumstances, the Lasso — see Example 4) can be viewed from an MDL perspective, and this becomes clearest if one investigates asymptotic convergence theorems relating to MDL. Second, it was found that MDL (and Bayes), without modification, can behave suboptimally under misspecification, i.e., when all models under consideration are wrong, but some are useful — see Sec. 6.3. Third, very recently, it was shown how some of the surprising phenomena underlying the *deep learning* revolution in machine learning can be explained from an MDL-related perspective; we briefly review these developments in Sec. 6.4. Finally, we note that G07 presented many explicit open problems, most of which have been resolved — we mention throughout the text whenever a new development solved an old open problem, deferring some of the most technical issues to the Appendix.

Notational preliminaries

We shall mainly be concerned with *statistical models* (families of probability distributions) of the form $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$ parametrized by some set Θ which is usually but not always a subset of Euclidean space; and *families of models* $\{\mathcal{M}_\gamma : \gamma \in \Gamma\}$, where each $\mathcal{M}_\gamma = \{p_\theta : \theta \in \Theta_\gamma\}$ is a statistical model, used to model the data $z^n := (z_1, \ldots, z_n)$ with each $z_i \in \mathcal{Z}$, for some outcome space $\mathcal{Z}$. Each p_θ represents a probability density function (pdf) or probability mass function, defined on sequences of arbitrary length. With slight abuse of notation we also denote the corresponding probability distribution by p_θ (rather than the more common P_θ). In the simple case that the data are i.i.d. according to each p_θ under consideration, we have $p_\theta(z^n) = \prod_{i=1}^n p_\theta(z_i)$.

We denote the maximum likelihood (ML) estimator given the model $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$ by $\hat{\theta}_{\mathrm{ML}}$, whenever it exists and is unique; the ML estimator relative to model $\mathcal{M}_\gamma$ is denoted by $\hat{\theta}_{\mathrm{ML}|\gamma}$. We shall, purely for simplicity, generally assume its existence and uniqueness, although nearly all results can be generalized to the case where it does not. We use $\breve{\theta}$ to denote more general estimators, and $\hat{\theta}_v$ to denote what we call the *MDL estimator with luckiness function* v, see (5).

2. The Fundamental Concept: Universal Modeling

MDL is best explained by starting with one of its prime applications, model comparison — we will generalize to prediction and estimation later, in Secs. 2.3 and 2.4. Assume then that we are given a finite or countably infinite collection of statistical models $\mathcal{M}_1, \mathcal{M}_2, \ldots$, each consisting of a set of probability distributions. The fundamental idea of MDL is to associate each $\mathcal{M}_\gamma$ with a *single* distribution $\bar{p}_\gamma$, often called a *universal distribution* relative to $\mathcal{M}_\gamma$. We call the minus-log-likelihood $-\log \bar{p}_\gamma(Z^n)$ the *code length of data* Z^n *under the universal code* $\bar{p}_\gamma$. This terminology, and how MDL is related to coding (lossless compression of data), is briefly reviewed in Secs. 2.3 and 2.4; but a crucial observation at this point is that the main MDL ideas can be understood abstractly, without resorting to the code length interpretation. We also equip the model indices $\Gamma := \{1, 2, \ldots, \gamma_{\max}\}$ (where we allow $|\Gamma| = \gamma_{\max} = \infty$) with a distribution, say π; if the number of models to be compared is small (e.g., bounded independently of n or at most a small polynomial in n), we can take π to be uniform distribution — for large (exponential in n) and infinite Γ, see Sec. 2.3 and Example 4. We then take, as our best explanation of the given data z^n, the model $\mathcal{M}_\gamma$ minimizing

$$-\log \pi(\gamma) - \log \bar{p}_\gamma(z^n), \tag{1}$$

or, equivalently, we maximize $\pi(\gamma)\bar{p}_\gamma(z^n)$; when π is uniform this simply amounts to picking the γ maximizing $\bar{p}_\gamma(z^n)$. (1) will later be generalized to π that are not distributions but rather more general "luckiness functions" — see Sec. 2.3.

(1) The Bayesian universal distribution

The reader may recognize this as being formally equivalent to the standard Bayesian way of model selection, the *Bayes factor method*[9] as long as the γ

are defined as Bayesian marginal distributions, i.e., for each γ, we set $\bar{p}_\gamma = p^{\text{BAYES}}_{w_\gamma}$, where

$$p^{\text{BAYES}}_{w_\gamma}(z^n) := \int p_\theta(z^n) w_\gamma(\theta) d\theta, \tag{2}$$

for some prior probability density w_γ on the parameters in Θ_γ, which has to be supplied by the user. When w_γ is clear from the context, we shall write $\bar{p}^{\text{BAYES}}_\gamma$ rather than $p^{\text{BAYES}}_{w_\gamma}$. Using Bayesian marginal distributions $\bar{p}^{\text{BAYES}}$ is indeed one possible way to instantiate MDL model selection, but it is not the only way: MDL can also be based on other distributions such as $\bar{p}^{\text{NML}} = p^{\text{NML}}_v$ (depending on a function v), $\bar{p}^{\text{PREQ}} = p^{\text{PREQ}}_{\breve{\theta}}$ (depending on an estimator $\breve{\theta}$) and others; in general we add a bar to such distributions if the "parameter" w, v or $\breve{\theta}$ is clear from the context. Before we continue with these other instantiations of $\bar{p}_\gamma$ we proceed with an example.

Example 1 (Bernoulli). Let $\mathcal{M} = \{p_\theta : \theta \in [0,1]\}$ represent the Bernoulli model, extended to n outcomes by independence. We then have for each $z^n \in \{0,1\}^n$ that $p_\theta(z^n) = \theta^{n_1}(1-\theta)^{n_0}$, where $n_1 = \sum_{i=1}^n z_i$ and $n_0 = n - n_1$. Most standard prior distributions one encounters in the literature are beta priors, for which $w(\theta) \propto \theta^\alpha (1-\theta)^\beta$, so that $p^{\text{BAYES}}_w(z^n) \propto \int \theta^{n_1+\alpha}(1-\theta)^{n_0+\beta} d\theta$. Note that p^{BAYES}_w is not itself an element of the Bernoulli model. One could use p^{BAYES}_w to compare the Bernoulli model, via (1), to, for example, a first-order Markov model, with Bayesian marginal likelihoods defined analogously. We shall say a lot more about the choice of prior below.

Example 2 (Gauss and general improper priors). A second example is the Gaussian location family $\mathcal{M}_{\text{GAUSS}}$ with fixed variance (say 1), in which $\mathcal{Z} = \mathbb{R}$, and $p_\theta(z^n) \propto \exp(\sum_{i=1}^n (z_i - \theta)^2/2)$. A standard prior for such a model is the uniform prior, $w(\theta) = 1$, which is *improper* (it does not integrate, hence does not define a probability distribution). Improper priors cannot be directly used in (2), and hence they cannot be directly used for model comparison as in (1) either. Still, we can use them in an indirect manner, as long as we are guaranteed that, for all $\mathcal{M}_\gamma$ under consideration, after some initial number of m observations, the Bayesian posterior $w_\gamma(\theta|z^m)$ is proper. We can then replace $p^{\text{BAYES}}_{w_\gamma}(z^n)$ in (2) by $p^{\text{BAYES}}_{w_\gamma}(z_{m+1}, \ldots, z_n|z^m) := \int p_\theta(z_{m+1}, \ldots, z_n) w_\gamma(\theta|z^m) d\theta$. We extend all these *conditional universal distributions* to distributions on $\mathcal{Z}^n$ by defining $p^{\text{BAYES}}_{w_\gamma}(z_1, \ldots, z_n) := p^{\text{BAYES}}_{w_\gamma}(z_{m+1}, \ldots, z_n|z^m) p_0(z^m)$ for some distribution p_0 on $\mathcal{Z}^m$ that is taken to be the same for all models $\mathcal{M}_\gamma$ under consideration. We can now use (1) again for model selection based on $p^{\text{BAYES}}_{w_\gamma}(z_1, \ldots, z_n)$, where we note that the choice of p_0 plays no role in the minimization, which is equivalent to minimizing $-\log \pi(\gamma) - \log p^{\text{BAYES}}_{w_\gamma}(z_{m+1}, \ldots, z_n|z^m)$.

Now comes the crux of the story, which makes MDL, in the end, quite different from Bayes: defining the $\bar{p}_\gamma$ as in (2) is just *one particular way* to define an MDL universal distribution — but it is by no means the only one. There are several other ways, and some of them are sometimes preferable to the Bayesian choice. Here we list the most important ones.

(2) NML or *Shtarkov*[10] distribution, and MDL estimators

This is perhaps the most fundamental universal distribution, leading also to the definition of an *MDL estimator*. In its general form, the NML distribution and "MDL estimators" depend on a function $v : \Theta \to \mathbb{R}_0^+$. The definition is then given by

$$p^{\text{NML}}_v(z^n) := \frac{\max_{\theta\in\Theta} p_\theta(z^n) v(\theta)}{\int \max_{\theta\in\Theta} p_\theta(z^n) v(\theta) dz^n} \overset{\text{(if } v \text{ constant)}}{=} \frac{p_{\hat{\theta}_{\text{ML}}(z^n)}(z^n)}{\int p_{\hat{\theta}_{\text{ML}}(z^n)}(z^n) dz^n}, \tag{3}$$

which is defined whenever the normalizing integral is finite. The logarithm of this integral is called the *model complexity* and is thus given by

$$\text{COMP}(\mathcal{M}; v) := \log \int \max_{\theta\in\Theta} (p_\theta(z^n) v(\theta)) dz^n \overset{\text{(if } v \text{ constant)}}{=} \log \int p_{\hat{\theta}_{\text{ML}}(z^n)}(z^n) dz^n. \tag{4}$$

Here the integral is replaced by a sum for discrete data, and max is replaced by sup if necessary. This means that any function $v : \Theta \to \mathbb{R}_0^+$ such that (4) is finite is allowed; we call any such v a *luckiness function*, a terminology we explain later. Note that v is not necessarily a probability density — it does not have to be integrable. For any luckiness function v, we define the *MDL estimator based on* v as

$$\hat{\theta}_v := \arg\max_{\theta\in\Theta} p_\theta(z^n) v(\theta) = \arg\min_{\theta\in\Theta} \{-\log p_\theta - [-\log v(\theta)]\}. \tag{5}$$

The v-MDL estimator is a penalized ML estimator, which coincides with the Bayes MAP estimator based on prior v whenever v is a probability density. Although this has only become clear gradually over the last 10 years, estimators of form (5) are the prime way of using MDL for estimation; there is, however, a second, "improper" way for estimating distributions within MDL though, see Sec.2.4. In practice, we will choose v that are sufficiently smooth so that, if the number of parameters is small relative to n, $\hat{\theta}_v$ will usually be almost indistinguishable from the ML estimator $\hat{\theta}_{\mathrm{ML}}$. COMP indeed measures something one could call a "complexity" — this is easiest to see if $v = 1$, for then, if $\mathcal{M}$ contains just a single distribution, we must have $\mathrm{COMP}(\mathcal{M}, v) = 0$, and the more distributions we add to $\mathcal{M}$, the larger $\mathrm{COMP}(\mathcal{M}, v)$ gets — this is explored further in Sec. 2.2.

Now suppose we have a collection of models $\mathcal{M}_\gamma$ indexed by finite Γ and we have specified luckiness functions v_γ on Θ_γ for each $\gamma \in \Gamma$, and we pick a uniform distribution π on Γ. As can be seen from the above, if we base our model choice on NML, we pick the model minimizing

$$-\log p_{\hat{\theta}_{v_\gamma}(z^n)}(z^n) - \log v_\gamma(\hat{\theta}_{v_\gamma}(z^n)) + \mathrm{COMP}(\mathcal{M}_\gamma; v_\gamma), \tag{6}$$

over γ, where $\mathrm{COMP}(\mathcal{M}_\gamma; v_\gamma)$ is given by

$$\begin{aligned}\mathrm{COMP}(\mathcal{M}_\gamma; v_\gamma) &= \log \int \max_{\theta \in \Theta_\gamma} (p_\theta(z^n) v_\gamma(\theta)) dz^n \\ &= \log \int p_{\hat{\theta}_{v_\gamma}(z^n)}(z^n) v_\gamma(\hat{\theta}_{v_\gamma}(z^n)) dz^n.\end{aligned} \tag{7}$$

Thus, by (6), MDL incorporates a trade-off between goodness of fit and model complexity as measured by COMP. Although the n-fold integral inside COMP looks daunting Suzuki and Yamanishi[11] show that in many cases (e.g., normal, Weibull–Laplace models) it can be evaluated explicitly with appropriate choice of v.

Originally, the NML distribution was defined by Shtarkov[10] for the special case with $v \equiv 1$, leading to the rightmost definition in (3), and hence the term NML (in the modern version, perhaps "normalized *penalized* ML" would be more apt). This is also the version that Rissanen[12] advocated as embodying the purest form of the MDL Principle. However, the integral in (3) is ill-defined for just about every parametric model defined on unbounded outcome spaces (such as $\mathbb{N}, \mathbb{R}$ or $\mathbb{R}^+$), including the simple normal location family. Using nonuniform v allows one to deal with such cases in a principled manner after all, see Sec. 2.5. For finite outcome spaces though, $v \equiv 1$ usually "works", and (3) is well defined, as we illustrate for the Bernoulli model (see Sec. 4 for more examples).

Example 3 (Continuation of Example 1). For the Bernoulli model, $\hat{\theta}_{\mathrm{ml}}(z^n) = n_1/n$ and $\mathrm{COMP}(\mathcal{M}, v)$ as in (7) with $v \equiv 1$ can be rewritten as $\log \sum_{n_1=0}^{n} \binom{n}{n_1} (n_1/n)^{n_1} (n_0/n)^{n_0}$, which, as we shall see in Sec. 2.2, is within a constant of $(1/2) \log n$. As reviewed in that sub-section, the resulting p_v^{NML} is asymptotically (essentially) indistinguishable from $p_{w_{\mathrm{J}}}^{\mathrm{BAYES}}$ where the latter is equipped with *Jeffreys' prior*, defined as $w_{\mathrm{J}}(\theta) \propto \sqrt{|I(\theta)|} = \theta^{-1/2}(1-\theta)^{-1/2}$, with $I(\theta)$ being the Fisher information at θ.

(3) The two-part (sub-)distribution[1]
Here one first discretizes Θ to some countable subset $\ddot{\Theta}$ which one equips with a probability mass function w; in contrast to the v above, this function must sum to 1. One then considers

$$p_w^{\mathrm{NML}}(z^n) := \frac{\max_{\ddot{\theta} \in \ddot{\Theta}} p_{\ddot{\theta}}(z^n) w(\ddot{\theta})}{\int \max_{\ddot{\theta} \in \ddot{\Theta}} p_{\ddot{\theta}}(z^n) w(\ddot{\theta}) dz^n}, \tag{8}$$

which is just a special case of (3). But since

$$\begin{aligned}\int \max_{\ddot{\theta} \in \ddot{\Theta}} p_{\ddot{\theta}}(z^n) w(\ddot{\theta}) dz^n &\le \int \sum_{\ddot{\theta} \in \ddot{\Theta}} p_{\ddot{\theta}}(z^n) w(\ddot{\theta}) dz^n \\ &= \sum_{\theta \in \ddot{\Theta}} w(\theta) \left(\int p_\theta(z^n) dz^n \right) \\ &= 1,\end{aligned} \tag{9}$$

we can approximate p_w^{NML} by the *sub*-distribution $p_w^{2\text{-P}}(z^n) := \max_{\ddot{\theta} \in \ddot{\Theta}} p_{\ddot{\theta}}(z^n) w(\ddot{\theta})$. This "distribution" adds or integrates to something smaller than 1. This can be incorporated into the general story by imagining that $p_w^{2\text{-P}}$ puts its remaining mass on a special outcome, say "$\diamond$", which in reality will never occur (while sub-distributions are thus "allowed", measures that add up to something *larger* than 1 have no place in MDL). The two-part distribution $p_w^{2\text{-P}}$ is historically the oldest universal distribution. The fact that it can be considered a special case of NML has only become fully clear very recently[13]; in

that same paper, an even more general formulation of (3) is given that has all Bayesian, two-part and NML distributions as special cases. Despite its age, the two-part code is still important in practice, as we explain in Sec. 2.3.

(4) The prequential plug-in distribution[14,15]
Here, one first takes any reasonable estimator $\breve{\theta}$ for the given model $\mathcal{M}$. One then defines

$$p_{\breve{\theta}}^{\text{PREQ}}(z^n) := \prod_{i=1}^{n} p_{\breve{\theta}(z^{i-1})}(z_i|z^{i-1}), \tag{10}$$

where for i.i.d. models, the probability inside the product simplifies to $p_{\breve{\theta}(z^{i-1})}(z_i)$. For the normal location family, one could simply use the ML estimator: $\breve{\theta}(z^m) := \hat{\theta}_{\text{ML}}(z^m) = \sum_{j=1}^{m} z_j/m$. With discrete data though, the ML estimator should be avoided, since then one of the factors in (10) could easily become 0, making the product 0, so that the model for which $p_{\breve{\theta}}^{\text{PREQ}}$ is defined can never "win" the model selection contest even if most other factors in the product (10) are close to 1. Instead, one can use a slightly "smoothed" ML estimate (a natural choice for $\breve{\theta}$ is to take an MDL estimator for some v as in (5), but this is not required). For example, in the Bernoulli model, one might take $\breve{\theta}(z^m) = (m_1 + (1/2))/(m+1)$, where $m_1 = \sum_{i=1}^{m} z_i$. With this particular choice, $p_{\breve{\theta}}^{\text{PREQ}}$ turns out to coincide *exactly* with $p_{w_{\text{J}}}^{\text{BAYES}}$ with Jeffreys' prior w_{J}. Such a precise correspondence between $\bar{p}^{\text{PREQ}}$ and $\bar{p}^{\text{BAYES}}$ is a special property of the Bernoulli and multinomial models though; with other models, the two distributions can usually be made to behave similarly, but not identically. The rationale for using $\bar{p}^{\text{PREQ}}$ is described in Sec. 2.4. In Sec. 3.2.1 we will say a bit more about hybrids between prequential plug-in and Bayes (the *flattened leader distribution*) and between prequential and NML (*sequential NML*).

Except for the just mentioned "hybrids", these first four universal distributions were all brought into MDL theory by Rissanen; they are extensively treated by G07, in which one chapter is devoted to each, and to which we refer for details. The following two are much more recent.

(5) The switch distribution $\bar{p}^{\text{SWITCH}}$ (Ref. 16)
In a particular type of nested model selection, this universal distribution behaves arguably better than the other ones. It will be treated in detail in Sec. 3.1.

(6) Universal distributions $\bar{p}^{\text{RIPR}}$ based on the Reverse Information Projection
These universal distributions[17] lead to improved error bounds and optional stopping behavior in hypothesis testing and allow one to forge a connection with group-invariant Bayes factor methods; see Sec. 3.3.

2.1. *Motivation*

We first give a very high-level motivation that avoids direct use of data compression arguments. For readers interested in data compression, Sec. 2.3 does make a high-level connection, but for more extensive material we refer to G07. We do, in Sec. 2.4, give a more detailed motivation in predictive terms, and, in Sec. 6, we shall review mathematical results indicating that MDL methods are typically consistent and enjoy fast rates of convergence, providing an additional motivation in itself.

Consider then models $\mathcal{M}_\gamma$, where for simplicity we assume discrete data, and let $\hat{\theta}_{\text{ML}|\gamma}$ be the maximum likelihood estimator within $\mathcal{M}_\gamma$. Define "the fit of the model to the data" in the standard way, as $F_\gamma(z^n) := p_{\hat{\theta}_{\text{ML}|\gamma}(z^n)}(z^n)$, the likelihood assigned to the data by the best-fitting distribution within the model. Now if we enlarge the model $\mathcal{M}_\gamma$, i.e., by adding several distributions to it, $F_\gamma(z^n)$ can only increase; and if we make $\mathcal{M}_\gamma$ big enough such that for each z^n, it contains a distribution p with $p(z^n) = 1$, we can even have $F_\gamma(z^n) = 1$ on all data. If we simply picked the γ maximizing $F_\gamma(z^n)$, we would be prone to severe overfitting. For example, if models are nested, then, except for very special data, we would automatically pick the largest one.

As we have seen, a central MDL idea is to instead associate each model $\mathcal{M}_\gamma$ with a *single* corresponding distribution $\bar{p}_\gamma$, i.e., we set $F_\gamma(z^n) := \bar{p}_\gamma(z^n)$. Then the total probability mass on all potential outcomes z^n cannot be larger than 1, which makes it impossible to assign overly high fit $F_\gamma(z^n)$ to overly many data sequences: no matter what distribution $\bar{p}_\gamma$ we chose, we must now have $\sum_{z^n} F_\gamma(z^n) = 1$, so a good fit on some z^n necessarily implies a worse fit on others, and we will not select a model simply because it accidentally contained *some* distribution that fitted our data very well — thus, measuring fit by a distribution $\bar{p}_\gamma$ instead of F_γ inherently prevents overfitting. This argument to measure fit relative to a model with a single $\bar{p}$ is similar to

Bayesian Occam's Razor arguments[18] used to motivate the Bayes factor; the crucial difference is that we do not restrict ourselves to $\bar{p}_\gamma$ of the form (2); inspecting the "Bayesian" Occam argument, there is, indeed, nothing in there which forces us to use distributions of Bayesian form.

The next step is thus to decide *which* $\bar{p}$ are best associated with a given $\mathcal{M}$. To this end, we define the *fitness ratio* for data z^n as

$$\mathrm{FR}(\bar{p}, z^n) := \frac{\bar{p}(z^n)}{\max_{\theta\in\Theta} p_\theta(z^n) v(\theta)}, \tag{11}$$

where $v\colon \Theta \to \mathbb{R}_0^+$ is a nonnegative function. To get a feeling for (11), it is best to first focus on the case with $v(\theta) \equiv 1$; it then reduces to

$$\mathrm{FR}(\bar{p}, z^n) = \frac{\bar{p}(z^n)}{p_{\hat{\theta}_{\mathrm{ML}}(z^n)}(z^n)}. \tag{12}$$

We next postulate that a good choice for $\bar{p}$ relative to the given model is one in which $\mathrm{FR}(\bar{p}, n)$ *tends to be as large as possible*. The rationale is that, overfitting having already been taken care of by picking some $\bar{p}$ that is a probability measure (integrates to 1), it makes sense to take a $\bar{p}$ whose fit to data (as measured in terms of likelihood) is proportional to the fit to data of the best-fitting distribution in $\mathcal{M}$: whenever *some* distribution in the model $\mathcal{M}$ fits the data z^n well, the likelihood $\bar{p}(z^n)$ should be high as well. One way to make "FR tends to be large" precise is by requiring it to be as large as possible in the worst-case, i.e., we want to pick the $\bar{p}$ achieving

$$\max_{\bar{p}} \min_{z^n \in \mathcal{Z}^n} \mathrm{FR}(\bar{p}, z^n), \tag{13}$$

where the maximum is over all probability distributions over samples of length n. It turns out that this maximin problem has a solution if and only if the complexity (4) is finite; and if it is fine, the unique solution is given by setting $\bar{p} = \bar{p}^{\mathrm{NML}}$, with $\bar{p}^{\mathrm{NML}}$ given by (3). The NML distribution thus has a special status as the most robust choice of universal $\bar{p}$ — even though $\bar{p}$ is itself a probability distribution, it meaningfully assesses fit in the worst-case over all possible distributions, and its interpretation does not require one to assume that the model $\mathcal{M}$ is "true" in any sense. The nicest sub-case is the one with $v(\theta) \equiv 1$, since then all distributions within the model $\mathcal{M}$ are treated on exactly the same footing; no data or distribution is intrinsically preferred over any other one. Unfortunately, for most popular models with infinite $\mathcal{Z}$, when taking $v(\theta) \equiv 1$, (13) usually has no solution since the integral $\int p_{\hat{\theta}_{\mathrm{ML}}(z^n)}(z^n) dz^n$ diverges for such models, making the complexity (4) infinite. For all sufficiently "regular" models (curved exponential families, see below), this problem can invariably be solved by restricting Θ to a bounded subset of its own interior — one can show that the complexity (4) is finite with $v \equiv 1$, and thus (13) has a solution given by (3) if $\hat{\theta}_{\mathrm{ml}}$ is restricted to a suitably bounded set. Yet, restricting Θ to a bounded subset of itself is not satisfactory, since it is unclear where exactly to put the boundaries. It is more natural to introduce a nonuniform v, which can invariably be chosen so that the complexity (4) is finite and thus (13) has a solution — more on choosing v at the end of Sec. 2.4.

A few remarks concerning this high-level motivation of MDL procedures are in order.

(1) It is clear that, by requiring FR to add to 1, we will be less prone to overfitting than by setting it simply to $p_{\hat{\theta}_{\mathrm{ML}}(z^n)}(z^n)$; whether the requirement to add (at most) to 1, making FR essentially a probability density function, is a *clever* way to avoid overfitting (leading to good results in practice) is not clear yet. For this, we need additional arguments, which we very briefly review. First, the sum-to-1 requirement is the only choice for which the procedure can be interpreted as selecting the model which minimizes code length of the data (the original interpretation of MDL); second, it is the only choice which has a predictive interpretation, which we review in Sec. 2.4 below; third, it is the only choice under which time-tested Bayesian methods fit into the picture; and fourth, with this choice we get desirable frequentist statistical properties such as consistency and convergence rates, see Sec. 6.

(2) The motivation above only applies to the NML universal distributions. How about the other five types? Originally, in the pure MDL approach mainly due to Rissanen, the NML was viewed as the optimal choice *per se*; other $\bar{p}$ should be used only for pragmatic reasons, such as them being easier to calculate. One would then design them so as to be as close as possible to the NML distributions in terms of the fitness ratio they achieve. In the following sub-section we show that all six of them satisfy the same

MDL/BIC asymptotics, meaning that their fitness ratio is never smaller than a constant factor of the NML one, either again in the worst-case over all z^n or in some weaker expectation sense. Thus, they are all "kind of ok" in a rather weak sense, and in practice one would simply revert to the one that is closest to NML and still usable in practice; with the Bayesian $\bar{p}^{\text{BAYES}}$, as we shall see, one can even get arbitrarily close to NML as n gets larger. This classical story notwithstanding, it has become more and more apparent that in practice one sometimes wants or needs properties of model selection methods that are not guaranteed by NML — such as near-optimal predictions of future data or strong frequentist Type-I error guarantees. This translates itself into universal codes $\bar{p}^{\text{SWITCH}}$ and $\bar{p}^{\text{RIPR}}$ that, for some special sequences, achieve much higher fitness ratio than $\bar{p}^{\text{NML}}$, while for all sequences having only very slightly smaller fitness ratio. This more recent and pragmatic way of MDL is briefly reviewed in Secs. 3.1 and 3.3. This raises the question how we should *define* a universal distribution: what choices for $\bar{p}_\gamma$ are still "universal" (and define an MDL method) and what choices are not? Informally, every distribution $\bar{p}_\gamma$ that for no $z^n \in \mathcal{Z}^n$ has $\bar{p}_\gamma(z^n) \ll \bar{p}_\gamma^{\text{NML}}(z^n)$ is "universal" relative to $\mathcal{M}_\gamma$. For parametric models such as exponential families, the "$\ll$" is partially formalized by requiring that at the very least, they should satisfy (14) below (G07 is much more precise on this).

(3) Third, we have not yet said how one should choose the "luckiness function" v — and one needs to make a choice to apply MDL in practice. The interpretation of v is closely tied to the predictive interpretation of MDL, and hence we postpone this issue to the end of Sec. 2.4.

(4) Fourth, the motivation so far is incomplete — we still need to explain why and how to incorporate the distribution π on model index γ. This is done in Sec. 2.3 below.

2.2. *Asymptotic expansions*

Now let $\bar{p}$ be defined relative to a single parametric model $\mathcal{M}$. It turns out that all universal codes we mentioned have in common that, for "sufficiently regular" k-dimensional parametric models, the log-likelihood for given data z^n satisfies the following celebrated asymptotics, often called the *MDL* or *BIC* expansion: for all "sufficiently regular" data sequences $z_1, z_2, \ldots$, there exists a constant $C \in \mathbb{R}$ independent of n such that for all n,

$$-\log \bar{p}(z^n) \leq -\log p_{\hat{\theta}_{\text{ML}}(z^n)}(z^n) + \frac{k}{2}\log n + C. \tag{14}$$

For $\bar{p}^{\text{NML}}$ and $\bar{p}^{\text{BAYES}}$, this holds for any choice of luckiness function v and prior w that is continuous and strictly positive on the parameter space Θ. For $\bar{p}_w^{2-\text{P}}$, this holds for clever choices of the discretization $\ddot{\Theta}$ and the probability mass function w; for $p_{\breve{\theta}}^{\text{PREQ}}$, this holds in a weaker expectation sense (see Sec. 3.2.1), as long as $\breve{\theta}$ is a suitably smoothed version of the ML estimator. Essentially, "sufficiently regular" parametric models are all exponential families (such as Bernoulli, multinomial, normal, gamma, beta, Poisson, etc.) and curved exponential families; corresponding results also hold for regression with (generalized) linear models. "Sufficiently regular data" are all sequences for which there is an INECCSI subset Θ_0 of the parameter space Θ such that, for all large n, the ML estimator of the sequence lies within Θ_0. Here INECCSI stands for a set whose *I*nterior is *N*on-*E*mpty and whose *C*losure is a *C*ompact *S*ubset of the *I*nterior of Θ. Essentially, this is any bounded subset of the same dimensionality as Θ that does not touch the boundaries of Θ itself; in the Bernoulli example, it would be any set of the form $[\epsilon, 1-\epsilon]$ for $\epsilon > 0$. For all universal distributions considered except $\bar{p}^{\text{PREQ}}$, as long as appropriate priors/estimators/luckiness functions are used, (14) will hold uniformly for all sequences in any INECCSI subset Θ_0, but the constant C may grow larger if we replace Θ_0 by a strictly larger INECCSI subset Θ_0' with $\Theta_0 \subsetneq \Theta_0' \subsetneq \Theta$. (for $\bar{p}^{\text{PREQ}}$ see Sec. 3.2.1). For the first four universal distributions, the inequality is actually equality up to a constant — (14) also holds with $\leq$ replaced by $\geq$, for a different constant. For the switch distribution $\bar{p}^{\text{SWITCH}}$, however, the left-hand side will be significantly smaller for a small but important subset of possible data sequences. Finally, since (14) thus also holds with $\bar{p} = \bar{p}^{\text{NML}}$ and $\leq$ replaced by $\geq$, exponentiating (14), we see that, if one restricts the minimum in (13) to all such "sufficiently regular" z^n, $\text{FR}(\bar{p}^{\mathbf{u}}(z^n))$ is guaranteed to be within a constant (independent of n) factor

of the optimal $\mathrm{FR}(\bar{p}^{\mathrm{NML}}, z^n)$, for $\mathbf{u} \in \{\text{BAYES}, 2-\text{P}, \text{PREQ}, \text{SWITCH}, \text{RIPR}\}$.

The NML/COMP expansion and the Jeffreys (Fisher information) integral

For the case that the model $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$ is a k-dimensional exponential family and $\bar{p}$ is the NML or Bayes distribution, we can be significantly more precise and evaluate the constant C in (14) up to $o(1)$: we get, under some weak additional regularity conditions on $\mathcal{M}$ and v,

$$\begin{aligned}\mathrm{COMP}(\mathcal{M}; v) \equiv & -\log p_v^{\mathrm{NML}}(z^n) - [-\log p_{\hat{\theta}_v(z^n)}(z^n) \\ & - \log v(\hat{\theta}_v(z^n))] = -\log p_v^{\mathrm{NML}}(z^n) \\ & + \log p_{\hat{\theta}_{\mathrm{ml}}(z^n)}(z^n) \cdot v(\hat{\theta}_{\mathrm{ML}}(z^n)) \\ = & \frac{k}{2}\log\frac{n}{2\pi} + \int_\Theta v(\theta)\cdot\sqrt{|I(\theta)|}d\theta + o(1),\end{aligned} \tag{15}$$

where k is the dimension of the model, $|I(\theta)|$ is the determinant of the $k \times k$ Fisher information matrix at parameter θ, the integral is over the parameter space Θ and the remainder term $o(1)$ vanishes as the sample size grows unbounded. This was first shown (essentially) by Rissanen,[12] for the case that Θ is restricted to an INECCSI subset of the full parameter space (so that $\bar{p}^{\mathrm{NML}}$ with $v \equiv 1$ is defined), and $v \equiv 1$. For this uniform v case, Myung *et al.*[19] gave a differential geometric interpretation of the Fisher information term, relating it to an intrinsic "volume" of the parameter space. The general result for nonuniform v, and without INECCSI restrictions, was very recently shown in a breakthrough paper by Suzuki and Yamanishi,[11] solving Open Problem 6 from G07.

Analogously to $\bar{p}^{\mathrm{NML}}$ (in fact much easier mathematically), we can expand $\bar{p}^{\mathrm{BAYES}}$ using a classical Laplace approximation; under the same conditions as before, with now the additional restriction that there exists an arbitrary INECCSI subset Θ_0 of Θ such that for all large n, the data have ML estimator within Θ_0, we find that

$$\begin{aligned}-\log p_w^{\mathrm{BAYES}}(z^n) = & -\log p_{\hat{\theta}_{\mathrm{ML}}(z^n)}(z^n) + \frac{k}{2}\log\frac{n}{2\pi} \\ & + \frac{1}{2}\log|I(\hat{\theta}_{\mathrm{ML}}(z^n))| \\ & - \log w(\hat{\theta}_{\mathrm{ML}}(z^n)) + o(1).\end{aligned} \tag{16}$$

From (15) and (16) we see that, if the *generalized Jeffreys integral* $\int v(\theta)\cdot\sqrt{|I(\theta)|}d\theta$ is finite (see the Appendix), then there is a special choice of prior w, the *generalized Jeffreys' prior*, with $w(\theta) = v(\theta)\sqrt{|I(\theta)|}/\int v(\theta)\sqrt{|I(\theta)|}d\theta$, under which $-\log\bar{p}^{\mathrm{NML}}(z^n)$ and $-\log\bar{p}^{\mathrm{BAYES}}(z^n)$ do not just coincide up to $O(1)$, but become asymptotically indistinguishable. If $v \equiv 1$, this w coincides with the well-known *Jeffreys' prior* w_{J} popular in Bayesian inference; the special case of this prior for the Bernoulli model was encountered in Example 1. Thus, the Bayesian universal distribution with the (generalized) Jeffreys' prior can be a very good alternative of p_v^{NML}.

2.3. *Unifying model selection and estimation*

Suppose we are given a countable collection of models $\{\mathcal{M}_\gamma : \gamma \in \Gamma\}$. Recall that the basic idea above was to associate each individual model $\mathcal{M}_\gamma$ with a single distribution $\bar{p}_\gamma$. It seems reasonable to do the same at the level of "meta-parameters" γ: we set $\overline{\mathcal{M}} := \{\bar{p}_\gamma : \gamma \in \Gamma\}$ and in complete analogy to (3), we define *the meta-universal distribution*

$$p_\pi^{\mathrm{NML}}(z^n) := \frac{\max_{\gamma\in\Gamma}\bar{p}_\gamma(z^n)\pi(\gamma)}{\int_{z^n}\max_{\gamma\in\Gamma}\bar{p}_\gamma(z^n)\pi(\gamma)dz^n} \tag{17}$$

for some nonnegative weight function π on Γ. It then makes sense to postulate that the best submodel $\mathcal{M}_\gamma$ for the given data z^n is given by the γ achieving the maximum in (17). Note that for determining this maximum, the denominator in (17) plays no role.

Let us assume that all $\bar{p}_\gamma$ have already been defined. Then we can use any π such that the overarching p_π^{NML} in (17) exists. We can now formulate a *general MDL Principle for model selection*: we start with a (potentially huge) set of candidate distributions $\mathcal{M}_{\mathrm{FULL}}$. We next carve up $\mathcal{M}_{\mathrm{FULL}}$ into interesting *sub-models* $\mathcal{M}_\gamma$ with $\gamma \in \Gamma$, so that $\bigcup_{\gamma\in\Gamma}\mathcal{M}_\gamma = \mathcal{M}_{\mathrm{FULL}}$. We then associate each $\mathcal{M}_\gamma$ with a universal distribution $\bar{p}_\gamma$, and we equip $\overline{\mathcal{M}}$ as defined above with luckiness function π [note that $\mathcal{M}_{\mathrm{FULL}}$, a countable union of (usually) uncountable sets, consists of all distributions under consideration, while $\overline{\mathcal{M}}$ is a countable set]. We then base the selection of a sub-model $\mathcal{M}_\gamma$ on (17). What we earlier called the "general MDL Principle" underneath (1) was the special case in which $\sum\pi(\gamma) = 1$, i.e., π is a probability mass function.

Via (9) we see that for any such probability mass function π, the denominator in (17) is well-defined, hence π is a valid luckiness function.

Now consider the special case in which every $\bar{p}_\gamma$ is chosen to be an NML distribution $p^{\mathrm{NML}}_{v_\gamma}$ for some luckiness functions v_γ. We take some function π': $\Gamma \to \mathbb{R}^+_0$ (which we will relate to the π above later on) and we set, for $\theta \in \Theta_\gamma$, $v_{\mathrm{FULL}}(\gamma, \theta) := \pi'(\gamma) v_\gamma(\theta)$. We can use $p^{\mathrm{NML}}_{v_{\mathrm{FULL}}}$ for parameter estimation on the joint parameters (γ, θ) just as we did earlier for parametric models, by using the MDL estimator $\widehat{(\gamma, \theta_\gamma)}_{v_{\mathrm{FULL}}}$ picking the (γ, θ_γ) minimizing, over $\gamma \in \Gamma, \theta_\gamma \in \Theta_\gamma$,

$$\begin{aligned} &-\log p_{\theta_\gamma}(z^n) - \log v_{\mathrm{FULL}}(\gamma, \theta_\gamma) + \mathrm{COMP}(\mathcal{M}_{\mathrm{FULL}}, v_{\mathrm{FULL}}) \\ &\quad = -\log p_{\theta_\gamma}(z^n) - \log v_\gamma(\theta) - \log \pi'(\gamma) \\ &\qquad + \mathrm{COMP}(\mathcal{M}_{\mathrm{FULL}}, v_{\mathrm{FULL}}), \end{aligned} \tag{18}$$

where again $\mathrm{COMP}(\mathcal{M}_{\mathrm{FULL}}, v_{\mathrm{FULL}})$ plays no role in the minimization. This MDL estimator really combines model selection (estimation of γ) and parametric estimation (estimation of θ_γ). If we now define $\pi(\gamma) := \pi'(\gamma)/\exp(\mathrm{COMP}(\mathcal{M}_\gamma, v_\gamma))$, we find that p^{NML}_π defined relative to model $\overline{\mathcal{M}}$ as in (17) is equal to $p^{\mathrm{NML}}_{v_{\mathrm{FULL}}}$ defined relative to the full union of models $\mathcal{M}_{\mathrm{FULL}}$, and the γ achieving the maximum in (17) coincides with the γ minimizing (18). This indicates that model selection and estimation is really the same thing with MDL: if we are given a single parametric model $\mathcal{M}_\gamma$ with luckiness v_γ, we pick the θ minimizing the first two terms in (18) for fixed γ; if we are interested in both θ and γ, we minimize over all terms; and if we are only interested in γ, we pick the γ achieving the maximum in (17), which, by construction, will give us the same γ as the joint minimization over (18).

Two-part versus one-part codes: The role of data compression

In the oldest (1978) version of MDL, only two-part codes on countable sets were used: the minimum over $\theta \in \Theta_\gamma$ was taken over a discretized grid $\ddot{\Theta}_\gamma$ and v_γ was a probability mass function over this grid; then for all γ, $\mathrm{COMP}(\mathcal{M}_\gamma, v_\gamma) \leq 0$ and $\mathrm{COMP}(\mathcal{M}, v) \leq 0$ [see (9)] and they were both approximated by 0. From the *Kraft inequality*[20] we see where the name "two-part code" comes from: this inequality says that for every probability mass function π on a countable set $\mathcal{A}$, there exists a lossless code such that for all $a \in \mathcal{A}$, the number of bits needed to encode a, is given by $-\log \pi(a)$. Thus the resulting method can be interpreted as picking the $(\ddot{\theta}, \ddot{\gamma})$ minimizing the two-stage code length of the data, where first the parameters (θ, γ) are encoded using $-\log \pi(\gamma) - \log v_\gamma(\theta)$ bits, and then z^n is encoded "with the help of γ", using $-\log p_{\theta_\gamma}(z^n)$ bits [in fact, the encoding of (γ, θ) itself has two sub-stages here so we really have a two-part code where the first part itself has two parts as well].

The discretization involved in using a probability mass function/code for continuous-valued θ makes things (unnecessarily, as was gradually discovered over the last 30 years) very complicated in general. Also, if one combines the choice of θ with the choice of γ, the approximation of $\mathrm{COMP}(\mathcal{M}_\gamma, v_\gamma)$ as 0 introduces some sub-optimalities. Thus, one would like to code the data in a way that avoids these two issues. It turns out that this can be achieved by replacing two-part by one-part codes for the data, namely, to use codes with length $-\log p^{\mathrm{NML}}_v(z^n)$: assuming for simplicity that data are discrete, the same Kraft inequality implies that there must also be a code, directly defined on z^n, which achieves code length for each z^n given by $-\log \bar{p}^{\mathrm{NML}}(z^n)$. Thus, even though for general luckiness functions v this code length cannot be decomposed into two sub-code lengths, it remains a valid code length and the name *MDL* for the resulting procedure remains, we feel, justified. In the past, it was sometimes thought by some MDL fans that two-part codes on countable sets would somehow lead to inherently better estimates $\hat{\theta}_w$ than estimators $\hat{\theta}_v$ for general luckiness functions as in (5). However, after 30 years it turned out there is nothing either conceptually or mathematically that indicates the need for two-part codes and countable sets: for any luckiness function v, the resulting procedure has a code length interpretation, and Grünwald and Mehta[13] show that all consistency and convergence results that hold for two-part estimators also hold for general MDL estimators (Sec. 6.1) — thus invalidating the conjecture in Open Problem 13 of G07 that postulated a special status for luckiness functions v that are probability mass functions on countable sets. For the luckiness function π on the discrete structure Γ, however, it *is* quite reasonable to choose a probability mass function: no probability mass is wasted [since the denominator in (17) plays no role in choosing γ], and designing π by thinking about the code lengths $-\log \pi(\gamma)$ comes very naturally, as the following example illustrates.

Example 4 (Variable selection: L_1-versus L_0-penalties). Suppose that each data point $Z_i = (X_i, Y_i)$ where Y_i denotes the variable to be predicted and $X_i = (X_{i1}, \ldots, X_{im}) \in \mathbb{R}^m$ is a vector of covariates or "features" that may or may not help for predicting Y_i. We consider a linear model $\mathcal{M} = \{p_\beta \colon \beta \in \mathbb{R}^m\}$, decomposed into sub-models $\mathcal{M}_\gamma$ expressing,

$$Y_i = \sum_{j=1}^{m} \gamma_j \beta_j X_{ij} + \epsilon_i,$$

where $\epsilon_1, \epsilon_2, \ldots$ represent zero-mean i.i.d. $N(0, \sigma^2)$ denotes the normally distributed noise and $\gamma = (\gamma_1, \ldots, \gamma_m) \in \{0,1\}^m$ is a binary vector indicating which variables are helpful for predicting the Z_i. Thus, if γ has k zero components, then $\mathcal{M}_\gamma$ is effectively an $(m-k)$-dimensional model. Our task is to learn, from the data, the vector γ^* indicating which variables are truly relevant, and/or such that predictions of new Y given new X based on $\mathcal{M}_{\gamma^*}$ are as good as possible.

In light of the above, a straightforward way to use MDL here is to pick the γ minimizing

$$-\log p_{v_\gamma}^{\mathrm{NML}}(y^n|x^n) - \log \pi(\gamma), \tag{19}$$

where we refer to Fig. 14.2 in G07 for an explanation why we can condition on x^n here. v_γ in (19) is an appropriately chosen luckiness function and π is really a probability mass function, such that $L_\pi(\gamma) := -\log \pi(\gamma)$ can be interpreted as the number of bits needed to encode γ using a particular code. In terms of coding, a natural choice for such a code would be to first encode the number of nonzero components k_γ in γ using a uniform code (that assigns equal code length to all possibilities). Since $0 \le k_\gamma \le m$, there are $m+1$ possibilities, so this takes $\log_2(m+1)$ bits. In a second stage, one encodes the location of these components. There are $\binom{m}{k_\gamma}$ possibilities here, so this takes $\log_2 \binom{m}{k_\gamma}$ bits using a uniform code. All in all, one needs

$$\log(m+1) + \log \binom{m}{k_\gamma} \tag{20}$$

"nits" (bits re-expressed in terms of natural logarithm log) to encode γ. Then (20) can be written as $-\log \pi(\gamma)$, with $\pi(\gamma) = 1/\big((m+1) \cdot \binom{m}{k_\gamma}\big)$, which, as predicted by the Kraft inequality, sums to 1 over $\gamma \in \{0,1\}^m$.

As to the left part of the code length (19), if the variance σ^2 is known, a natural luckiness function v_γ to use is a $(m-k_\gamma)$-dimensional Gaussian with mean 0 and variance $\sigma^2\Sigma$ for some (usually diagonal) covariance matrix Σ_γ. This gives (see Chap. 12 of G07)

$$\begin{aligned} -\log p_{v_\gamma}^{\mathrm{NML}}(y^n|x^n) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \hat{\beta}_{\mathrm{ML}|j} \gamma_j x_i \right)^2 \\ + \frac{n}{2} \log 2\pi\sigma^2 + \frac{1}{2} \log |\mathbf{X}^T\mathbf{X} \\ + \Sigma^{-1}| + \frac{1}{2} \log |\Sigma|, \end{aligned} \tag{21}$$

where $\mathbf{X} = (X_1, \ldots, X_n)$ and $|\cdot|$ stands for determinant. We thus end up finding the γ minimizing the sum of (21) and (20). If, as here, the noise is normal with fixed variance, then for this choice of luckiness function, $p_{v_\gamma}^{\mathrm{NML}}(y^n|x^n)$ actually *coincides* with $p_{w_\gamma}^{\mathrm{BAYES}}(y^n|x^n)$ for a particular prior w_γ, thus one has a Bayesian interpretation as well (Bartlett *et al.*[21] show that such a precise correspondence between NML and Bayes only holds in quite special cases, see the Appendix). If the variance σ^2 is unknown, one can treat it as a nuisance parameter and equip it with the improper Haar prior, leading to a modification of the formula above; see Example 6. Even if the noise is not known to be normally distributed, one can often still use the above method — pretending the noise to be normally distributed and accepting that one is misspecified — by varying the *learning rate*, as briefly explored in Sec. 6.3.

Note that the code/prior we used here induces *sparsity*: if there exists a γ with mostly zero components that already fits the data quite well, we will tend to select it, since, for $k \ll n$, $\log \binom{n}{k}$ increases approximately linearly in k. *That does not mean that we necessarily believe that the "truth" is sparse — it just expresses that we hope that we can already make reasonably good predictions with a small number of features.*

An alternative, and very popular, approach to this problem is the celebrated *Lasso*,[22] in which we consider only the full model $\mathcal{M} = \mathcal{M}_{(1,1,\ldots,1)}$ and we pick the $\beta \in \mathbb{R}^m$ minimizing

$$\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \beta_j \gamma_j x_i \right)^2 + \frac{\lambda}{2\sigma^2} \sum_{j=1}^{m} |\beta_j| \tag{22}$$

for some regularization parameter $\lambda > 0$ [the factor σ^2 plays no role in the minimization; it is incorporated only to facilitate comparison with (21)]. It is known this will tend to select β with many zero

components, thus also inducing sparsity, and it can be implemented computationally much more efficiently than the two-part approach sketched above, effectively replacing L_0-penalties by L_1-penalties. With our "modern" view of MDL, this can be thought of as a form of MDL too, where we simply impose the luckiness function $v(\beta) = \exp(-(\lambda/\sigma^2) \times \sum_{j=1}^m |\beta_j|)$ and use the estimator $\hat{\theta}_v$ given by (5). The luckiness function v depends on λ; the optimal choice of λ is then once again related to the optimal learning rate; see Sec. 6.3. Finally, we note that there exists a third MDL approach one can use here: one starts out with an NML approach similar to (19) but then performs a continuous relaxation of the resulting optimization problem; the resulting "relaxed" NML criterion is then once again tractable and similar to an L_1-optimization problem such as (22); this approach has been described by Miyaguchi and Yamanishi[23] who extend the idea to group Lasso and other settings.

2.4. *Log-loss prediction and universal distributions*

Now consider the simple case again with a finite set of models $\{\mathcal{M}_\gamma : \gamma \in \Gamma\}$ where Γ is small compared to n and we use the uniform prior π, picking the γ maximizing $\bar{p}_\gamma$. It was the fundamental insight of Rissanen[14] and Dawid[15] that such model choice by maximizing $\bar{p}_\gamma(z^n)$ for a single distribution $\bar{p}_\gamma$ can be motivated in a different way as well — in essence, it selects the model with the best predictive performance on unseen data. This approach shows that MDL is quite similar in spirit to cross-validation, the main difference with leave-one-out cross-validation being that the *cross* in cross-validation is replaced by a *forward* and that the loss function used to measure prediction error is restricted to be the logarithmic score, also commonly known as *log-loss* (which, however, is often used in cross-validation as well).

Formally, the log-loss of predicting a single outcome $z \in \mathcal{Z}$ with a distribution p is defined as $-\log p(z)$: the larger the probability density, the smaller the loss. If one predicts a sequence of n outcomes $z^n = (z_1, \ldots, z_n)$ with n predictions $p_1, p_2, \ldots, p_n$, then the *cumulative log-loss* is defined as the sum of the individual losses: $\sum_{i=1}^n -\log p_i(z_i)$.

Now, if we adopt a probabilistic world view and represent our beliefs about z^n by a probability distribution $\bar{p}$, then the obvious way to make sequential predictions is to set $p_i := \bar{p}(Z_i = \cdot | z^{i-1})$, so that $-\log p_i(z_i) = -\log \bar{p}(z_i|z^{i-1})$. For arbitrary probability distributions, we have, by the formula for conditional probability: for all $z^n \in \mathcal{Z}^n$, $p(z^n) = \prod_{i=1}^n p(z_i|z^{i-1})$. Taking logarithms gives

$$\sum_{i=1}^n -\log \bar{p}(z_i|z^{i-1}) = -\log \bar{p}(z^n). \tag{23}$$

In other words, for every possible sequence, *the cumulative log-loss obtained by sequentially predicting z_i based on the previously observed data z^{i-1} is equal to the minus-log-likelihood.*

Conversely, if we are given an arbitrary *sequential prediction strategy* $\bar{s}$ which when input with a sequence z^{i-1} of arbitrary length $i-1$ outputs a prediction for the next outcome z_i in the form of a probability distribution $\bar{s}_{z^{i-1}}$ on $\mathcal{Z}$, we can *define* $\bar{p}(z_i|z^{i-1}) := \bar{s}_{z^{i-1}}(z_i)$ and then further define $\bar{p}(z^n) := \prod_{i=1}^n \bar{p}(z_i|z^{i-1})$. A simple calculation shows that we must have $\int_{z^n \in \mathcal{Z}^n} \bar{p}(z^n) dz^n = 1$, so we have *constructed* a probability distribution $\bar{p}$ which once again satisfies (23). The fundamental insight here is that, when the log-loss is used, *every probability distribution defines a sequential prediction strategy and — perhaps more surprisingly — vice versa, every sequential prediction strategy defines a probability distribution, such that on all sequences of outcomes, the minus-log-likelihood is equal to the cumulative loss.*

Example 5 Consider again the Bernoulli model $\mathcal{M} = \{p_\theta : \theta \in [0,1]\}$. Each element $p_\theta \in \mathcal{M}$ defines a prediction strategy which, no matter what happened in the past, predicts that the probability that the next outcome $Z_i = 1$ is equal to θ. It incurs cumulative loss, on sequence z^n with n_1 ones and $n_0 = n - n_1$ zeros, given by $n_1(-\log\theta) + n_0(-\log(1-\theta))$. The Bayesian universal distribution $p_{w_{\mathrm{J}}}^{\mathrm{BAYES}}$ with Jeffreys' prior that we considered in Example 1 satisfies, as was already mentioned, $p_{w_{\mathrm{J}}}^{\mathrm{BAYES}}(Z_{m+1} = 1|z^m) = (m_1 + (1/2))/(m+1)$, so it "learns" the probability of 1 based on past data and does not treat the data as i.i.d. any more. The asymptotic expansion (16) then shows that its cumulative loss is of order

$$\begin{aligned} &-\log p_{\hat{\theta}_{\mathrm{ML}}(z^n)}(z^n) + \frac{1}{2}\log n + O(1) \\ &\quad = -n_1 \log(n_1/n) - n_0 \log(n_0/n) \\ &\qquad + \frac{1}{2}\log n + O(1). \end{aligned}$$

We may now ask: given a parametric model $\mathcal{M}$, what distribution (i.e., prediction strategy) in $\mathcal{M}$ leads to the best predictions of data $z_1, \ldots, z_n$? For simplicity we will assume that data are i.i.d. according to all $p_\theta \in \mathcal{M}$. We have to distinguish between the best sequential prediction strategy *with hindsight* and the best prediction strategy that can be formulated before actually seeing the data. The former is given by the $p_\theta \in \mathcal{M}$ achieving

$$\begin{aligned}\min_{\theta\in\Theta} \sum_{i=1}^{n} -\log p_\theta(z_i|z^{i-1}) &= \min_{\theta\in\Theta} -\log p_\theta(z^n) \\ &= -\log p_{\hat{\theta}_{\mathrm{ML}}(z^n)}(z^n),\end{aligned}$$

i.e., the best predictions with hindsight are given by the ML distribution $\hat{\theta}_{\mathrm{ML}}(z^n)$. However, $\hat{\theta}_{\mathrm{ML}}(z^n)$ is only knowable after seeing all the data z^n, whereas in reality, we have, at each time i, to make a prediction $\bar{p}(Z_i|z^{i-1})$ relying only on the previously seen data z^{i-1}. We might thus aim for a prediction strategy (distribution) $\bar{p}$ which will tend to have a small *regret* (additional prediction error)

$$\begin{aligned}\mathrm{REG}(\bar{p}, z^n) = &\sum_{i=1}^{n} -\log \bar{p}(z_i|z^{i-1}) \\ &- \left[\min_{\theta\in\Theta} \sum_{i=1}^{n} -\log p_\theta(z_i|z^{i-1})\right] \\ = &-\log \bar{p}(z^n) + \log p_{\hat{\theta}_{\mathrm{ML}}(z^n)}(z^n). \qquad (24)\end{aligned}$$

But what does "tend" mean here? One strict way to implement the idea is to require (24) to be small in the worst case — one looks for the distribution $\bar{p}$ achieving

$$\min_{\bar{p}} \max_{z^n\in\mathcal{Z}^n} \mathrm{REG}(\bar{p}, z^n), \qquad (25)$$

where the minimum is over all probability distributions on $\mathcal{Z}^n$. But comparing (24) and (25) with (13), using that $-\log$ is strictly decreasing, we see that the $\bar{p}$ achieving (25) is just the NML distribution with $v \equiv 1$, which was already our "favorite" distribution to use in MDL model comparison any way! And, just like before, if (25) has no solution, we may add a $-\log v$ luckiness term to (24) so as to regularize the problem, and then the optimal prediction strategy will be given by p_v^{NML}. We also see that with $v \equiv 1$, $\mathrm{COMP}(\mathcal{M}, v)$ is equal to the minimax regret (25); and with nonuniform v, $\mathrm{COMP}(\mathcal{M})$ will become equal to the minimax *luckiness regret*, i.e., (24) with a $-\log v$ term added.

We now also see where the idea to use the prequential plug-in distribution $\bar{p}^{\mathrm{PREQ}}$ instead of $\bar{p}^{\mathrm{NML}}$ comes from: if calculating p_v^{NML} is too difficult, or if the *horizon* n (which is needed to calculate p_v^{NML}) is unknown, we might simply pick any estimator $\breve{\theta}$ which we think is "reasonable" and replace our prediction $p_v^{\mathrm{NML}}(Z_i|z^{i-1})$ by $p_{\breve{\theta}(z^{i-1})}(Z_i)$ — if the estimator was chosen cleverly, we can expect the resulting cumulative regret to be small. Reconsidering (14), we see that all the universal distributions, viewed as prediction strategies, with the right choice of luckiness functions, priors and/or estimates, can be made to achieve a logarithmic (in n) worst-case regret — since the cumulative log-loss achieved by the best predictor in hindsight usually grows linearly in n, a logarithmic regret is quite satisfactory. Returning to our Bernoulli example, we see that the cumulative log-loss obtained by $\hat{\theta}_{\mathrm{ML}}$, the best with hindsight, is equal to $nH(n_1/n) = nH(\hat{\theta}_{\mathrm{ML}}(z^n))$, where $H(\theta)$ is the binary entropy, $H(\theta) := -\theta\log\theta - (1-\theta)\log(1-\theta)$. Note that, in line with the above discussion, $nH(\hat{\theta}_{\mathrm{ML}})$ is linear in n unless $\hat{\theta}_{\mathrm{ML}}$ tends to 0 or 1, but the regret of $\bar{p}^{\mathrm{BAYES}}$ with Jeffreys' prior is logarithmic in n.

We thus get a novel interpretation of MDL: it associates each model $\mathcal{M}_\gamma$ with a sequential prediction strategy $\bar{p}_\gamma$ that is designed to achieve small regret compared to the hindsight-optimal prediction strategy within $\mathcal{M}_\gamma$; it then picks the model for which the corresponding prediction strategy achieves the smallest cumulative loss on the data.

Related works

Dawid (see Ref. 15 and many subsequent works) suggests to use this *prequential model choice* also with respect to loss functions other than the logarithmic loss; minimax-optimal cumulative prediction strategies without making stochastic assumptions about the data, with log-loss but (mainly) with other loss functions, are one of the main topics in *machine learning theory*; see for example Ref. 24; but there they are generally not used for model comparison or selection.

Why the logarithmic score?

Why does it make sense to minimize cumulative log-loss? Outside of the MDL world, the log-loss is often used for two reasons: first, it is (essentially) the only *local proper scoring rule*.[25] Second, it has an interpretation in terms of *money*: for every sequential

prediction strategy, there is a corresponding "sequential investment" strategy such that the smaller the cumulative log-loss, the larger the monetary gains made with this strategy ("Kelly Gambling"[17,20]).

Within the MDL field however, the use of the log-loss comes from the Kraft inequality, which directly relates it to *lossless data compression*. As we already saw before Example 4, for any sequential prediction strategy, i.e., every distribution p on sequences of length n, there is a *lossless code* C such that, for all sequences of length n,

$$-\log_2 p(z^n) = \text{No. of bits needed to code } z^n \text{ using } C.$$

Conversely, for any code C, there is a corresponding distribution p such that the above holds (see Chap. 3 of G07 for a very extensive explanation). Thus, the original MDL idea to "take the model that compresses the data most" is first made more formal by replacing it by "associate each model with a code that compresses well whenever some distribution in this model compresses well", and this turns out to be equivalent to "associate each model $\mathcal{M}_\gamma$ with a distribution $\bar{p}_\gamma$ that assigns high likelihood whenever some distribution in the model assigns high likelihood".

MDL prediction and "improper" estimation. As is clear from the prequential interpretation of MDL given above, once a universal distribution $\bar{p}$ has been fixed, one can use it to *predict* Z_i given z^{i-1} by $\bar{p}(Z_i|z^{i-1})$. At least for i.i.d. data, we can estimate the underlying "true" distribution p^* based on such predictions directly, by simply interpreting $\bar{p}(Z_i|z^{i-1})$ as an estimate of p^*! This is different from the previous form of MDL estimation described in Sec. 2.3, which was based on MDL (penalized ML) estimators $\hat{\theta}_v$. Note that this standard MDL estimator is "in-model" or *proper* (to use machine learning terminology[26]), whereas $\bar{p}(Z_i|z^{i-1})$ is *out-model* or *improper*: in general, there may be no $p \in \mathcal{M}$ such that $\bar{p}(\cdot|z^{i-1}) = p$. For example, with Bayes universal distributions, $p^{\text{BAYES}}_{w_{Z_i|z^{i-1}}}$ will be a mixture of distributions in $\mathcal{M}$ rather than a single element; see G07 for more discussion.

2.5. *The luckiness function*

The choice of a luckiness function is somewhat akin to the choice of a prior in Bayesian statistics, yet — as explained at length in Chap. 17 of G07 — there are very important differences, both technically (luckiness functions are not always integrable) and philosophically. Basically, a luckiness function just determines for what type of data one will be "lucky" [$v(\hat{\theta}_v(z^n))$ large] and get small cumulative regret based on small samples (and presumably, good model selection results as well), and for what data one will be less lucky and get good results only when the dataset grows much larger — v may thus be chosen for purely pragmatic reasons. For example, as in Example 4 (see the italicized text there), if one assigns a large value $\pi(\gamma)$ to some model $\mathcal{M}_\gamma$ within a large collection of models $\{\mathcal{M}_\gamma : \gamma \in \Gamma\}$ where γ is *sparse*, one may do this because one hopes that this sub-model will already lead to *reasonable* predictions of future data, even though one feels that, at the same time, when more data becomes available, a model $\mathcal{M}_{\gamma'}$ with a much larger number of nonzero parameters may at some point almost certainly become better (Example 4). Such an interpretation is not possible with a Bayesian prior π, where a large value of $\pi(\gamma)$ indicates a strong belief that $\mathcal{M}_\gamma$ is true, or at least, that predictions based on acting as if it will be true will be optimal — a Bayesian with high prior on γ considers $\mathcal{M}_\gamma$ *likely* rather than just *useful* — a distinction worked out in detail by Grünwald.[27] Nevertheless, just like a Bayesian prior, the luckiness function has to be chosen by the user/statistician, and often contains a subjective element. Still, in contrast to Bayesian priors, since we invariably take a worst-case log-loss stance in MDL, there often is a uniquely preferable choice of luckiness function v for parametric models $\mathcal{M}$. First, if $\text{COMP}(\mathcal{M}, v) < \infty$ with uniform v and no clear prior knowledge or preference is available, then uniform v is usually preferable over other v, since it achieves the worst-case optimal prediction performance. Second, if $\text{COMP}(\mathcal{M}_\gamma, v) = \infty$ with uniform v for some $\gamma \in \Gamma$ we can often still set the first few, say m, outcomes aside and pick a luckiness function $v_\gamma(\theta) := p_\theta(z_1, \ldots, z_m)$ for $\theta \in \Theta_\gamma$. The corresponding estimator (for fixed γ) $\hat{\theta}_{v_\gamma}$ based on the data $z_{m+1}, \ldots, z_n$ as given by (5) will then be equal to the ML estimator based on the full data, $\hat{\theta}_{v_\gamma}(z^n_{m+1}) = \hat{\theta}_{\text{ML}|\gamma}(z^n)$, and by choosing m large enough, one can often get that $\text{COMP}(\mathcal{M}_\gamma, v_\gamma)$ is finite after all for all $\gamma \in \Gamma$ and one may then compare models by picking the γ maximizing $p^{\text{NML}}_{v_\gamma}(z_{m+1}, \ldots, z_n)$. Thus, the luckiness function is now determined by the first few "start-up" data points, and one uses NML based on the remaining

data points with an estimator that coincides with the ML estimator based on the full data. G07 argues why this data-driven luckiness function is the best default choice available; note that it is analogous to our use of improper Bayes priors as described in Example 2.

3. Novel Universal Distributions

3.1. *The switch distribution and the AIC–BIC dilemma*

The *AIC–BIC dilemma* (see for example Ref. 28 and the many citations therein) is a classic conundrum in the area of statistical model selection: if one compares a finite number of models, the two standard benchmark methods, with (different) asymptotic justifications, are AIC and BIC. Suppose one first selects a model using AIC or BIC. One then predicts a future data point based on, e.g., maximum likelihood estimation, or by adopting the Bayes/MDL predictive distribution, within the chosen model. If one compares a finite number of models, then AIC tends to select the one which is optimal for prediction (compared to BIC, the predictions converge faster, by a factor of order $\log n$, to the optimal ones). On the other hand, BIC is consistent: with probability 1, it selects the smallest model containing the true distribution, for all large n; the probability that AIC selects an overly large model does not go to 0 for large n. Both the predictive-optimality and the consistency property are desirable, but, like AIC and BIC, common methods all fail on one of the two. For example, MDL, with each of the four classical distributions, and Bayes factor model selection will behave like BIC for large n and be consistent but prediction-sub-optimal; for any fixed k, leave-k-out and k-fold cross-validation will tend to behave like AIC and have the reverse behavior. Yang[28] shows that, in general, this dilemma cannot be solved: every consistent method has to be slightly prediction-sub-optimal in some situations; he also shows that prediction by model averaging cannot solve this dilemma either.

Nevertheless, as first shown by van Erven *et al.*[16] (who thereby solved Open Problems 8 and 17 of G17), one can design universal distributions that "almost" get the best of both worlds: basing MDL model selection on them using (1) one gets a criterion which is strongly consistent while at the same time losing only an exceedingly small-order $\log\log n$ factor in terms of prediction quality compared to the AIC-type methods. Although it can be applied to arbitrarily large model collections, the idea of this so-called *switch distribution* $\bar{p}^{\text{SWITCH}}$ is best explained by considering the simplest case with just two nested models $\mathcal{M}_0 \subset \mathcal{M}_1$: one starts with two standard universal distributions (say, Bayesian or luckiness-NML) $\bar{p}_0$ for $\mathcal{M}_0$ and $\bar{p}_1$ for $\mathcal{M}_1$. For every $i > 0$, $\bar{p}_1$ defines a conditional distribution $\bar{p}_1(z_i, \ldots, z_n | z^{i-1})$. One now picks a "prior" distribution π on the integers [typically one that decreases polynomially, e.g., $\pi(i) = 1/(i(i+1))$], and one defines a new universal distribution for $\mathcal{M}_1$ by

$$\bar{p}^{\text{SWITCH}}(z^n) := \sum_{i=1}^{n} \pi(i) \bar{p}_1(z_i, \ldots, z_n | z^{i-1}) \cdot \bar{p}_0(z^{i-1}).$$

This distribution is best understood from the prequential interpretation of MDL (Sec. 2.4). It will satisfy

$$\begin{aligned}
\sum_{i=1}^{n} -\log \bar{p}^{\text{SWITCH}}(z_i | z^{i-1}) &= -\log \bar{p}^{\text{SWITCH}}(z^n) \\
&= -\log \sum_{i=1}^{n} \pi(i) \bar{p}_1(z_i, \ldots, z_n | z^{i-1}) \bar{p}_0(z^{i-1}) \\
&\leq \min_{i \in \{1,\ldots,n\}} -\log \pi(i) \bar{p}_1(z_i, \ldots, z_n | z^{i-1}) \cdot \bar{p}_0(z^{i-1}) \\
&\leq \min_{i} -\log \bar{p}_1(z_i, \ldots, z_n | z^{i-1}) \cdot \bar{p}_0(z^{i-1}) \\
&\quad + 2 \log n.
\end{aligned}$$

In other words, the cumulative log-loss achieved by $\bar{p}^{\text{SWITCH}}$ is "almost" (within an order $\log n$ term) as small as that of the strategy that first predicts by $\bar{p}_0$ and then *switches* from $\bar{p}_0$ to $\bar{p}_1$ at the switching point i that is optimal with hindsight. By clever choice of the prior π, one can get the extra term down to order $\log\log n$. In cases where the data are actually sampled from a distribution in $\mathcal{M}_1$ that is "close" (defined suitably) to $\mathcal{M}_0$, the predictions based on $\bar{p}^{\text{SWITCH}}$ will, with high probability, be substantially better than those based on $\bar{p}_1$ — a dramatic example (that makes very clear why this happens) is given in the first figure of Ref. 29. If the data come from a distribution that is "far" from $\mathcal{M}_0$, they will tend to be worse than those based on $\bar{p}_1$ by a negligible amount. Working out the math shows that associating $\mathcal{M}_1$ with $\bar{p}^{\text{SWITCH}}$ and $\mathcal{M}_0$ with $\bar{p}_0$ indeed gives a strongly consistent model selection criterion that is almost [to within an

$O(\log\log n)$ factor] prediction-optimal, thus almost solving the AIC–BIC dilemma. van Erven *et al.*[29] describe in great detail why the standard NML or Bayesian universal model $\bar{p}_1$ does not lead to the optimal cumulative log-loss if the data come from a distribution close to, but not in, $\mathcal{M}_0$.

In case the number of models on the list is larger than two or even infinite, one has to associate each model with a separate switch distribution. The technique for doing so is described by van Erven *et al.*[29] who also give an efficient implementation and prove consistency and prediction-optimality of the switch distribution in a weak, cumulative sense for both finite and infinite numbers of models. van der Pas and Grünwald[30] mathematically show the "almost" prediction-optimality for a finite number of models.

3.2. *Hybrids between NML Bayes and prequential plug-in*

A Problem for NML: Unknown horizon Bayesian universal distributions with fixed priors have the property that the probability assigned to any initial sequence $z^{n'}$, where $n' < n$, is independent of the total length of the sequence. For other universal models, such as NML, this is not always the case. Take for example the Bernoulli model extended to sequences by independence: For sequence length $n = 2$, the normalizing term in the NML equals $1 + (1/2)^2 + (1/2)^2 + 1 = 5/2$. For sequence length $n = 3$, the normalizing term equals $2 + 6 \times (1/3)(2/3)^2 = 78/27$. For $n = 2$, the NML probability of the sequence 00 is $1/(5/2) = 0.4$. However, for the sequence length $n = 3$, the probability of the initial sequence 00 is obtained as the sum of the probabilities of the sequences 000 and 001, which becomes $1/(78/27) + (4/27)/(78/27) \approx 0.397 < 0.4$. As shown by Ref. 21 (see also Ref. 31), there do exist cases in which NML is, like Bayes, horizon-independent, but these are very rare — see the Appendix.

The above discrepancy between the initial sequence probabilities for different sequence lengths may be a problem in situations where we need to obtain predictions without necessarily knowing the total length of the sequence, or the *horizon*. Another related issue is that even if the total sequence length was given, it can be computationally expensive to obtain marginal and conditional probabilities along the initial sequences. One possible solution would be to restrict to Bayesian universal distributions. However, while these solve the horizon issue, they are (a) still often computationally inefficient and (b) they lack NML-style worst-case regret interpretations. This has spurned research into universal codes that can be calculated without knowing the horizon in advance and that behave better as regards to (a) or (b), which we now review.

3.2.1. *Prequential plug-in and the* $(k/2)\log n$ *formula*

The most straightforward way to deal with issue (a) is to use the prequential $\bar{p}^{\mathrm{PREQ}}$ which, by construction, is horizon-independent. However, for the prequential $\bar{p}^{\mathrm{PREQ}}$ (10) the BIC asymptotics (14) only hold in expectation if the data are sampled from one of the distributions in the model $\mathcal{M}$. This makes the result much weaker than for the other five universal distributions considered, for which the asymptotics hold for every individual sequence in some large set, i.e., without making any stochastic assumptions at all. One might thus wonder what happens for general data. Extending the earlier works by Takeuchi and Barron[32] and Kotłowski *et al.*[33] shows that, if data are sampled from a distribution p, and $p_{\bar{\theta}}$ is the distribution in $\mathcal{M}$ that is closest in Kullback–Leibler (KL) divergence to p, then (14) holds in expectation, with a correction term involving the variances of both distributions; for one-dimensional models, we get

$$\frac{\mathrm{VAR}_p(Z)}{\mathrm{VAR}_{p_{\bar{\theta}}}(Z)} \cdot \frac{1}{2} \log n, \tag{26}$$

a formula that can be extended to multidimensional models and individual sequence settings. Solving Open Problem 2 from G07, Grünwald and Kotłowski[34] show that, essentially, there exists *no* "in-model" estimator that can achieve the standard asymptotics in general; a correction such as (26) is always needed, whatever estimator one tries. Here an "in-model" estimator (or "proper" estimator, see end of Sec. 2.4) is an estimator that always outputs a distribution inside the model $\mathcal{M}$; the ML and Bayes MAP estimators are in-model, but the Bayes predictive distribution is not in-model, since it is a mixture distribution over all distributions in $\mathcal{M}$.

Solving Open Problem 3 from G07, Kotłowski *et al.*[33] also provide a new universal distribution, in

which for any given estimator $\breve{\theta}$, $p_{\breve{\theta}}^{\mathrm{PREQ}}(Z_{i+1}|z^i) = p_{\breve{\theta}(z^i)}(Z_{i+1})$ is turned into a slightly "flattened" version $p_{\breve{\theta}}^{\mathrm{PREQ}*}(Z_{i+1}|z^i)$, which is not in $\mathcal{M}$ any more (it is not an in-model estimator), but it does achieve the standard $(k/2)\log n$ asymptotics without correction. For example, in case of the normal location family with fixed variance σ^2, it coincides with a Bayesian predictive distribution based on a standard conjugate prior, which in this case is a normal with mean $\breve{\mu}$ (the Bayes MAP estimate) but variance $\sigma^2 + O(1/n)$. More generally, $p_{\breve{\theta}}^{\mathrm{PREQ}*}(Z_{i+1}|z^i)$ becomes a hybrid between the estimator $\breve{\theta}$ and a Bayes predictive distribution, but it has the advantage over the latter that it can be calculated without performing an integral over the parameter space. It thus provides an alternative to the NML distribution that is horizon-free and that is often faster to compute than $\bar{p}^{\mathrm{BAYES}}$.

Roos and Rissanen[35] and Rissanen *et al.*[36] developed other prequential, horizon-free universal codes that are non-Bayesian, yet remain more closely to NML in spirit than $p_{\breve{\theta}}^{\mathrm{PREQ}*}$. They work out the details for discrete models including Bernoulli as well as linear regression models. For Bernoulli models, the resulting universal code coincides with the so-called one-step lookahead model proposed earlier by Takimoto and Warmuth.[37] For linear regression models the asymptotic consistency of the resulting model selection criterion was studied by Määttä *et al.*[38] and Määttä and Roos.[39] Relatedly, Watanabe and Roos[40] show that no horizon-independent strategy can be asymptotically minimax in the multinomial case and propose simple Bayesian universal models with a horizon-dependent Dirichlet prior that achieve asymptotic minimaxity and simplify earlier proposals. Among the proposed priors is $\mathrm{Dir}(\alpha, \ldots, \alpha)$ with $\alpha = 1/2 - \ln 2/2 \ln n$ which converges to the Jeffreys' prior $\mathrm{Dir}(1/2, \ldots, 1/2)$ but has a mild dependency on the horizon n.

3.3. *Hypothesis testing: Universal distributions based on the reverse information projection*

Suppose we compare just two models, $\mathcal{M}_0$ and $\mathcal{M}_1$, as explanations for data z^n, a situation similar to classical *null hypothesis testing*, the standard method for evaluating new treatments in the medical sciences and scientific hypotheses in most applied sciences such as psychology, biology and the like: we can think of $\mathcal{M}_0$ and $\mathcal{M}_1$ as two hypotheses, where, usually, $\mathcal{M}_0$ represents the status quo ("treatment not effective", "coin unbiased"). In case[a] $\mathcal{M}_0 = \{P_0\}$ represents a *simple* (singleton) hypothesis, there is a strong additional motivation for using MDL as a hypothesis testing method, and in particular, for quantifying the evidence against $\mathcal{M}_0$ in terms of

$$D(z^n) = -\log \bar{p}_1(z^n) - [-\log p_0(z^n)],$$

the code length or cumulative-log-loss difference (see Sec. 2.4) between encoding (or sequentially predicting) the data with p_0 and with $\bar{p}_1$. This additional motivation is given by the *no hypercompression inequality* (G07), a mathematical result stating that, no matter how $\bar{p}_1$ is defined, as long as it is a probability distribution, we have for all $K > 0$, and $0 \leq \alpha \leq 1$,

$$\begin{aligned} &P_0(D(Z^n) \leq -K) \leq 2^{-K}, \\ &\text{i.e., } P_0\left(\frac{p_0(Z^n)}{\bar{p}_1(Z^n)} \leq \alpha\right) \leq \alpha. \end{aligned} \tag{27}$$

This expresses, in terms of sequential log-loss prediction (compression), that, if P_0 is true, then the probability that one can predict data better, by K or more loss units, by predictions based on $\bar{p}_1$ rather than p_0, is exponentially small in K — and this holds independently of the sequence length n. In terms of more classical quantities, it states that, no matter how we chose $\bar{p}_1$, if P_0 holds true then the likelihood ratio is a p-value. In fact it is a conservative p-value, giving usually somewhat less evidence against $\mathcal{M}_0$ than a standard p-value, for which the rightmost inequality in (27) is an equality. The inequality (27) goes in the right direction to retain the cornerstone of classical Neyman–Pearson testing: if one sets significance level α before seeing the data and one chooses $\mathcal{M}_1$ whenever $D(Z^n) \leq -\log \alpha$, i.e., $p_0(Z^n)/\bar{p}_1(Z^n) \leq \alpha$, then the probability, under the null P_0, of making a false decision is bounded by α. But the fact that the rightmost inequality in (27) is usually strict has pleasant practical repercussions: as explored by Ref. 17, the Type-I error guarantees are retained under *optional continuation*. This is the (common)

[a]In this sub-section we view, for notational convenience, the elements of $\mathcal{M}_j$ as probability distributions P_θ with densities or mass functions p_θ.

practice to decide, on the basis of an initial sample z^n, whether or not to gather new data and do a second test. One may for example decide to gather new data if the result based on z^n was hopeful yet not conclusive. This is highly problematic for standard, strict p-value-based hypothesis testing, but with MDL testing with a simple $\mathcal{M}_0$, one can simply multiply the likelihood ratios of the two (or more) tests performed, or equivalently, add the code length differences for each test performed. The resulting code length difference/likelihood ratio will still lead to valid Type-I error bounds.[17]

But, all this holds only for simple $\mathcal{M}_0$. Yet the tests most used in practice, such as the t-test and contingency table tests, all involve *composite* $\mathcal{M}_0 = \{P_\theta : \theta \in \Theta_0\}$. For composite $\mathcal{M}_0$, the no-hypercompression inequality (27) usually only holds for *some* $P_0 \in \mathcal{M}_0$, but for Type-I error guarantees and the like we would want to have it hold for *all* P_θ with $\theta \in \Theta_0$. That is, we would like to employ universal distributions $\bar{p}_1$ and $\bar{p}_0$ such that we have

$$\text{For all } \theta \in \Theta_0 : P_\theta(D(Z^n) \leq -K) \leq 2^{-K},$$
$$\text{i.e., } P_\theta\left(\frac{\bar{p}_0(Z^n)}{\bar{p}_1(Z^n)} \leq \alpha\right) \leq \alpha. \tag{28}$$

In general, this will not hold for standard choices (NML, Bayes, prequential plug-in, etc.) of $\bar{p}_1$ and $\bar{p}_0$. However, Grünwald *et al.*[17] show that, for any given (arbitrary) $\bar{p}_1$, one can, under very mild conditions, *construct* a $\bar{p}_0$ such that (28) holds, thereby solving Open Problems 9 and 19 of G07. This $\bar{p}_0$ is the *Reverse Information Projection*[41,42] of $\bar{p}_1$ onto $\mathcal{P}_{\text{BAYES}}(\mathcal{M}_0)$, where $\mathcal{P}_{\text{BAYES}}$ is the set of densities $\bar{p}_0$ for z^n that can be written as Bayes marginal distributions $\bar{p}_0^{\text{BAYES}}(z^n) = \int p_\theta(z^n) w_0(\theta) d\theta$ for some prior w_0 on Θ_0 — for every prior w_0, $\mathcal{P}_{\text{BAYES}}(\mathcal{M}_0)$ contains a separate distribution on Z^n. The RIPr is defined as the density achieving $\min_{\bar{p}_0 \in \mathcal{P}_{\text{BAYES}}(\mathcal{M}_0)} \times D(\bar{p}_1 \| \bar{p}_0)$, where $D(\cdot\|\cdot)$ is the Kullback–Leibler divergence. Thus, one constructs a $\bar{p}_0$ with the desired no-hypercompression property, and at the same time, it will minimize KL divergence to $\bar{p}_1$, which implies that if data were sampled from $\bar{p}_1$, it would yield optimal log-loss predictions. This, in turn, implies that the $\bar{p}_0$ constructed this way will satisfy the standard asymptotics (14) as long as the $\bar{p}_1$ on which it is based does. Based on the likelihood ratio between $\bar{p}_1$ and its RIPr $\bar{p}_0$, one is also allowed to do optional continuation while retaining Type-I Error guarantees. Thus, even if one is an adherent of classical, frequentist testing theory, there are strong reasons for MDL-style testing based on the RIPr universal distribution. Grünwald *et al.*[17] further extend the reasoning to give guidelines on how $\bar{p}_1$ can be chosen to get further good frequentist properties.

Example 6 (Right-Haar priors and the Bayesian *t*-test). In a series of papers (highlights include Refs. 43 and 44), Berger and collaborators established Bayes factor testing methods for composite $\mathcal{M}_0 = \{P_\theta : \theta \in \Theta_0\}$ where the only free parameters in Θ_0 are "nuisance" parameters that are shared by Θ_1 and are governed by a group structure. A prime example is the unknown variance in the t-test. Berger uses a special type of improper prior, the so-called *right-Haar* prior, which can be defined for every such type of nuisance parameter. While Bayes factors usually do not combine well with improper priors, the Bayes factors for group invariance parameters equipped with the right-Haar prior behave remarkably well. Grünwald *et al.*[17] show that, even though the right-Haar priors are usually improper, they can also be understood from a purely MDL perspective: if $\bar{p}_1^{\text{BAYES}}$ and $\bar{p}_0^{\text{BAYES}}$ are equipped with the right-Haar prior on the nuisance parameters, and the prior on the additional parameters in $\bar{p}_1^{\text{BAYES}}$ satisfies some additional requirements, then both $\bar{p}_1^{\text{BAYES}}$ and $\bar{p}_0^{\text{BAYES}}$ can be interpreted as sequential prediction strategies, and the log of the Bayes factor can be interpreted as the code length/cumulative-log-loss difference. Moreover, $\bar{p}_0^{\text{BAYES}}$ is (essentially) the RIPr for $\bar{p}_1^{\text{BAYES}}$ and the no-hypercompression inequality (28) that is so desirable from a frequentist perspective holds uniformly for all $\theta_0 \in \Theta_0$.

Let us consider the one-sample Bayesian t-test as an example. Here $\mathcal{M}_0 = \{p_{0,\sigma} : \sigma > 0\}$ is the set of all normal distributions with mean 0; the variance σ^2 is a free parameter. $\mathcal{M}_1 = \{p_{\mu,\sigma} : \mu \in \mathbb{R}, \sigma > 0\}$ is the set of all normal distributions with as free parameters μ and σ. The question of interest is to establish whether $\mu = 0$ or not; σ is an unknown "nuisance" parameter — it determines the scale of the data but is not itself of intrinsic interest. In the *Bayesian t-test* one equips both $\mathcal{M}_0$ and $\mathcal{M}_1$ with the improper right-Haar prior, $w(\sigma) = 1/\sigma$. To complete the definition of $\bar{p}_1^{\text{BAYES}}$, $\mathcal{M}_1$ is equipped with a conditional prior density (given σ) on the effect size $\delta := \mu/\sigma$. This second density has to be symmetric around zero and proper (this is what we

called the "additional requirement on the prior on Θ_1", instantiated to the case where the nuisance parameter is a variance). One now proceeds by testing using the Bayes factor $\bar{p}_0^{\,\mathrm{BAYES}}/\bar{p}_1^{\,\mathrm{BAYES}}$. In this special case, the procedure was already suggested by Jeffreys,[45] and the right-Haar prior coincides with Jeffreys' prior on the variance. Berger *et al.* extend the method to general group-invariant parameter vectors such as the joint mean and variance in the two-sample t-test, testing a Weibull against a log-normal and many other scenarios.

4. Graphical Models

Graphical models are a framework for representing multivariate probabilistic models in a way that encompasses a wide range of well-known model families, such as Markov chains, Markov random fields and Bayesian networks; for a comprehensive overview, see Ref. 46. A key property of a graphical model is parsimony, which can mean, for instance, a low-order Markov chain or more generally a sparse dependency graph that encodes conditional independence assumptions. Choosing the right level of parsimony in graphical models is an ideal problem for MDL model selection.

In Bayesian network model selection the prevailing paradigm is, unsurprisingly, the Bayesian one. Especially the works of Geiger and Heckerman[47] and Heckerman *et al.*[48] have been extremely influential. The main workhorse of this approach is the so-called Bayesian Dirichlet (BD) family of scores which is applicable in the discrete case where the variables being modeled are categorical. Given a data sample, such scores assign a goodness value to each model structure. Exhaustive search for the highest scoring structure is possible when the problem instance (characterized by the number of random variables) is of limited size, but heuristic search techniques such as variants of local search or "hill-climbing" can be used for larger problem.

Different forms of the BD score imply different Dirichlet priors (different hyper-parameters) for the local multinomial distributions that comprise the joint distribution. For example, in the commonly used BDeu score, the priors are determined by a single hyper-parameter, α. For a variable X_i with r distinct values and parents Pa_i that can take q possible combinations of values (*configurations*), the BDeu prior is $\mathrm{Dir}(\alpha/rq, \ldots, \alpha/rq)$. One of the main motivations for adopting this prior is that it leads to *likelihood equivalence*, i.e., it assigns equal scores to all network structures that encode the same conditional independence assumptions. In light of the fact that Bayesian model selection embodies a particular form/variation of MDL, these methods fit, at least to some extent, in the MDL framework as well. However, there also exist more "pure", non-Bayesian MDL methods for model selection in Bayesian networks; we mention Refs. 49 and 50 as early representative examples. These early methods are almost invariably based on the two-part coding framework. More recently, several studies have proposed new model selection criteria that exploit the NML distribution. One approach is a continuous relaxation of NML-type complexities proposed by Miyaguchi *et al.*[51] in which the model selection problem takes on a tractable Lasso-type L_1-minimization form (see also Example 4). In other approaches, NML [or usually, approximations (but not relaxations) thereof] are used directly for encoding parts of the model; we now describe these latter approaches in a bit more detail.

4.1. *Factorized NML and variants*

Silander *et al.*[52] propose the *factorized NML* (fNML) score for Bayesian network model selection which was designed to be *decomposable* meaning that it can be expressed as a sum that includes a term for each variable in the network. This property facilitates efficient search among the super-exponential number of possible model structures; see, e.g., Ref. 48. The fNML score factors the joint likelihood not only in terms of the variables but also in terms of distinct configurations of the parent configurations. Each factor in the product is given by a multinomial NML probability, for which a linear-time algorithm by Kontkanen and Myllymäki[53] can be used.

A similar idea where a Bayesian network model selection criterion is constructed by piecing together multiple NML models under the multinomial model was proposed recently by Silander *et al.*[54] In the proposed *quotient NML* (qNML) score, the local scores corresponding to each variable in the network are defined as log-quotients of the form

$$\log \frac{\mathrm{NML}_{\mathrm{FULL}}(X_i \cup \mathrm{Pa}_i)}{\mathrm{NML}_{\mathrm{FULL}}(\mathrm{Pa}_i)},$$

where $\mathrm{NML}_{\mathrm{FULL}}$ refers to an NML distribution defined by using a fully connected network to model the variable X_i and its parents Pa_i in the numerator and the same thing for the parent set Pa_i in the denominator. Technically, this amounts to collapsing the configurations of the variables into distinct values of a single categorical variable. Even though the resulting categorical variable may have a huge number of possible values, the linear time algorithm[53] or efficient approximations (see the next sub-section) can be used to implement the computations. A notable property of the qNML score is that, unlike the fNML score, it is likelihood-equivalent (see above).

Eggeling *et al.*[55] apply similar ideas to a different model class, namely parsimonious Markov chains. There too, the likelihood is decomposed into factors depending on the configurations of other variables, and each part in the partitioning is modeled independently using the multinomial NML formula. The authors demonstrate that the fNML-style criterion they propose leads to parsimonious models with good predictive accuracy for a wide range of different scenarios, whereas the corresponding Bayesian scores are sensitive to the choice of the prior hyper-parameters, which is important in the application where parsimonious Markov chains are used to model DNA binding sites.[56]

In all these papers, both simulated and real-world data experiments suggest that the MDL-based criteria are quite robust with respect to the parameters in the underlying data source. In particular, the commonly used Bayesian methods (such as the BDeu criterion) that are being used as benchmarks are much more sensitive and fail when the assumed prior is a poor match to the data-generating model, whereas the MDL methods are invariably very close to the Bayesian methods with the prior adapted to fit the data. This poses interesting questions concerning the proper choice of priors in the Bayesian paradigm.

In fact, the prevalence of the Bayesian paradigm and the commonly used BD scores is challenged by two recent observations: First, Silander *et al.*[52] show that the Dirichlet prior with hyper-parameters $(1/2, \ldots, 1/2)$, which is the invariant Jeffreys' prior for the multinomial model, but not likelihood-equivalent when used in the BD score, is very close to the fNML model and consequently, enjoys better robustness properties than the BDeu score which is the likelihood-equivalent BD score variant. Second, Suzuki[57] shows that the BDeu criterion is *irregular*, i.e., prone to extreme overfitting behavior in situations where a deterministic relationship between one variable and a set of other variables holds in the data sample. The MDL scores discussed above are regular in this respect and their robustness properties seem to be better than those of the BD scores, see Ref. 54.

4.2. *Asymptotic expansions for graphical models*

Asymptotic results concerning MDL-based criteria in graphical models are interesting for several reasons. For one, they lead to efficient scores that can be evaluated for thousands of different model structures. Second, asymptotic expansions can lead to insights about the relative complexity of different model structures.

Various asymptotic forms exist for the point-wise and the expected regret depending on the model class in question. For convenience we repeat the classical expansion of the NML (as well as the Bayesian marginal likelihood with Jeffreys' prior) regret/model complexity that applies for regular model classes $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$ for which $\mathrm{COMP}(\mathcal{M}, v)$ is finite with uniform v (see Sec. 2.2 above):

$$\mathrm{COMP}(\mathcal{M}, v) = \frac{k}{2} \log \frac{n}{2\pi} + \int_\Theta \sqrt{|I(\theta)|} d\theta + o(1), \tag{29}$$

where k is the dimension of the model, $|I(\theta)|$ is the determinant of the Fisher information matrix at parameter θ, the integral is over the parameter space Θ and the remainder term $o(1)$ vanishes as the sample size tends to infinity.

For discrete data scenarios, by far the most interesting case is the multinomial model (extension of the Bernoulli distribution to an i.i.d. sequence of r-valued categorical random variables) since it is a building block of a number of MDL-criteria such as fNML and qNML (see above). There are many asymptotic expansions for the NML regret under the multinomial model. Probably the most useful is the one proposed by Szpankowski and Weinberger[58]:

$$n\left(\log \alpha + (\alpha + 2) \log C_\alpha - \frac{1}{C_\alpha}\right) - \frac{1}{2} \log \left(C_\alpha + \frac{2}{\alpha}\right), \tag{30}$$

where n is the sample size, $\alpha = \frac{r}{n}$ and $C_\alpha = \frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{4}{\alpha}}$. This simple formula is remarkably accurate over a wide range of finite values of n and r (see Ref. 54). Note that the leading term is proportional to n (rather than $\log n$ as usual) because the formula is derived for the regime $r = \Theta(n)$ where the alphabet size grows proportionally to the sample size. If r grows slower than n or not at all, the leading term tends to the classical form (29), where the leading term is $\frac{k}{2}\log n$. In practice, the approximation (30) is applicable for a wide range of r/n ratios.

Roos,[59] and Zou and Roos[60] studied the second term in the expansion (29), namely the Fisher information integral, under Markov chains and Bayesian networks using Monte Carlo sampling techniques. This approach reveals systematic differences between the complexities of models even if they have the same number of parameters.

5. Latent Variable and Irregular Models

Although thus far we have highlighted exponential family and regression applications, NML and other universal distributions can of course be used for model selection and estimation in complete generality — and many practical applications are in fact based on highly irregular models. Often, "classical" two-part distributions (based on discretized models) are used, since NML distributions often pose computational difficulties. However, Yamanishi and collaborators have managed to come up with tractable approximations of NML-type distributions for some of the most important irregular (i.e., non-exponential family) models such as hierarchical latent variable models,[61] and the related Gaussian mixture models.[62,63] Suzuki *et al.*[64] provide an NML approach to nonnegative matrix factorization. Two-part codes (and corresponding MDL estimators) for mixture families that come close to achieving the minimax regret were considered very recently by Miyamoto *et al.*[65]

When it comes to asymptotic approximations for code lengths/log-likelihoods based on NML and other universal distributions — all approximations so far (in Sec. 2.2) were derived essentially assuming that the model under consideration is an exponential family. Extensions to curved exponential families and generalized linear models are relatively straightforward (see G07 for details). For more irregular models, Watanabe has proposed the widely applicable information criterion (WAIC) and the widely applicable Bayesian information criterion (WBIC), see Refs. 66 and 67, where the latter can be viewed as an asymptotic expansion of the log-likelihood based on a Bayesian universal distribution. It coincides with BIC when applied to regular models but is applicable even for singular (irregular) models. The asymptotic form of WBIC is

$$\mathrm{WBIC}(\mathcal{M}) = -\log p_{\theta_0}(z^n) + \lambda \log n + O_p(\sqrt{\log n}), \tag{31}$$

where θ_0 is the parameter value minimizing the Kullback–Leibler divergence from the model to the true underlying distribution, and $\lambda > 0$ is a rational number called the real log-canonical threshold (see Ref. 67), which can be interpreted as the effective number of parameters (times two).

6. Frequentist Convergence of MDL and Its Implications

Rissanen first formulated the MDL Principle as — indeed — a *Principle*: one can simply start by *assuming*, as an axiom, that modeling by data compression (or, equivalently, sequential predictive log-loss minimization) is the right thing to do. One can also take a more conventional, frequentist approach, and check whether MDL procedures behave desirably under standard frequentist assumptions. We now review the results that show that, in general, they do — thus providing a frequentist justification of MDL ideas: with some interesting caveats, MDL model selection is typically *consistent* (the smallest model containing the true distribution is eventually chosen, with probability one) and MDL prediction and estimation achieves good *rates of convergence* (the Hellinger distance between the estimated and the true density goes to zero, with high probability, quite *fast*). In this section we review the most important convergence results. In particular, Sec. 6.1 shows that the link between data compression and consistent estimation is in fact very strong; and Sec. 6.4 shows that, by taking MDL as a *principle*, one can get useful intuitions about deep questions concerning deep learning; and the intuitions can then, as a second step, be once again validated by frequentist results.

Thus, let us assume, as is standard in frequentist statistics, that data are drawn from a distribution in

one of the models under $\mathcal{M}_\gamma$ under consideration. We consider consistency and convergence properties of the main MDL procedures in their main applications: model selection, prediction and estimation.

Model selection

For model selection between a finite number of models, all universal codes mentioned here are consistent in wide generality; for example, this has been explicitly proven if the data are i.i.d. and all models on the list are exponential families, but results for more complex models with dependent data have also been known for a long time; see G07 for an overview of results. If the collection of models is countably infinite, then results based on associating each $\mathcal{M}_\gamma$ with $\bar{p}_\gamma^{\text{BAYES}}$ have also been known for a long time; such results typically hold for "almost all" (suitable defined) distributions in all $\mathcal{M}_\gamma$; again, see G07 for a discussion of the (nontrivial) "almost all" requirement. These countable-Γ consistency results were extended to the switch distribution by van Erven *et al.*[29]

Prediction and "improper" estimation

As to sequential prediction (Sec. 2.4), the rate of convergence results are very easy to show (see Chap. 15 of G07), but these typically only demonstrate that the *cumulative*-log-loss prediction error of sequentially predicting with a universal distribution $\bar{p}$ behaves well as n increases. Thus, since the sum of prediction errors is small, say (for parametric models) of order $\log n$, for *most* t the individual prediction error at the tth sample point must be of order $1/t$, since $\sum_{t=1}^{n} 1/t - \log t = O(1)$. Still, it remains an open question how to prove for individual t what exactly the expected prediction error is at that specific n. Since one can view each prediction as an "improper" estimate (end of Sec. 2.4), the convergence rates of the resulting estimators, which estimate the underlying distribution based on a sample of size t as $\bar{p}(Z_{t+1}|z^t)$, usually also behave well in a cumulative sense, but again it is very hard to say anything about individual t. The asymptotic expansions (15) and (16) imply that, for fixed parametric models $\mathcal{M}_\gamma$, $\bar{p}_\gamma^{\text{BAYES}}$ and $\bar{p}_\gamma^{\text{NML}}$ achieve optimal cumulative prediction and estimation errors. If, however, they are defined relative to a full model class $\mathcal{M} = \bigcup_{\gamma\in\Gamma}\mathcal{M}_\gamma$ consisting of at least two nested models, then they may fail to achieve optimal rates by a $\log n$ factor. van Erven *et al.*[29] show that sequential prediction/estimation based on the switch distribution achieves the minimax-optimal cumulative prediction/estimation error rates even in such cases. van der Pas and Grünwald[30] show that, if only two models are compared, then the optimal obtainable rate for individual n for any consistent procedure is achieved as well.

6.1. *Frequentist convergence of MDL estimation*

Very strong results exist concerning the convergence of MDL estimation based on an MDL estimator $\hat{\theta}_v$ as given by (5). A first, classical result was already stated by the ground-breaking work,[68] establishing that consistency and good convergence rates can be obtained for the special case of a two-part-code estimator $\hat{\theta}_w$ based on a probability mass function w, as long as w satisfies $\sum_{\theta\in\ddot{\Theta}} w(\theta)^\eta < \infty$ for some $\eta < 1$ and w puts sufficient prior mass in a KL neighborhood of the true θ. These results were greatly extended by Zhang[69,70] and further, very recently, by Grünwald and Mehta.[13] The latter consider $\hat{\theta}_v$ for general v. Let $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ be a statistical model and suppose data are i.i.d. $\sim p$ with $p = p_{\theta^*} \in \mathcal{M}$. They find that a sufficient condition for consistency is that $v(\theta^*) < \infty$ and that for some $\eta < 1$, the following *generalized model complexity*

$$\begin{aligned}\text{COMP}_\eta(\mathcal{M}, v) &:= \text{COMP}(\mathcal{M}_\eta, v)\\ &= \log \int p(z^n)^{1-\eta} \\ &\quad \cdot \frac{(p_{\hat{\theta}_v}(z^n))^\eta v(\hat{\theta}_v(z^n))}{\left(\int p(\underline{z})^{1-\eta}(p_{\hat{\theta}_v(z^n)}(\underline{z}))^\eta d\underline{z}\right)^n} dz^n\end{aligned} \tag{32}$$

is bounded and $o(n)$, i.e., it grows slower than linear (the slower it grows, the faster the MDL estimator converges to the true distribution in Hellinger distance). This condition strictly and significantly weakens the Barron–Cover requirement. The result holds without any further conditions; for example, $\mathcal{M}$ may be a countable union of parametric models or even a huge nonparametric model. Note that $\text{COMP}_1(\mathcal{M}, v)$ is the model complexity that we have encountered before in (7). Ironically, for any $\eta < 1$, slow growth ($o(n)$) of $\text{COMP}_\eta(\mathcal{M})$ is sufficient for consistency of $\hat{\theta}_v$, but for $\eta = 1$, which would be more fully in line with the MDL ideas, it is not.

6.2. *From MDL to Lasso*

As illustrated in Example 4, when used for high-dimensional variable selection, the original MDL approach would be to use a mixed two-part/one-part code as in (1) with a $-\log \pi(\gamma)$ term to account for the model index $\gamma \in \Gamma$. In such settings, there may well be $p > n$ variables of interest, each of which may or may not be included in the model, so that the minimization over Γ requires trying out $2^p \gg 2^n$ choices — which is practically infeasible. For this reason, in practice people have strongly preferred the *Lasso* and related methods based on L_1-penalties, which take linear rather than exponential time in p (note that the classic MDL essentially penalizes by an L_0-penalty). However, Barron and Luo,[71] and Barron *et al.*[72] showed that, under some conditions on the true distribution (such as Gaussian noise), the Lasso method can be re-interpreted in terms of code length minimization after all; see also Ref. 73, and, for further extensions, Refs. 74 and 75. For a different approach to unify model selection with very high-dimensional models with the luckiness NML, see Miyaguchi and Yamanishi.[23]

Although some of the details may differ, it seems that most of these works are subsumed by the aforementioned result of Grünwald and Mehta[13] who show that general penalized estimators can be re-interpreted as minimizing a one-part code length as long as $\text{COMP}_1(\mathcal{M}, v)$ is bounded, and can be proven consistent under the (still quite weak) condition that COMP_η as in (32) is bounded for some $\eta < 1$. Thus, the connection between MDL and general (including Lasso and other L_1-penalties, but also with entirely different penalties) penalization methods is substantially stronger than it seemed before these developments took place.

Supervised machine learning

Importantly, all the works mentioned here except Ref. 13 cannot show convergence under misspecification — for example, when applied to the Lasso, they would require an assumption of normal noise (corresponding to the squared error used in the Lasso fit, which is equivalent to the log-loss under a normal distribution for the noise). In practice though, the Lasso (with the squared error) is often used in cases in which one cannot assume normally distributed errors. Reference 13 contains results that can still be used in such cases [although the formula for $\text{COMP}_\eta(\mathcal{M}, v)$ changes], based on ideas which we sketch in the following sub-section.

More generally, one of the major areas within machine learning is *supervised learning* in which one assumes that data $(X_1, Y_1), (X_2, Y_2), \ldots$ are i.i.d. $\sim P_0$, with $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$, and one aims to use the data to learn a predictor function $f: \mathcal{X} \to \mathcal{Y}'$ that has small *expected loss* or *risk*, defined as $\mathbf{E}_{(X,Y)\sim P}[\ell(f(X), Y)]$, where $\ell: \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}$ is some loss function of interest and f is a member of some "predictor model" $\mathcal{F}$. For example, the statistical notion of "regression with random design" corresponds, in machine learning, to a supervised learning problem with $\mathcal{Y} = \mathcal{Y}' = \mathbb{R}$ and $\ell(y', y) = (y' - y)^2$. Early MDL convergence results do not cover this "supervised" situation: they are not equipped to handle either random design or loss functions beyond the log-loss. Some of the more recent works mentioned above are able to handle random design but not general loss functions (for example, for Lasso-type applications they require the noise to be normally distributed). Reference 13 seems to be the first that can fully handle supervised learning scenarios: the convergence results can be used with random design, and they can also be used with large classes of loss functions including squared error (without normality assumption) and zero/one-loss. This is achieved by associating predictors f with densities $p_f(x, y) \propto \exp(-\ell(f(x), y))$, so that the log-loss relative to density p_f on data (x, y) becomes linearly related to the loss of f on (x, y); the analysis then proceeds via analyzing convergence of MDL for the densities $\{p_f : f \in \mathcal{F}\}$ as a misspecified probability model.

6.3. *Misspecification*

As beautifully explained by Rissanen,[76] one of the main original motivations for MDL-type methods is that they have a clear interpretation independent of whether any of the models under consideration is "true" in the sense that it generates the data: one chooses a model minimizing a code length, i.e., a prediction error on unseen data, which is meaningful and presumably might give something useful irrespective of whether the model is true (Rissanen even argues that the whole notion of a "true model" is misguided). This model-free paradigm also leads one to define the NML distribution as minimizing prediction error in a stringent worst-case-over-all data sense [Eq. (13)] rather than a stochastic sense.

Nevertheless, it is of interest to see what happens if one samples data from a distribution for which all models under consideration are wrong, but some are quite useful in the sense that they lead to pretty good predictions. Doing this leads to rather unpleasant surprises: as first noted by Grünwald and Langford,[77] MDL (and Bayesian inference) can become inconsistent: one can give examples of $\{\mathcal{M}_\gamma : \gamma \in \Gamma\}$ with countably infinite Γ and a "true" data generating distribution P_0 such that, when data are sampled i.i.d. from P_0, MDL will tend to select a sub-optimal model for all large n — while all sub-models $\mathcal{M}_\gamma$ are wrong, one of them, $\mathcal{M}_{\tilde{\gamma}}$ is optimal in several intuitive respects (closest in KL divergence to P_0, leading to best predictions under a number of loss functions), yet it will not be selected for large n. While the models considered by Grünwald and Langford[77] were quite artificial, Grünwald and van Ommen[78] showed that the same can happen in a more natural linear regression setting; moreover, they also showed that even if Γ is finite, although then *eventually* MDL will select the best sub-model, for even relatively large n it may select arbitrarily bad sub-models. De Heide[79] shows that the problem also occurs with MDL and Bayesian regression with some real-world datasets.

It turns out that the root of the problem is related to the no-hypercompression property (27). If the collection of models $\mathcal{M} = \bigcup_{\gamma\in\Gamma} \mathcal{M}_\gamma$ contains the density p_0 of the "true" distribution P_0, then any distribution $p \in \bigcup_{\gamma\in\Gamma} \mathcal{M}_\gamma$ will satisfy no-hypercompression relative to the true p_0:

$$P_0\left(\frac{p_0(Z^n)}{p(Z^n)} \leq \alpha\right) \leq \alpha. \tag{33}$$

This property underlies the proof of all MDL consistency and rate-of-convergence results, such as those by Barron and Cover,[68] Zhang,[69] and Grünwald and Mehta.[13] However, if the model class $\mathcal{M}$ does not contain the true p_0, then, in order to prove consistency, one needs (33) to hold with the P_0 outside the brackets unchanged, but the p_0 inside the brackets replaced by $\tilde{p}$, the distribution/density in $\mathcal{M}$ that is closest to P_0 in KL divergence (why it should be KL is explained at length by Grünwald and van Ommen[78]). Unfortunately though, (33) does not necessarily hold with p_0 replaced by $\tilde{p}$. If it does not, MDL (and Bayesian methods, whose consistency relies on similar properties) may become inconsistent. Grünwald and van Ommen[78], based on earlier ideas in Refs. 80 and 81, propose a solution that works for Bayesian universal distributions: it replaces the likelihoods $p_\theta(z^n)$ for every $p = p_\theta$ with $p \in \mathcal{M}$ by the *generalized likelihood* $p_\theta^\eta(z^n)$ for some $\eta > 0$; usually $\eta < 1$ — this η has the same mathematical function as the η appearing in (32). It turns out that with such a modification, if η is chosen small enough, a version of the no-hypercompression inequality (33) holds after all. References 78 and 81 also provide a method for learning η from the data, the "Safe Bayesian" algorithm (note that η cannot be learned from the data by standard MDL or Bayesian methods). The recent work of Grünwald and Mehta[13] suggests that the modification of likelihoods by exponentiating with η should work for general MDL methods as well.

6.4. *PAC-MDL bounds and deep learning*

One of the great mysteries of modern *deep learning* methods in machine learning is the following[82]: deep learning is based on *neural network* models which can have many millions of parameters. Although typically run on very large training samples z^n, n is usually still so small that the data can be fit perfectly, with zero error on the training set. Still, the trained models often perform very well on future test sets of data. How is this possible? At first sight this contradicts the tenet, shared by MDL and just about any other method of statistics, that good generalization requires the models to be "small" or "simple" [small $\text{COMP}(\mathcal{M})$ in MDL analyses, small VC dimension or small entropy numbers in statistical learning analyses] relative to the sample size. One of several explanations (which presumably all form a piece of the puzzle) is that the local minimum of the error function found by the training method is often very *broad* — if one moves around in parameter space near the minimum, the fit hardly changes. Hochreiter and Schmidhuber[83] already observed that describing weights in sharp minima requires high precision in order to not incur non-trivial excess error on the data, whereas flat minima can be described with substantially lower precision, thus forging a connection to the MDL idea; in fact related ideas already appear in Ref. 84. In these papers, the MDL Principle is used in a manner that is less direct than what was done thus far in this paper: we (and, usually, Barron and Rissanen) *directly* hunt for the shortest description of the data.

In contrast, the aforementioned authors simply note that, *no matter how* a vector of parameters for a model was obtained, if, with the obtained vector of parameters, the data can be compressed substantially, for example by coding first the parameters and then the data with the help of the parameters, then, if we believe the MDL Principle, with these parameters the model (network) should generalize well to future data. In modern practice, neural networks are often trained with stochastic gradient descent (SGD), and it has been empirically found that networks that generalize well do tend to have parameters lying in very flat minima.

While this use of the MDL Principle seems less precise than what we reviewed earlier in this paper, it can once again be given a frequentist justification, and this justification is mathematically precise after all: the so-called *PAC-Bayesian generalization bounds*[85] show that the generalization performance of any classifier can be directly linked to a quantity that gets smaller as soon as one needs (a) less bits to describe the parameter and as soon as one needs (b) less bits to describe the data given the parameters; both the results and their proofs are very similar to the MDL convergence results by Barron and Cover,[68] Zhang,[69,70] and Grünwald and Mehta.[13] Although in general, the formulation is not as straightforward as a simple sum of the two description lengths (a) and (b), the connections between both the two-part code length and the Bayesian code length are quite strong, as was already noticed by Blum and Langford.[86] In particular, for discrete Θ, such PAC-Bayes bounds contain a term $-\log \pi(\theta)$ which can be interpreted as the number of bits needed to encode θ using the codes based on some distribution π; for general, uncountable Θ, this term gets replaced by a KL divergence term that can still be related to a code length via a so-called "bits back argument" pioneered by Hinton and van Camp.[84] Dziugaite and Roy,[87] and Zhou *et al.*,[82] inspired by earlier work by Langford and Caruana,[88] indeed show that, for some real-world datasets, one can predict nontrivial generalization using deep neural nets by looking at the number of bits needed to describe the parameters and applying PAC-Bayesian bounds.

7. Concluding Remarks

We have given a self-contained introduction to MDL, incorporating and highlighting recent developments. Of necessity, we had to make a choice as to what to cover in detail, and there are many things we omitted. We would like to end with briefly mentioning three additional developments. First, there has always been the question about how MDL relates to other complexity notions such as those considered in the *statistical learning theory* literature[26]: Vapnik–Chervonkis dimension, entropy numbers, Rademacher complexity and so on. A major step towards understanding the relation was made by Grünwald and Mehta[13] who show that for probability models with members of the form $p_\theta(z) \propto \exp(-\eta \, \mathrm{LOSS}_\theta(z))$, where LOSS is an arbitrary bounded loss function, the NML complexity can be precisely bounded in terms of the Rademacher complexity defined relative to LOSS. Second, we should note that Rissanen's own views and research agenda have steered in a direction somewhat different from the developments we describe: Rissanen[89] published *Information and Complexity in Statistical Modeling*, which proposes foundations of statistics in which no underlying "true model" is ever assumed to exist. As Rissanen writes, "even such a well-meaning statement as "all models are wrong, but some are useful", is meaningless unless some model is 'true'." Rissanen expands MDL and NML ideas in the direction of the *Kolmogorov structure function*, taking the idea of distinguishable distributions underlying Ref. 19 as the fundamental; while presumably compatible with the developments we describe here, the emphasis of this work is quite different.

We end with a word about applications: since 2007, numerous applications of MDL and MDL-like techniques have been described in the literature; as discussed in Sec. 6.2, highly popular methods such as Lasso and Bayes factor methods can often be seen as "MDL-like". Even as to specific "pure" MDL applications (such as based on NML and two-part codes), the number and scope of applications are simply too large to give a succinct representative overview. However, there is one particular area which we would like to mention specifically, since that area had hardly seen any MDL applications before 2007 whereas nowadays such applications are flourishing: this is the field of *data mining*. Some representative publications are Refs. 90–92. Most of this work centers on the use of two-part codes, but sometimes NML and other sophisticated universal distributions/codes are used as well.[93]

Appendix A. When the Original ($v \equiv 1$) NML is Undefined: Details, Open Problems and Their Solutions

The original NML distribution $\bar{p}^{\mathrm{NML}}$ with uniform v relies on the existence of the Shtarkov integral $\int_{z^n \in \mathcal{Z}^n} p_{\hat{\theta}_{\mathrm{ML}}(z^n)}(z^n) dz^n$; its asymptotic expansion (15) relies on the existence of the *Jeffreys integral* $\int \sqrt{|I(\theta)|} d\theta$ being finite, the latter being equivalent to the requirement that Jeffreys' prior is proper. Both are quite strong requirements; for infinite sample spaces $\mathcal{Y}$, they "usually" — that is, in most models one considers in practice, such as normal, exponential, Poisson, etc. — do not hold; but once one restricts the parameter space to an INECCSI set, they generally do hold. This may lead one to conjecture that the Shtarkov integral is finite *if and only if* the corresponding Jeffreys integral is finite. Resolving this conjecture was posed as an open problem by G07; Grünwald and Harremoës[94] and Bar-Lev *et al.*[95] show that, in general, the conjecture is wrong; though, for exponential families, under a very mild additional condition, it holds true.

From a more practical perspective, one would of course like to know what universal distribution to use if the standard MDL is undefined. Several proposals floated around in the early 2000s; for an overview, see Chap. 11 of G07. By now, the dominant method has become to factor in a nonuniform weight function v and calculate the *luckiness NML* as in (11). This method was originally called *luckiness NML-2* by G07, which (among many other methods) identified several "luckiness" versions of NML that had been proposed by various authors; luckiness NML-2 turned out both more practically useable and mathematically analyzable than other methods, and in this text we simply call it *luckiness NML*. In particular, Suzuki and Yamanishi[11] show that, for exponential family models, the n-dimensional integral in the luckiness NML can be replaced by a $2k$-dimensional one, and in many cases can be performed explicitly. As we indicated in Sec. 2.5, one can sometimes set the first m examples aside as start-up data to define a luckiness function, leading to *conditional NML*. Again, G07 defined different forms of conditional NML, and again, conditional NML-2 (directly based on luckiness NML-2) turned out to be the most natural one: Bartlett *et al.*[21] show that for some important classes of models, the NML distributions $\bar{p}^{\mathrm{NML}}$ and the Bayes marginal distributions $\bar{p}^{\mathrm{BAYES}}$ with improper Jeffreys' prior *exactly*, and not just asymptotically, coincide for each n. Moreover, for the case of one-dimensional families, they completely characterize the class of models for which this holds: essentially, it holds for exponential families that are also location or scale families, i.e., the normal and gamma distributions, and monotone transformations thereof (such as, e.g., the Rayleigh distributions); as well as for one curious additional family. This correspondence between objective Bayesian and conditional NML-2 approaches notwithstanding, Kojima and Komaki[96] show that "conditional NML-3", which G07 considered the most intuitive version, but at the same time, mathematically overly complicated for practical use, can be given a practical implementation after all, thereby solving the Open Problem 7 of G07.

Acknowledgments

We would like to thank Kenji Yamanishi, Matthijs van Leeuwen and Bin Yu for providing references to new MDL work, Jun-Ichi Takeuchi for prompting us to write this paper and Jorma Rissanen and Andrew Barron for many fruitful conversations over the last 20 years. This work is part of the research programme *Safe Bayesian Inference* with Project No. 617.001.651, which is financed by the Dutch Research Council (NWO).

References

1. J. Rissanen, Modeling by the shortest data description, *Automatica* **14** (1978) 465–471.
2. J. Rissanen, *Stochastic Complexity in Statistical Inquiry* (World Scientific, Hackensack, 1989).
3. A. Barron, J. Rissanen and B. Yu, The minimum description length principle in coding and modeling, *IEEE Trans. Inform. Theory* **44**(6) (1998) 2743–2760.
4. P. Grünwald, *The Minimum Description Length Principle* (The MIT Press, Cambridge, 2007).
5. P. D. Grünwald, I. J. Myung and M. A. Pitt, *Advances in Minimum Description Length: Theory and Applications* (The MIT Press, 2005).
6. C. S. Wallace and D. M. Boulton, An information measure for classification, *Comput. J.* **11** (1968) 185–195.

7. P. M. B. Vitányi and M. Li, Minimum description length induction, Bayesianism, and Kolmogorov complexity, *IEEE Trans. Inform. Theory* **IT-46**(2) (2000) 446–464.
8. T. F. Sterkenburg, Universal prediction: A philosophical investigation, Ph.D. thesis, University of Groningen (2018).
9. R. E. Kass and A. E. Raftery, Bayes factors, *J. Am. Stat. Assoc.* **90**(430) (1995) 773–795.
10. Yu. M. Shtarkov, Universal sequential coding of single messages, *Probl. Inf. Transm.* **23**(3) (1987) 3–17.
11. A. Suzuki and K. Yamanishi, Exact calculation of normalized maximum likelihood code length using Fourier analysis, arXiv:1801.03705 [math.ST].
12. J. Rissanen, Fisher information and stochastic complexity, *IEEE Trans. Inform. Theory* **42**(1) (1996) 40–47.
13. P. Grünwald and N. Mehta, A tight excess risk bound via a unified PAC-Bayesian-Rademacher-Shtarkov-MDL complexity, in *Proc. Thirtieth Conf. Algorithmic Learning Theory (ALT 2019)* (2019), arXiv:1720.07732 [stat.ME].
14. J. Rissanen, Universal coding, information, prediction and estimation, *IEEE Trans. Inform. Theory* **30** (1984) 629–636.
15. A. P. Dawid, Present position and potential developments: Some personal views, statistical theory, the prequential approach, *J. R. Stat. Soc. A* **147**(2) (1984) 278–292.
16. T. van Erven, P. D. Grünwald and S. de Rooij, Catching up faster in Bayesian model selection and model averaging, in *Advances in Neural Information Processing Systems*, Vol. 20 (Curran Associates, Inc. 2008), pp. 417–424.
17. P. Grünwald, R. de Heide and W. Koolen, Safe testing, arXiv: 1906. 07801 [math.ST].
18. C. E. Rasmussen and Z. Ghahramani, Occam's razor, in *Advances in Neural Information Processing Systems*, Vol. 13 (The MIT Press, 2000), pp. 294–300.
19. I. J. Myung, V. Balasubramanian and M. A. Pitt, Counting probability distributions: Differential geometry and model selection, *Proc. Natl. Acad. Sci. USA* **97** (2000) 11170–11175.
20. T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 1991).
21. P. Bartlett, P. Grünwald, P. Harremoës, F. Hedayati and W. Kotlowski, Horizon-independent optimal prediction with log-loss in exponential families, in *Proc. 26th Conf. Learning Theory (COLT 2013)* (2013).
22. T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Springer Verlag, 2001).
23. K. Miyaguchi and K. Yamanishi, High-dimensional penalty selection via minimum description length principle, *Mach. Learn.* **107**(8–10) (2018) 1283–1302.
24. N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning and Games* (Cambridge University Press, Cambridge, 2006).
25. A P. Dawid, The geometry of proper scoring rules, *Ann. Inst. Stat. Math.* **59**(1) (2007) 77–93.
26. S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms* (Cambridge University Press, 2014).
27. P. Grünwald, Safe probability, *J. Stat. Plan. Inference* **195** (2018) 47–63.
28. Y. Yang, Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation, *Biometrica* **92**(4) (2005) 937–950.
29. T. van Erven, P. D. Grünwald and S. de Rooij, Catching up faster by switching sooner: A predictive approach to adaptive estimation with an application to the AIC–BIC dilemma, *J. R. Stat. Soc. B (Stat. Methodol.)* **74**(3) (2012) 361–417.
30. S. van der Pas and P. D. Grünwald, Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in nested model selection, *Stat. Sin.* **28** (2018) 229–253.
31. A. Barron, T. Roos and K. Watanabe, Bayesian properties of normalized maximum likelihood and its fast computation, in *Proc. IEEE Int. Symp. Information Theory* (IEEE Press, 2014), pp. 1667–1671.
32. J. Takeuchi and A. R. Barron, Robustly minimax codes for universal data compression, in *Proc. Twenty-First Symp. Information Theory and Its Applications (SITA '98)* (1998).
33. W. Kotłowski, P. Grünwald and S. de Rooij, Following the flattened leader, in *Proc. 23rd Conf. Learning Theory (COLT)* (2010), pp. 106–118.
34. P. Grünwald and W. Kotłowski, Prequential plug-in codes that achieve optimal redundancy rates even if the model is wrong, in *Proc. 2010 Int. Symp. Information Theory (ISIT)* (2010).
35. T. Roos and J. Rissanen, On sequentially normalized maximum likelihood models, in *Proc. First Workshop Information Theoretic Methods in Science and Engineering (WITMSE-2008)* (Tampere International Center for Signal Processing, 2008).
36. J. Rissanen, T. Roos and P. Myllymäki, Model selection by sequentially normalized least squares, *J. Multivariate Anal.* **101**(4) (2010) 839–849.
37. E. Takimoto and M. Warmuth, The last-step minimax algorithm, in *Proc. Eleventh Int. Conf. Algorithmic Learning Theory (ALT-2000)* (2000).

38. J. Määttä, D. F. Schmidt and T. Roos, Subset selection in linear regression using sequentially normalized least squares: Asymptotic theory, *Scand. J. Stat.* **43**(2) (2016) 382–395.
39. J. Määttä and T. Roos, Robust sequential prediction in linear regression with Student's t-distribution, in *Proc. Fourteenth Int. Symp. Artificial Intelligence and Mathematics (ISAIM-2016)* (2016).
40. K. Watanabe and T. Roos, Achievability of asymptotic minimax regret by horizon-dependent and horizon-independent strategies, *J. Mach. Learn. Res.* **16**(11) (2015) 2357–2375.
41. J. Q. Li, Estimation of mixture models, Ph.D. thesis, Yale University, New Haven, CT (1999).
42. J. Q. Li and A. R. Barron, Mixture density estimation, in *Advances in Neural Information Processing Systems*, eds. S. A. Solla, T. K. Leen and K.-R. Müller, Vol. 12 (The MIT Press, Cambridge, 2000), pp. 279–285.
43. J. O. Berger, L. R. Pericchi and J. A. Varshavsky, Bayes factors and marginal distributions in invariant situations, *Sankhyā, Indian J. Stat. A (1961–2002)* **60**(3) (1998) 307–321.
44. S. C. Dass and J. O. Berger, Unified conditional frequentist and Bayesian testing of composite hypotheses, *Scand. J. Stat.* **30**(1) (2003) 193–210.
45. A. Ly, J. Verhagen and E. J. Wagenmakers, Harold Jeffreys' default Bayes factor hypothesis tests: Explanation, extension, and application in psychology, *J. Math. Psychol.* **72** (2016) 19–32.
46. D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (The MIT Press, 2009).
47. D. Geiger and D. Heckerman, A characterization of the Dirichlet distribution with application to learning Bayesian networks, in *Proc. 11th Conf. Uncertainty in Artificial Intelligence (UAI-1995)* (1995), pp. 196–207.
48. D. Heckerman, D. Geiger and D. Chickering, Learning Bayesian networks: The combination of knowledge and statistical data, *Mach. Learn.* **20**(3) (1995) 197–243.
49. W. Lam and F. Bacchus, Learning Bayesian belief networks: An approach based on the MDL principle, *Comput. Intell.* **10**(3) (1994) 269–293.
50. R. R. Bouckaert, Probabilistic network construction using the minimum description length principle, in *Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, M. Clarke, R. Kruse and S. Moral, eds. Lecture Notes in Computer Science, Vol. 747 (Springer, 2005), pp. 41–48.
51. K. Miyaguchi, S. Matsushima and K. Yamanishi, Sparse graphical modeling via stochastic complexity, in *Proc. 2017 SIAM Int. Conf. Data Mining (SDM2017)* (2017), pp. 723–731.
52. T. Silander, T. Roos and P. Myllymäki, Learning locally minimax optimal Bayesian networks, *Int. J. Approx. Reason.* **51**(5) (2010) 544–557.
53. P. Kontkanen and P. Myllymäki, A linear-time algorithm for computing the multinomial stochastic complexity, *Inf. Process. Lett.* **103**(6) (2007) 227–233.
54. T. Silander, J. Leppä-aho, E. Jääsaari and T. Roos, Quotient normalized maximum likelihood criterion for learning Bayesian network structures, in *Proc. Int. Conf. Artificial Intelligence and Statistics* (2018), pp. 948–957.
55. R. Eggeling, T. Roos, P. Myllymäki and I. Grosse, Robust learning of inhomogeneous PMMs, in *Proc. Seventeenth Int. Conf. Artificial Intelligence and Statistics* (2014), pp. 229–237.
56. R. Eggeling, T. Roos, P. Myllymäki and I. Grosse, Inferring intra-motif dependencies of DNA binding sites from chip-seq data, *BMC Bioinformatics* **16** (2015) 375:1–375:15.
57. J. Suzuki, Jeffreys' and BDeu priors for model selection, in *Proc. 9th Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-2016)* (2016).
58. W. Szpankowski and M. J. Weinberger, Minimax pointwise redundancy for memoryless models over large alphabets, *IEEE Trans. Inform. Theory* **58**(7) (2012) 4094–4104.
59. T. Roos, Monte Carlo estimation of minimax regret with an application to MDL model selection, in *Proc. IEEE Information Theory Workshop 2008 (ITW-2008)* (IEEE Press, 2008), pp. 284–288.
60. Y. Zou and T. Roos, On model selection, Bayesian networks, and the Fisher information integral, *New Generat. Comput.* **35**(1) (2017) 5–27.
61. T. Wu, S. Sugawara and K. Yamanishi, Decomposed normalized maximum likelihood codelength criterion for selecting hierarchical latent variable models, in *Proc. ACM Int. Conf. Knowledge Discovery and Data Mining* (2017).
62. S. Hirai and K. Yamanishi, An upper bound on normalized maximum likelihood codes for Gaussian mixture models, arXiv:1709.00925 [CS.IT].
63. S. Hirai and K. Yamanishi, Efficient computation of normalized maximum likelihood codes for Gaussian mixture models with its applications to clustering, *IEEE Trans. Inform. Theory* **59**(11) (2013) 7718–7727.
64. A. Suzuki, K. Miyaguchi and K. Yamanishi, Structure selection for convolutive non-negative matrix factorization using normalized maximum likelihood coding, in *Proc. 2016 IEEE 16th Int. Conf. Data Mining (ICDM)* (IEEE, 2016), pp. 1221–1226.
65. K. Miyamoto, A. R. Barron and J. Takeuchi, Improved MDL estimators using local exponential

family bundles applied to mixture families, in *Proc. Int. Symp. Information Theory 2019* (2019).
66. S. Watanabe, Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *J. Mach. Learn. Res.* **11** (2010) 3571–3594.
67. S. Watanabe, A widely applicable Bayesian information criterion, *J. Mach. Learn. Res.* **14** (2013) 867–897.
68. A. R. Barron and T. M. Cover, Minimum complexity density estimation, *IEEE Trans. Inform. Theory* **37**(4) (1991) 1034–1054.
69. T. Zhang, From ϵ-entropy to KL entropy: Analysis of minimum information complexity density estimation, *Ann. Stat.* **34**(5) (2006) 2180–2210.
70. T. Zhang, Information theoretical upper and lower bounds for statistical estimation, *IEEE Trans. Inform. Theory* **52**(4) (2006) 1307–1321.
71. A. Barron and X. Luo, MDL procedures with $L1$ penalty and their statistical risk, in *Proc. 2008 Workshop on Information Theoretic Methods in Science and Engineering* (2008).
72. A. Barron, C. Huang, J. Li and X. Luo, The MDL principle, penalized likelihoods, and statistical risk, in *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday,* eds. P. Grünwald *et al.* (Tampere International Center for Signal Processing, 2008), pp. 33–63.
73. S. Chatterjee and A. Barron, Information theoretic validity of penalized likelihood, in *Proc. 2014 IEEE Int. Symp. Information Theory (ISIT)* (IEEE, 2014), pp. 3027–3031.
74. M. Kawakita and J.-I. Takeuchi, Barron and Cover's theory in supervised learning and its application to Lasso, in *Proc. 33rd Int. Conf. Machine Learning* (2016), pp. 1958–1966.
75. W. Brinda and J. Klusowski, Finite-sample risk bounds for maximum likelihood estimation with arbitrary penalties, *IEEE Trans. Inform. Theory* **64**(4) (2018) 2727–2741.
76. J. Rissanen, Complexity of models, in *Complexity, Entropy and the Physics of Information*, ed. W. H. Zurek (Addison-Wesley, 1991), pp. 117–125.
77. P. Grünwald and J. Langford, Suboptimal behavior of Bayes and MDL in classification under misspecification, *Mach. Learn.* **66**(2–3) (2007) 119–149.
78. P. Grünwald and T. van Ommen, Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it, *Bayesian Anal.* **12**(4) (2017) 1069–1103.
79. R. de Heide, The Safe–Bayesian Lasso, Master's thesis, Leiden University (2016).
80. P. D. Grünwald, Viewing all models as "probabilistic", in *Proc. Twelfth ACM Conf. Computational Learning Theory (COLT' 99)* (1999), pp. 171–182.
81. P. Grünwald, The safe Bayesian: Learning the learning rate via the mixability gap, in *Proc. 23rd Int. Conf. Algorithmic Learning Theory (ALT '12)* (Springer, 2012), pp. 169–183.
82. W. Zhou, V. Veitch, M. Austern, R. Adams and P. Orbanz, Compressibility and generalization in large-scale deep learning, arXiv:1804.05862 [Stat.ML].
83. S. Hochreiter and J. Schmidhuber, Flat minima, *Neural Comput.* **9**(1) (1997) 1–42.
84. G. E. Hinton and D. van Camp, Keeping the neural networks simple by minimizing the description length of the weights, in *Proc. Sixth Annu. Conf. Computational Learning Theory* (ACM, 1993), pp. 5–13.
85. D. McAllester, PAC-Bayesian stochastic model selection, *Mach. Learn.* **51**(1) (2003) 5–21.
86. A. Blum and J. Langford, PAC-MDL bounds, in *Proc. Sixteenth Conf. Learning Theory (COLT' 03)* (2003), pp. 344–357.
87. G. K. Dziugaite and D. M. Roy, Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data, in *Proc. Thirty-Third Conf. Uncertainty in Artificial Intelligence (UAI '17)* (2017).
88. J. Langford and R. Caruana, (Not) bounding the true error, in *Advances in Neural Information Processing Systems 14*, eds. T. G. Dietterich, S. Becker, and Z. Ghahramani (The MIT Press, 2002), pp. 809–816.
89. J. Rissanen, *Information and Complexity in Statistical Modeling* (Springer-Verlag, New York, 2007).
90. J. Vreeken, M. van Leeuwen and A. Siebes, Krimp: Mining itemsets that compress, *Data Min. Knowl. Disc.* **23**(1) (2011) 169–214.
91. D. Koutra, U. Kang, J. Vreeken and C. Faloutsos, Summarizing and understanding large graphs, *Stat. Anal. Data Min., ASA Data Sci. J.* **8**(3) (2015) 183–202.
92. K. Budhathoki, J. Vreeken and J. Origo, Causal inference by compression, *Knowl. Inf. Syst.* **56**(2) (2018) 285–307.
93. N. Tatti and J. Vreeken, Finding good itemsets by packing data, in *Eighth IEEE Int. Conf. Data Mining* (IEEE, 2008), pp. 588–597.
94. P. Grünwald and P. Harremoës, Finiteness of redundancy, regret, Shtarkov sums, and Jeffreys integrals in exponential families, in *Proc. IEEE Int. Symp. Information Theory* (IEEE, 2009), pp. 714–718.
95. S. K. Bar-Lev, D. Bshouty, P. Grünwald and P. Harremoës, Jeffreys versus Shtarkov distributions associated with some natural exponential families, *Stat. Methodol.* **7**(6) (2010) 638–643.
96. M. Kojima and F. Komaki, Relations between the conditional normalized maximum likelihood distributions and the latent information priors, *IEEE Trans. Inform. Theory* **62**(1) (2016) 539–553.