Chapter 12

Model Selection with Informative Normalized Maximum Likelihood: Data Prior and Model Prior

Jun Zhang
University of Michigan

12.1. Introduction

Model selection has, in the last decade, undergone rapid growth for evaluating models of cognitive processes, ever since its introduction to the mathematical/cognitive psychology community (Myung, Forster, & Browne, 2000). The term "model selection" refers to the task of selecting, among several competing alternatives, the "best" statistical model given experimental data. To avoid ambiguity, "best" here has a now-standard operational definition – the commonly accepted criterion is that models must not only show reasonable goodness-of-fit in accounting for existing data, but also demonstrate some kind of simplicity so that it would not capture sampling noise in the data. This criteria, emphasizing generalization as opposed to fitting as the goal of modeling, embodies Occam's Razor, the principle of offering parsimonious explanation of data with fewest assumptions. Though mathematical implementations may differ, resulting in the various methods such as AIC, BIC, MDL, etc., each invariably boils down to balancing two aspects of model evaluation, one measuring its goodness-of-fit over existing data and the other measuring its complexity or capability for generalization.

The Minimum Description Length (MDL) Principle (Rissanen, 1978, 1983, 1996, 2001) is an information theoretic approach to inductive inference with roots in algorithmic coding theory. It has become one of the most popular means for model selection (Grünwald, Myung, & Pitt, 2005; Grünwald, 2007). Under this approach, data are viewed as codes to be compressed by the model. The goal of model selection is to identify the model, from a set of candidate models, that permits the shortest description length (code) of the data. The state-of-the-art of MDL approach to model selection has evolved into using the so-called Normalized Maximum Likelihood, or NML for short (Rissanen, 1996, 2001), as the criterion for model selection. In this chapter, this framework is revisited, and then modified by formally introducing the notion of "data prior". This turns the (non-informative) NML framework into the "informative" NML framework, which carries Bayesian interpretations. Informative NML subsumes (the traditional, non-informative) NML for the case of data prior being uniform, much in the same way that Bayesian inference subsumes maximal likelihood inference for the case of prior over hypotheses (parameters) being uniform.

12.2. A Revisit to NML

12.2.1. *Construction of normalized maximal likelihood*

Denote the set of probability distributions f over some sample space $\mathcal{X}$ as[1]

$$\mathcal{B} = \{f : \mathcal{X} \to [0,1],\ f > 0,\ \sum_{x\in\mathcal{X}} f(x) = 1\}\ .$$

We will use the term "model class", denoted by $\mathcal{M}_\gamma$ with a structural index γ, to specifically refer to a parametric family $\mathcal{M}_\gamma$ of probability distributions all of functional form

$$\mathcal{M}_\gamma = \{f(\cdot|\theta) \in \mathcal{B},\ \forall\theta \in \Theta \subseteq \Re^m\}\ \ ;$$

in other words, for any fixed θ,

$$f(x|\theta) > 0\ , \qquad \sum_{x\in\mathcal{X}} f(x|\theta) = 1\ .$$

The NML distribution $p^*(x)$ computed from the entire model class is, by definition,

$$p^*(x) = \frac{f(x|\hat{\theta}(x))}{\sum_{y\in\mathcal{X}} f(y|\hat{\theta}(y))}, \tag{12.1}$$

where $\hat{\theta}(\cdot)$ denotes the maximum likelihood estimator

$$\hat{\theta}(x) = \text{argmax}_\theta f(x|\theta)\ . \tag{12.2}$$

[1] We assume, for ease of exposition, that sample space $\mathcal{X}$ is discrete and hence use the summation notation $\sum_{x\in\mathcal{X}}\{\cdot\}$. When $\mathcal{X}$ is uncountable, then f is taken to be the probability density function with the summation sign replaced by $\int_X\{\cdot\}d\mu$ where $\mu(dx) = d\mu$ is the background measure on $\mathcal{X}$.

Note that, in general $p^*(x)$ itself may not be a member of the family $\mathcal{M}_\gamma$ of the distributions in question,

$$p^*(\cdot) \notin \mathcal{M}_\gamma,$$

because it is obtained by a) selecting *one* parameter $\hat{\theta}(x)$ (and hence one distribution in $\mathcal{M}_\gamma$) for *each point* x of the sample space $\mathcal{X}$ (i.e., for each data point), then b) using the corresponding value of the distribution function $f(x|\hat{\theta}(x))$, and finally c) normalizing across all possible data points $x \in \mathcal{X}$. The NML distribution is a *universal* distribution, in the sense of being generated from the family $\mathcal{M}_\gamma$ (i.e., an entire class) of probability distributions; it (generally) does not, however, correspond to any individual distribution within that family. See Figure 1.

12.2.2. *Code length, universal distribution, and complexity measure*

In algorithmic coding theory, the negative logarithm of a distribution corresponds to the "code length". Under this interpretation, $p^*(x)$ is identified as the length of an ideal code for a model class

$$\text{ideal code length} = -\log p^*(x) = -\log f(x|\hat{\theta}(x)) + \log \sum_{y \in \mathcal{X}} f(y|\hat{\theta}(y)). \tag{12.3}$$

For arguments of such coding scheme being "ideal" in the context of model selection, see Myung, Navarro, and Pitt (2006). It suffices to point out that as a criterion for model selection, the two terms in (12.3) describe on the one hand the goodness-of-fit of a model with its best-fitting parameter (first term) and on the other the complexity of a model class (second term). Therefore, the general philosophy of NML falls in the same spirit of properly balancing two opposing tensions in model construction, namely, better approximation versus lower complexity, to achieve the goal of best generalizability.

Note that in (12.1) the probability that the universal distribution p^* assigns to the observed data x is proportional to the maximized likelihood $f(x|\hat{\theta}(x))$, and the normalizing constant

$$C_\gamma = \sum_{y \in \mathcal{X}} f(y|\hat{\theta}(y)) \tag{12.4}$$

is the sum of maximum likelihoods of all potential data that could be observed in an experiment. It is for this reason that p^* is often called the normalized maximum likelihood (NML) distribution associated with model

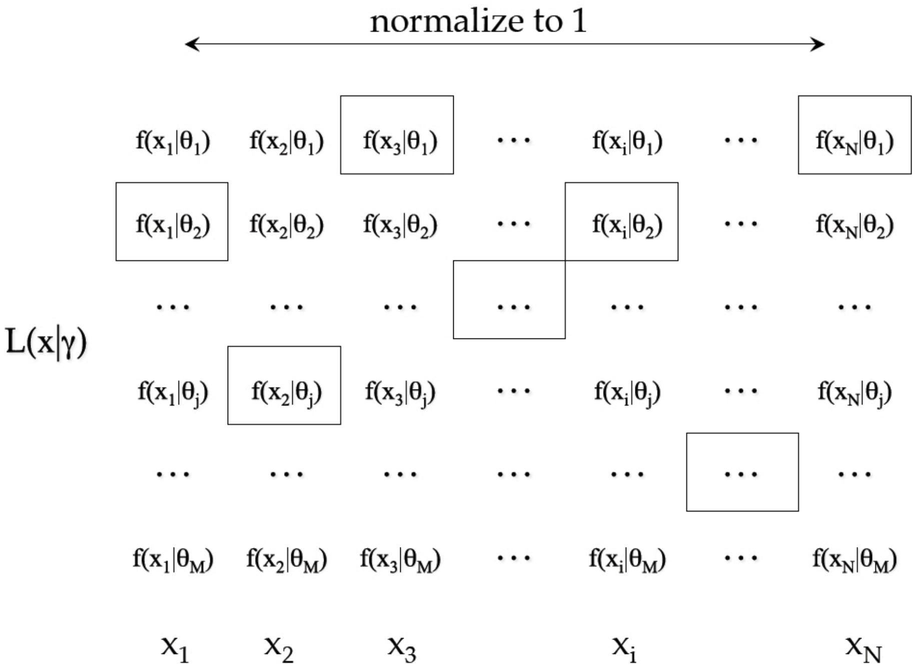

normalize to 1
L(x|γ)
f(x1|θ1) f(x2|θ1) f(x3|θ1) ··· f(xi|θ1) ··· f(xN|θ1)
f(x1|θ2) f(x2|θ2) f(x3|θ2) ··· f(xi|θ2) ··· f(xN|θ2)
··· ··· ··· ··· ··· ··· ···
f(x1|θj) f(x2|θj) f(x3|θj) ··· f(xi|θj) ··· f(xN|θj)
··· ··· ··· ··· ··· ··· ···
f(x1|θM) f(x2|θM) f(x3|θM) ··· f(xi|θM) ··· f(xN|θM)
x1 x2 x3 xi xN

Fig. 12.1. Figure 1. Schematic diagram of normalized maximum likelihood (NML) for a model class M_γ whose likelihood functions $f(\cdot|\theta)$ are parameterized by θ. Each row represents the probability density (mass) indexed by a particular θ_j value as parameter, so each row sums to 1. On the bottom, x represents all possible data, with each data point x_i "selecting" (across the corresponding column) a particular $\hat{\theta}$ with the largest likelihood value, indicated by a box. The ML function $f(x|\hat{\theta}(x))$, which is also denoted $L(x|\gamma) \equiv L_\gamma(x)$, is a map from x to the largest likelihood value shown in the box. Their sum, denoted $\hat{C}_\gamma$, may not equal 1. Normalizing $f(x|\hat{\theta}(x))$ by $\hat{C}_\gamma$ gives the NML function.

class $\mathcal{M}_\gamma$. The NML distribution is specified once the functional (i.e., parametric) form of the model class is given. It is determined prior to an experiment, that is, prior to any specific data point x being given. Model complexity, as represented by the second term of (12.3), is operationalized as the logarithm of the *sum of all best-fits* a model class can provide collectively. This complexity measure therefore formalizes the intuition that the model that fits almost every data pattern very well would be much more complex than a model that provides a relatively good fit to a small set of data patterns but does poorly otherwise.

The NML distribution p^* is derived as a solution to a *minimax problem*: Find the distribution that minimizes the worst-case average regret

(Rissanen, 2001):

$$p^* \longleftarrow \inf_{q \in \mathcal{B}} \sup_{g \in \mathcal{B}} \mathrm{E}_g \left\{ \log \frac{f(y|\hat{\theta}(y))}{q(y)} \right\} \tag{12.5}$$

where p, q ranges over the entire $\mathcal{B}$, the set of all probability distributions, and $\mathrm{E}_g\{\cdot\}$ denotes the taking of expectation

$$\mathrm{E}_g\{F(y)\} = \sum_{y \in \mathcal{X}} g(y) F(y) \ .$$

The solution, p^*, is *not* constrained to be in the set $\mathcal{M}_\gamma$. The basic idea of this minimax approach to model selection is to identify a single probability distribution that is "universally" representative of an entire parametric family of distributions and that mimics the behavior of any member of that family in the sense formulated in (12.5) (Barron, Rissanen, & Yu, 1998; Hansen & Yu, 2001). Since its computation does not invoke or even assume the existence of a true, data-generating distribution, the NML distribution is said to be "agnostic" from the truth distribution (Myung, Navarro, & Pitts, 2006), though such claim about "agnosticity" is the subject of some debate (Karabatsos & Walker, 2006; Grünwald & Navarro, 2009; Karabatsos & Walker, 2009). The debate is centered around whether the Bayesian approach under a non-informative Dirichlet process prior can be viewed as identical to that of maximal likelihood estimator, and whether the choice of a particular form of penalty function is a priori motivated.

12.2.3. *NML and Bayesianism with non-informative prior*

Under asymptotic expansion, the negative logarithm of the NML distribution can be shown (Rissanen, 1996) to be:

$$-\log p^*(x) = -\log f(x|\hat{\theta}(x)) + \frac{k}{2} \log \left(\frac{n}{2\pi} \right) + \log \int_\Theta \sqrt{\det \mathrm{I}(\theta)}\, d\theta + o(1) \tag{12.6}$$

where n denotes the sample size, k is the number of model parameters, and $\mathrm{I}(\theta)$ is the Fisher information matrix

$$\mathrm{I}(\theta) = \sum_{x \in \mathcal{X}} f(x|\theta) \frac{\partial \log f(x|\theta)}{\partial \theta^i} \frac{\partial \log f(x|\theta)}{\partial \theta^j} \ .$$

The expression (12.6) was called the "Fisher information approximation (FIA) to the NML criterion" (Pitt, Myung, & Zhang, 2002). The first two terms are known as the Bayesian Information Criterion (BIC; Schwartz, 1978). The third term of (12.6) involving the Fisher information also appeared from a formulation of Bayesian parametric model selection (Balasubramanian, 1997). This hints at the deeper connection between NML approach and Bayesian approach to model selection. We elaborate here.

In Bayesian model selection, the goal is to choose, among a set of candidate models, the one with the largest value of the marginal likelihood for observed data x, defined as

$$p_{\text{Bayes}}(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta \tag{12.7}$$

where $\pi(\theta)$ is a prior on the parameter space. A specific choice is the Jeffrey's prior $\pi_J(\theta)$, which is non-informative

$$\pi_J(\theta) = \frac{\sqrt{\det I(\theta)}}{\int_{\Theta} \sqrt{\det I(\theta)} d\theta} \ .$$

An analysis by Balasubramanian (1997) shows that if $\pi_J(\theta)$ is used in (12.7), then an asymptotic expansion of $-\log p_{\text{Bayes}}(x)$ yields an expression with the same three leading terms as in (12.6). In other words, for large n, Bayesian model selection with (the non-informative) Jeffrey's prior and NML become virtually indistinguishable. This observation parallels the findings by Takeuchi & Amari (2005) that the asymptotic expressions of various estimators, including MDL, projected Bayes estimator, bias-corrected MLE, each of which indexes a point (value of θ) in the model manifold, were related to the choice of priors; this in turn has an information geometric interpretation (Matsuzoe, Takeuchi, & Amari, 2006).

Note that in the NML approach, data is assumed to be drawn from the sample space according to a uniform distribution: the summation $\sum_{x\in\mathcal{X}}$ treats every data x with the same weight. In algorithmic coding applications, this is not a problem because here the data are the symbols under transmission which can be pre-defined to occur equally likely by the encoder and the decoder. In model selection applications where data will most likely be generated from a non-uniform distribution, care must be taken to calculate such quantities like (12.4). If the summation is taken over the stream of data that follow each other (i.e., as the data generation process is being realized), then the multiplicity in any sample value x will be naturally taken into account. On the other hand, if the summation is taken a priori (i.e., the data generation process is being assumed), then proper weighting of the data stream is called for. In this contribution, we explore a generalization of the NML formulation about model complexity measure by explicitly considering the modeler's prior belief about data and prior belief of the model classes ("prior" in comparison with data collecting data and model fitting).

12.3. NML with Informative Priors

Recall that the normalizing constant in (12.1) is obtained by first finding the maximum likelihood value for each sample point and then summing all such maximum likelihood values across the sample space. An implicit assumption behind this definition of model complexity is that every sample point is equally likely to occur *a priori* (i.e., before data collection). In terms of the Bayesian language, this amounts to assuming no prior information about possible data patterns. In this sense, NML may be viewed as a "non-informative" MDL method.

In practice, however, it is common that information about the possible patterns of data is available prior to data collection. For example, in a memory retention experiment, one can expect that the proportion of words recalled is likely to be a decreasing function of time rather than an increasing function, that retention performance will be in general worse under free recall than under cued recall, that the rate in which information is forgotten or lost in memory will be greater for uncommon, low frequency words than for common, high frequency words, etc. Such prior information implies that not all data patterns are equally likely. It would be advantageous to incorporate such information in the model selection process. The exposition below explores the possibility of developing an "informative" version of NML.

12.3.1. *Universal distribution with data-weighting*

Recall that in point estimation, a given data point $x \in \mathcal{X}$ selects, within the entire model class $\mathcal{M}_\gamma$, a particular distribution with parameter $\hat{\theta}$:

$$x \to \hat{\theta} \rightsquigarrow f(\cdot|\hat{\theta}) \in \mathcal{M}_\gamma \ .$$

Here $\hat{\theta} : \mathcal{X} \to \Theta$ is some estimating function, for example, the MLE as given by (12.2). The $\rightsquigarrow$ sign is taken to mean "selects". The expression $f(y|\hat{\theta}(x))$, when viewed as a function of y for any fixed x, is a probability distribution that belongs to the family $\mathcal{M}_\gamma$ (i.e., is one of its elements). Evaluated at $y = x$, we denote $f(x|\hat{\theta}(x)) \equiv L_\gamma(x)$, viewed now as a function of the data x explicitly (recall that γ is the index for model class $\mathcal{M}_\gamma$). Note that $L_\gamma(x)$ is not a probability distribution; $\sum_x L_\gamma(x) \neq 1$ in general. The NML distribution $p^*(x)$, which is the normalized version of $L_\gamma(x)$, is derived as the solution of the minimax problem (12.5), over the yet-to-be determined distribution $q(x)$, with regret given as $\log(L_\gamma(x)/q(x))$. Now, instead of using this regret function, we use $\log(s(x)L_\gamma(x)/q(x))$ and consider a more general minimax problem

$$\inf_{q\in\mathcal{B}} \sup_{g\in\mathcal{B}} \mathrm{E}_g \left\{ \log \frac{s(y)\, L_\gamma(y)}{q(y)} \right\} \;, \tag{12.8}$$

where $s(x)$ is any positively-valued function of x.

Proposition 12.1. *The solution to the minimax problem (12.8) is given by $q(\cdot) = p(\cdot|\gamma)$ where*

$$p(x|\gamma) \equiv \frac{s(x)\, L_\gamma(x)}{\hat{C}_\gamma} = \frac{s(x)\, L_\gamma(x)}{\sum_{y\in\mathcal{X}} s(y)\, L_\gamma(y)} \;; \tag{12.9}$$

the minimaximizing bound is $\log \hat{C}_\gamma$ where

$$\hat{C}_\gamma = \sum_{y\in\mathcal{X}} s(y)\, L_\gamma(y) \;. \tag{12.10}$$

Proof. Our proof follows that of Rissenan (2001) with only slight modifications. First, noting the elementary relation

$$\inf_{q\in\mathcal{B}} \sup_{g\in\mathcal{B}} G(g,q) \geq \sup_{g\in\mathcal{B}} \inf_{q\in\mathcal{B}} G(g,q)$$

for any functional $G(g,p)$. Applying this to (12.8), the quantity $\{\cdot\}$ under minimaximizing,

$$\begin{aligned} \mathrm{E}_g \left\{ \log \frac{s(y)L_\gamma(y)}{q(y)} \right\} &= \mathrm{E}_g \left\{ \log \frac{g(y)}{q(y)} \right\} - \mathrm{E}_g \left\{ \log \frac{g(y)}{s(y)L_\gamma(y)} \right\} \\ &= \mathrm{E}_g \left\{ \log \frac{g(y)}{q(y)} \right\} - \mathrm{E}_g \left\{ \log \frac{g(y)}{p(x|\gamma)} \right\} + \log \hat{C}_\gamma = D(g||q) - D(g||p) + \log \hat{C}_\gamma \;, \end{aligned}$$

where $D(\cdot||\cdot)$ is the non-negative Kullback-Leibler divergence

$$D(g||q) = \mathrm{E}_g \left\{ \log \frac{g(y)}{q(y)} \right\} = \sum_{y\in\mathcal{X}} g(y) \log \frac{g(y)}{q(y)} \;.$$

Therefore

$$\begin{aligned} \inf_{q\in\mathcal{B}} \sup_{g\in\mathcal{B}} \mathrm{E}_g \left\{ \log \frac{s(y)\, L_\gamma(y)}{q(y)} \right\} &\geq \sup_{g\in\mathcal{B}} \inf_{q\in\mathcal{B}} \left(D(g||q) - D(g||p) + \log \hat{C}_\gamma \right) \\ = \sup_{g\in\mathcal{B}} \left(-D(g||p) + \log \hat{C}_\gamma \right) = \log \hat{C}_\gamma & \end{aligned}$$

where the infimum (over q) in the last-but-one step is achieved for $q = g$ and the supremum (over g) in the last step is achieved for $g = p$. Therefore, the solution to (12.8) is achieved when $q = p(\cdot|\gamma)$. □

Remark 12.1. The distribution $p^*(x)$, that is, non-informative NML (12.1), is known (Shtarkov, 1987) also to be the solution of the following slightly different minimax problem:

$$\inf_{q \in \mathcal{B}} \sup_{y \in \mathcal{X}} \log \frac{f(y|\hat{\theta}(y))}{q(y)} \ .$$

We can modify the above to yield a minimax problem (with given $s(y)$)

$$\inf_{q \in \mathcal{B}} \sup_{y \in \mathcal{X}} \log \frac{s(y) f(y|\hat{\theta}(y))}{q(y)} \ ,$$

and show that (12.9) is also its solution. The proof of this statement follows readily from the proof of Proposition 12.1.

We call (12.9) the informative NML distribution, which depends on an arbitrary positively-valued function $s(\cdot)$. Clearly, for all densities g,

$$\mathrm{E}_g \left\{ \log \frac{s(y)\, L_\gamma(y)}{p(y|\gamma)} \right\} = \mathrm{E}_g \log \hat{C}_\gamma = \log \hat{C}_\gamma$$

is constant. When $s(y) = const$, then

$$\hat{C}_\gamma \rightsquigarrow \mathrm{const} \sum_y L_\gamma(y) = \mathrm{const}\, C_\gamma$$

with

$$p(x|\gamma) \rightsquigarrow p^* = \frac{L_\gamma(x)}{\sum_y L_\gamma(y)} \ ,$$

both reducing to the (non-informative) NML solution derived by Rissanen (2001). The difference between $p(x|\gamma)$ and p^* is, essentially, the use of $s(x)L_\gamma(x)$ in place of $L_\gamma(x)$, that is, the maximal likelihood value $L_\gamma(x)$ at a data point x is weighted by a non-uniform, data-dependent factor $s(x)$. The data-dependency of the universal distribution (which in general still lies outside the manifold of the model class) qualifies it for the term "informative" NML (just as the parameter-dependency of a prior distribution in the Bayesian formulation qualifies it as an "informative prior").

Note that the function $s(x)$ in Proposition 12.1 can be any positively-valued function defined on $\mathcal{X}$. And the choice of $s(x)$ would affect the complexity measure $\hat{C}_\gamma$, which is also always positive.

12.3.2. *Prior over data and prior over model class*

The maximal likelihood values $L_\gamma(x)$ from model class γ over data x are a series of positive values; normalization over x gives the (non-informative) NML distribution $p^*(x)$ in Rissanen's (2001) analysis. Here, it is presupposed that the modeler has a prior belief π_γ about the plausibility of various model classes γ (with $\pi_\gamma > 0, \sum_\gamma \pi_\gamma = 1$), and a prior belief $\pi(x)$ about the credibility of the data x (with $\pi(x) > 0, \sum_x \pi(x) = 1$). These two types of prior beliefs may not be "compatible", in some sense yet to be specified more accurately below.

Let us take

$$s(x) = \frac{\pi(x)}{\sum_\gamma \pi_\gamma \, L_\gamma(x)} \ . \tag{12.11}$$

The meaning of such $s(x)$ will be elaborated later — it is related to, but not identical with, the so-called "luckiness prior" (Grünwald, 2007).

Note that (12.9) can be re-written as

$$p(x|\gamma) = \frac{p(\gamma|x)\,\pi(x)}{\sum_{y\in\mathcal{X}} p(\gamma|y)\,\pi(y)} \ , \tag{12.12}$$

where $p(\gamma|x)$ is defined by

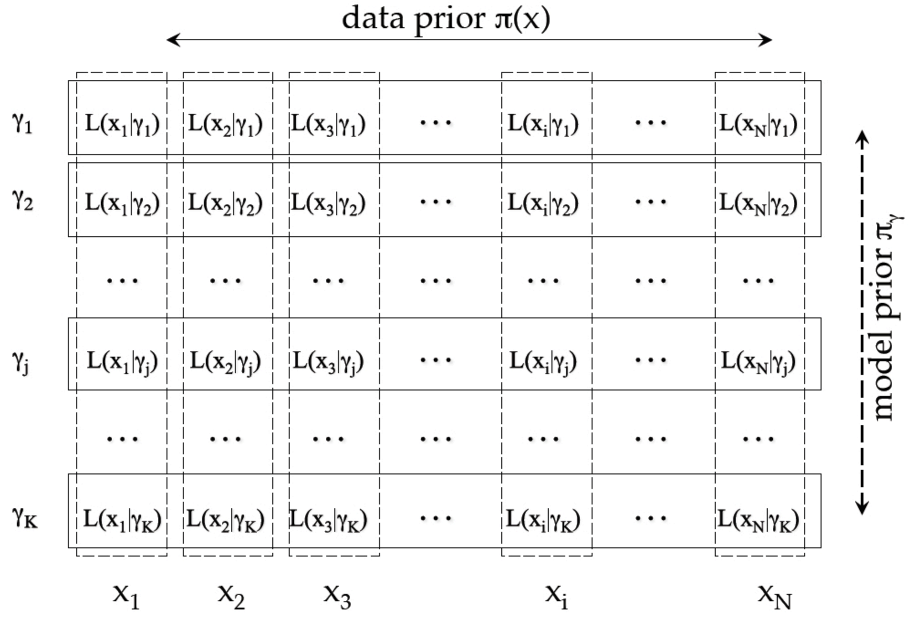
$$p(\gamma|x) \equiv \frac{\pi_\gamma \, L_\gamma(x)}{\sum_\gamma \pi_\gamma \, L_\gamma(x)} \ .$$

Since the denominator of the right-hand side of the above expression involves a summation over γ (and not x), we can then obtain

$$p(\gamma|x) = \frac{\pi_\gamma \, p(x|\gamma)}{\sum_\gamma \pi_\gamma \, p(x|\gamma)} \ . \tag{12.13}$$

The two equations (12.12) and (12.13) clearly have Bayesian interpretations: when $\pi(x)$ is taken to be the modeler's initial belief about the data *prior to* modeling, the solution to the minimax problem, now in the form of (12.12), can be viewed as the *a posterior* description of the data from the perspective of the model class $\mathcal{M}_\gamma$, with $p(\gamma|x)$ as likelihood functions *about the various model classes.* Likewise, when π_γ is taken to be the modeler's initial belief about the model class $\mathcal{M}_\gamma$ *prior to* an experiment, $p(\gamma|x)$ as given by (12.13) can be viewed as the *a posterior* belief about the various model classes after experimentally obtaining and fitting data x, whereas the informative NML solution $p(x|\gamma)$ serves as the likelihood functions *about the data.* So, informative NML has *two* interpretations, a) as the posterior of the data given model, in (12.12), or b) as the likelihood function of the model given data, in (12.13).

The above two interpretations correspond to two ways (see Figure 2) the maximum likelihood values $L_\gamma(x)$ can be normalized: a) across data x to become the probability distribution over data $p(x|\gamma)$; and b) across model class $\mathcal{M}_\gamma$ to become the probability distribution over model class $p(\gamma|x)$. This demonstrates a duality between data and model from the modeler's perspective.



	x_1	x_2	x_3	$\cdots$	x_i	$\cdots$	x_N
γ_1	$L(x_1\|\gamma_1)$	$L(x_2\|\gamma_1)$	$L(x_3\|\gamma_1)$	$\cdots$	$L(x_i\|\gamma_1)$	$\cdots$	$L(x_N\|\gamma_1)$
γ_2	$L(x_1\|\gamma_2)$	$L(x_2\|\gamma_2)$	$L(x_3\|\gamma_2)$	$\cdots$	$L(x_i\|\gamma_2)$	$\cdots$	$L(x_N\|\gamma_2)$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
γ_j	$L(x_1\|\gamma_j)$	$L(x_2\|\gamma_j)$	$L(x_3\|\gamma_j)$	$\cdots$	$L(x_i\|\gamma_j)$	$\cdots$	$L(x_N\|\gamma_j)$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
γ_K	$L(x_1\|\gamma_K)$	$L(x_2\|\gamma_K)$	$L(x_3\|\gamma_K)$	$\cdots$	$L(x_i\|\gamma_K)$	$\cdots$	$L(x_N\|\gamma_K)$

Fig. 12.2. Figure 2. Illustration of data prior, model prior, and the ML values $L(x|\gamma) \equiv L_\gamma(x)$ for data points $x_1, x_2, \cdots, x_N$ across various model classes $M_{\gamma_1}, M_{\gamma_2}, \cdots, M_{\gamma_k}$. When model prior π_γ is given, ML values (viewed as columns) are used as the likelihood function of a particular data point for different model classes, in order to derive posterior estimates of model classes. When data prior $\pi(x)$ is given, ML values (viewed as rows) are used as the likelihood function of a particular model class for different data points, in order to derive posterior estimates of data (informative NML solution). Data prior and model prior can be made to be compatible (see Proposition 12.2).

12.3.3. *Model complexity measure*

Let us now address the model complexity measure associated with the informative NML approach. Substituting (12.11) into the expression for the complexity measure $\hat{C}_\gamma$, we have

$$\pi_\gamma \hat{C}_\gamma = \sum_x p(\gamma|x)\pi(x) \ . \tag{12.14}$$

Explicitly written out

$$\hat{C}_\gamma = \sum_x \frac{\pi(x) L_\gamma(x)}{\sum_\gamma \pi_\gamma L_\gamma(x)} \ .$$

From (12.14), we obtain

$$\sum_\gamma \pi_\gamma \hat{C}_\gamma = 1 \ .$$

This indicates that the new model complexity measure proposed here, $\hat{C}_\gamma$, is normalized after weighted by π_γ. The fact that $\pi_\gamma > 0$ implies that

$$\hat{C}_\gamma < \infty \ .$$

This solves a long-standing problem, the so-called "infinity problem" (Grünwald, Myung, & Pitt, 2005) associated with C_γ in the non-informative NML.

Recall that the non-informative NML follows the "two-part code" idea of MDL, that is, one part that codes the description of the hypothesis space (the functional form of the model class), the other part that codes the description of the data as encoded with respect to these hypotheses (the maximum likelihood value of the MLE). As such, the original minimax problem (12.5) has a clear interpretation of the "ideal code" from algorithmic coding perspective, with C_γ as the complexity measure of the model class. Here the normalization factor $\hat{C}_\gamma$ associated with the informative NML solution (12.9) has an analogous interpretation. The only difference is that the complexity measure now is dependent on the prior belief of the data $\pi(x)$ and the prior belief of the model classes π_γ, in addition to its dependency on the best-fits provided by each model class for all potential data.

The $s(x)$ factor introduced in the minimax problem given in (12.8) is related (but with important differences, see next subsection) to the "luckiness prior" introduced by Grünwald (2007). In the current setting, with $s(x)$ taking the specific form of (12.11), we have the following interpretation: for the occurrence of any data point x, the denominator $\sum_\gamma \pi_\gamma L_\gamma(x)$ gives expected occurrence of x from the modeler's prior knowledge about all models he/she builds, whereas the numerator $s(x)$ gives the modeler's prior knowledge about the data occurrence from a known data-acquisition procedure. Since the knowledge of the modeler/experimenter about model building and experimentation may come from different sources, the "luckiness" of acquiring data x as resulting from an experiment thus can be operationalized as the ratio of these two probabilities associated with different types of uncertainty about data.

Note that if and only if

$$\pi(x) = \text{const} \sum_\gamma \pi_\gamma L_\gamma(x), \tag{12.15}$$

the luckiness factor $s(x) = \text{const}$; this is the case when the informative NML solution (12.9) reduces to the non-informative NML solution (12.1), both in this formulation, and in the approach reviewed by Grünwald (2007). We say that the modeler's prior belief over data and prior belief over model class are mutually *compatible* when (12.15) is satisfied (over all possible data values x). It is easy to see that luckiness is a constant (i.e., same across all data points) if and only if model prior and data prior are compatible.

Proposition 12.2. *The following three statements are equivalent.*

(a) Luckiness $s(x)$ is constant;
(b) Model prior π_γ and $\pi(x)$ are compatible;
(c) Informative NML is identical with non-informative NML.

12.3.4. *Data prior versus "luckiness prior"*

The data-dependent factor $s(x)$ introduced here, while in the same spirit of the so-called "luckiness prior" as in Grünward (2007, pp. 308-312), carries subtle differences. In Grünward's case, the corresponding minimax problem is

$$\inf_{q\in\mathcal{B}} \sup_{g\in\mathcal{B}} \mathrm{E}_g \left\{ \log p(y|\hat{\theta}(y)) - \log q(y) - a(\hat{\theta}(y)) \right\}$$

and the extra factor $a(\hat{\theta}(y))$ is a function of the maximum likelihood estimator $\hat{\theta}(\cdot)$. In the present case, the minimax problem is

$$\inf_{q\in\mathcal{B}} \sup_{g\in\mathcal{B}} \mathrm{E}_g \left\{ \log p(y|\hat{\theta}(y)) - \log q(y) + \log s(y) \right\},$$

with $s(y)$ a function defined on the sample space directly (and not through "pull-back"). However, both approaches to informative NML afford

Bayesian interpretations. The approached described in Grünwald (2007) will lead to the *luckiness-tilted Jeffreys' prior* (p.313, ibid.),

$$\pi_{J,a}(\theta) = \frac{\sqrt{\det \mathrm{I}(\theta)}\, e^{-a(\theta)}}{\int_{\Theta} \sqrt{\det \mathrm{I}(\theta)}\, e^{-a(\theta)} d\theta} \ ,$$

which has the information geometric interpretation as an invariant volume form under a generalized conjugate connection on the manifold of probability density functions (Takeuchi & Amari, 2005; Zhang & Hasto, 2006; Zhang, 2007). The approach adopted in this paper gives rise to a dual interpretation between model and data. Just as the maximum likelihood principle can be used to select the parameter (among all "competing" parameters) of a certain model class, the NML principle has been used to select a model class out of a set of competing models. Just as there is a Bayesian counterpart to the ML principle for parameter selection, what is proposed here is the Bayesian counterpart to NML, i.e., the use of maximum $p(\gamma|x)$ value (with fixed x, i.e., the given data) for model selection (among all possible model classes). The same, old debate and argument surrounding ML and Bayes can be brought back here — we are back to square one. Except that we are now operating at a higher level of explanatory hierarchy, namely, at the level of model classes (whereby each class is represented by a universal distribution through its maximum likelihood values after proper normalization); yet the duality between model and data still manifests itself.

12.4. General Discussions

To summarize this chapter, from the maximum likelihood function $f(x|\hat{\theta}(x)) \equiv L_\gamma(x)$ (where $\hat{\theta}$ is the MLE for the model class $\mathcal{M}_\gamma$), one can *either* construct the (non-informative) NML as a universal distribution of the model class γ through normalizing with respect to x, as Rissanen (2001) did, *or* derive the posterior distribution for model selection (12.13) through normalizing with respect to γ, as is done here. This has significant implications for model selection. In the former case, model selection is through the comparison of NML values for various model classes. Because the NML solution (12.1) is a probability distribution (in fact, universal distribution representing the particular model class) with total mass 1, then necessarily no single model can dominate (i.e., be the preferred choice) across all data! In other words, for any data x where model class γ_1 is preferred to model class γ_2, there exists some other data x' where model class γ_2 outperforms model class γ_1. Here, in our situation, we use an (informative) universal distribution which is interpreted as the likelihood function, with respect to a prior belief about all model classes — model selection is through computing the Bayes factor which combines the two data scenario. The dominance or superiority of one model class over another in accounting for all data is permitted under the current method.

12.5. Conclusion

Normalized maximal likelihood is a probability distribution over the sample space associated with a parametric model class. At each sample point, the value of an NML distribution is obtained by taking the likelihood value of the maximum likelihood estimator (ML value) corresponding to that sample point (data), and then normalizing across the sample space to give rise to the unit probability measure. Here, the minimax problem that leads to the above (non-informative) NML as its solution is revisited by our introducing an arbitrary weighting function over sample space. The solution then becomes "informative NML", which involves both a prior distribution over the sample space ("data prior") and a prior distribution over model class ("model prior"), with obvious Bayesian interpretations of the ML values. This approach avoids the so-called "infinity problem" of the non-informative NML, namely the unboundedness of the logarithm of the normalization factor (which serves as an index for model complexity), while at the same time providing a notion of consistency between the modeler's prior beliefs about models and data.

Acknowledgments

This chapter is based on preliminary results grown out of a discussion between the author and Jay Myung during 2005, which was reported to the 38th Annual Meeting of the Society for Mathematical Psychology held at the University of Memphis, TN (Zhang & Myung, 2005). The work has since been greatly expanded – the author benefited from subsequent discussions on this topic with Richard Shiffrin, who encouraged the development of the notion of a "data prior", and with Woojae Kim and Jay Myung, who caught an error in an earlier, circulating draft. The views of this chapter (and any mistakes therein) represent solely those of the author, and may be different from those commentators.

References

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

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description length principle in coding and modeling, *IEEE Transactions on Information Theory, 44*, 2743–2760.

Grünwald, P., Myung, I. J., & Pitt, M. A. (2005). *Advances in Minimum Description Length: Theory and Applications.* Cambridge, MA: MIT Press.

Hansen, M. H., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association, 96*, 746–774.

Grünwald, P. (2007). *The Minimum Description Length Principle.* Cambridge, MA: MIT Press.

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

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

Karabatsos, G., & Walker, S. G. (2009). Rejoinder on the normalized maximum likelihood and Bayesian decision theory: Reply to Grünwald and Navarro (2009). *Journal of Mathematical Psychology, 53*, 52.

Matsuzoe, H., Takeuchi, J., & Amari, S. (2006). Equiaffine structures on statistical manifolds and Bayesian statistics. *Differential Geometry and Its Applications, 24*, 567–578.

Myung, I. J. (2000) The importance of complexity in model selection. *Journal of Mathematical Psychology, 44*, 190–204.

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Guest editors' introduction: Special issue on model selection. *Journal of Mathematical Psychology, 44*, 1–2.

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

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

Rissanen, J. (1978). Modeling by the shortest data description. *Automata, 14*, 465–471.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics, 11*, 416–431.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics, 14*, 1080–1100.

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

Rissanen, J. (2000). MDL denoising. *IEEE Transactions on Information Theory, 46*, 2537–2543.

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

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

Takeuchi, J., & Amari, S. (2005). α-Parallel prior and its properties. *IEEE Transaction on Information Theory, 51*, 1011–1023.

Zhang, J. (2007). A note on curvature of -connections on a statistical manifold. *Annals of Institute of Statistical Mathematics, 59*, 161–170.

Zhang, J., & Hasto, P. (2006). Statistical manifold as an affine space: A functional equation approach. *Journal of Mathematical Psychology, 50*, 60–65.

Zhang, J., & Myung, J. (2005). Informative normalized maximal likelihood and model complexity. Talk presented to the 38th Annual Meeting of the Society for Mathematical Psychology, University of Memphis, TN.