Schwarz, Wallace, and Rissanen: Intertwining Themes in Theories of Model Selection

Aaron D. Lanterman
Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1308 W. Main, Urbana, IL 61801, USA. E-mail: lanterma@ifp.uiuc.edu.

May 29, 2000

Abstract

Investigators interested in model order estimation have tended to divide themselves into widely separated camps; this survey of the contributions of Schwarz, Wallace, Rissanen, and their coworkers attempts to build bridges between the various viewpoints, illuminating connections which may have previously gone unnoticed and clarifying misconceptions which seem to have propagated in the applied literature. Our tour begins with Schwarz's approximation of Bayesian integrals via Laplace's method. We then introduce the concepts underlying Rissanen's minimum description length principle via a Bayesian scenario with a known prior; this provides the groundwork for understanding his more complex non-Bayesian MDL which employs a "universal" encoding of the integers. Rissanen's method of parameter truncation is contrasted with that employed in various versions of Wallace's minimum message length criteria. Rissanen's more recent notion of stochastic complexity is outlined in terms of Bernardo's information-theoretic derivation of the Jeffreys prior.

Key Words: Bayesian model selection, minimum description length, minimum message length, stochastic complexity, parameter truncation

1 Introduction

Since its introduction by Fisher, the method of maximum-likelihood has proven an effective method of vector parameter estimation when the dimension of the parameter space is fixed. When the dimension of the parameter space itself needs to be estimated, maximum-likelihood techniques tend to be "greedy," consistently picking the models of greatest complexity to yield overly tight fits to the data.

The challenges of model order estimation were addressed by Akaike (1973). Five years later, alternative approaches were proposed by Schwarz (1978) and Rissanen (1978). Schwarz took a Bayesian approach of integrating out nuisance parameters via Laplace's method. Adopting a quite different tactic rooted in coding theory and theoretical computer science (Li & Vitányi, 1997), Rissanen proposed the minimum description length principle, which seeks to minimize the number of bits needed to describe the data over the available models. Although his 1978 paper may be the most ubiquitously cited reference for the origins of MDL, the seeds of estimation-via-coding were planted a decade earlier in the computer science literature by Wallace & Boulton (1968). Wallace later refined his ideas and termed them minimum message length (Wallace & Freeman, 1987a). A trio of technical reports from Monash University (Oliver & Hand, 1994; Oliver & Baxter, 1994; Baxter & Oliver, 1994) and the closely related dissertation by Baxter (1996b), although written with an emphasis on Wallace's ideas, discuss some of the similarities and differences between the works of Schwarz, Wallace, and Rissanen.

Most Bayesian inference procedures are based on minimizing an expected cost function. Wallace's MML criteria (Wallace & Freeman, 1987; 1992) crafts Bayesian inference procedures without having to specify an explicit cost function (such as squared error or probabilities of detection and false alarm). Although prior distributions also arise in Rissanen's work, they merely function as technical tools; as we shall see, they are transforms of lengths of messages encoded via codebooks that are explicitly devised *without* incorporating prior knowledge about the parameters. Rissanen goes to great lengths to avoid using classical Bayesian priors in modeling what he considers "subjective" knowledge about parameters and staunchly resists Bayesian interpretations of his criteria. Preferring to make as few assumptions about the parameters as possible, his minimum description length approach, developed in the early 80's, represents parameters using a "universal" coding scheme for the integers (Rissanen 1983; 1986; 1987a; 1987b). Rissanen's most recent work takes a different tactic, exploiting what he calls a normalized maximum-likelihood code which amounts to employing a Bernardo-Jeffreys prior that is restricted to subsets of the parameter space, with the subset index encoded using his universal integer code (Rissanen 1995a; 1995b; 1996). Grünwald (2000) offers a highly readable account of Rissanen's ideas, aimed at non-specialists. In asymptotic settings, Hansen & Yu (1998) bring MDL inference to life with a lucid presentation of numerous applications. In spite of Rissanen's philosophical objections to using prior distributions to represent *a priori* knowledge about parameters, there is nothing to prevent devoted Bayesians from assimilating Rissanen's MDL techniques into their own agenda. We investigate this path in Section 4.3.

Traditional maximum-likelihood and Bayesian techniques have a well-developed theory; many of the traditional debates have been somewhat resolved, and numerous books are available to the practitioner. Unfortunately, model order estimation remains a subject of tremendous controversy; there is little agreement on what the "best" approach is, and indeed little agreement on if there is, in fact, such a thing as a "best" approach. Each of the pioneering investigators in the field seems to have staked out his own individual niche; attempts to synthesize the different mindsets into a coherent whole are rare. We are aware of only two books on model order selection. The volume by Linhart & Zucchini (1986) consists mainly of Akaike-type methods, and Rissanen's book (1989) focuses on his own approach, which he has made profound extensions to since its publication; we recommend supplementing it with a more recent set of lecture notes Rissanen (1999b) prepared for a course at the Tampere Univ. of Technology. Although not technically a "book," Grünwald's (1998) 300+ page thesis may be the clearest, most coherent reference on coding-theoretic model selection that we are aware of; in particular, Chapter 7 compares MDL and MML. Special issues of The Computer Journal[1] (Vol. 42, No. 4, 1999), Statistics and Computing (Vol. 10, No. 1, 2000) and The Journal of Mathematical Psychology (2000) are devoted to model selection topics including MDL, MML, Akaike's techniques, and cross-validation approaches.

Considering the challenging nature of the original works and the preponderance of misinterpretations and misunderstandings in the applied literature, we hope this paper will be a welcome addition to the surveys given in the technical reports and dissertations cited above, perhaps shining light on some dark corners and paving the way for further understanding and developments.

Although there is a rich body of work relating to Akaike's methods (Akaike, 1974; Stone, 1977; Taylor, 1987; Sakamoto et al., 1986; Sakamoto, 1992; Hurvich et al., 1998; Bozdogan, 2000; Cavanaugh & Shumway, 1998), especially in time series analysis (Akaike, 1979; Atkinson, 1978; Shibata, 1976), we will not consider his work further here, and instead focus on the work of Schwarz, Wallace, and Rissanen. All three researchers obtain similar asymptotic results. Akaike's results, which are based on bias-correction terms for estimating cross-entropies, are quite different.

[1] Articles may be downloaded from the web at www3.oup.co.uk/computer_journal/hdb/Volume_42/Issue_04.

2 Organization and Contributions

Over the course of this review, we occasionally attempt to point out some connections between results in the literature that have either not been well understood or which seem to have gone unnoticed. The literature contains myriad potential pitfalls which must be illuminated in order to fully appreciate the work of the cited authors, and more importantly, to avoid potential misunderstandings.

Taylor series expansions of logposteriors in the spirit of Clarke & Barron (1990; 1994) and Balasubramanian (1997) will appear throughout; they form the glue that unifies the various approaches. Section 3 presents Schwarz's approach to model order estimation (Schwarz, 1978), which involves computing Bayesian integrals via Laplace's method, an application of Taylor series approximation. We point out that different formulas result depending on whether or not the full logposterior or just the loglikelihood are expanded to second-order terms. We present Schwarz's classic asymptotic approximation and compare it to a less commonly known approximation suggested by Draper (1995). We also tackle the frequent misinterpretation of the omnipresent $-(d/2)\log N$ term as a logprior on model order.

The basic ideas underlying minimum description length inference (Rissanen 1978; 1987a) are illustrated in Section 4. Section 4.3 applies Rissanen's method of parameter truncation to a Bayesian situation where a prior is available. The resulting formula differs from the Schwarz result by a term which is linear in the number of parameters. This presentation is helpful as a prelude to understanding Rissanen's more complex non-Bayesian MDL principle, which we consider in Section 7.

Section 5 reviews the work of Wallace and coworkers in minimum message length inference (Wallace & Freeman 1987a; 1992), a variant of MDL inference which emphasizes Bayesian formulations and employs a different approach to parameter truncation than that taken by Rissanen. We catalogue and clarify the different versions of MML that have appeared in the literature, and relate these versions of MML inference to the different versions of Schwarz's method presented in Section 3. We point out that Takeuchi's (1997) flavor of MDL is actually closer to Wallace's MML than Rissanen's MDL, although it employs Rissanen-style rectangular quantization regions.

Section 6 reviews the results of the preceding sections and interprets them as penalized loglikelihoods.

In many applications, no obvious prior distribution is available. Hence, Section 7 explores the non-Bayesian version of Rissanen's MDL (1983), which exploits a so-called "universal" prior on the integers, and Section 8 presents his more recent concept of *stochastic complexity* (Rissanen, 1996).

3 Schwarz's Application of Laplace's Method

We will begin with Schwarz's approach to model order estimation, as his approximate Bayesian approach may be easier to understand than the substantially more complicated coding-theoretic arguments used by Rissanen and Wallace. Consider the multihypothesis testing problem of determining which model $m \in \mathcal{M}$ generated a given data set x, where $\mathcal{M}$ indexes the models and the models have *a priori* probability $p_m(m)$. For each model, we need the likelihood and prior densities denoted by $p_l(x|\theta, m)$ and $p_p(\theta|m)$. The parameter space of θ may differ for each m, $\theta \in \mathbf{\Theta}_m$, where $\mathbf{\Theta}_m$ is the parameter space associated with model m.

The Bayesian procedure proceeds by integrating out the nuisance variable θ to find the probability of x under the model m

$$p(x, m) = p_m(m) \int_{\mathbf{\Theta}_m} p_l(x|\theta, m) p_p(\theta|m) d\theta. \tag{1}$$

and picking the class m which maximizes $p(m|x) \propto p(x,m)$.

In a few specialized cases, such as auditory-nerve discharge rate estimation (Mark & Miller, 1992) and multinomial graphical models (Heckerman, 1995), the Bayesian integral (1) can be performed analytically. In many cases, however, the integral is formidable, and the approximation technique employed in Schwarz (1978) becomes attractive. Direct computation of (1), if feasible, will yield better results than the approximations presented below; hence, we should only turn to the approximations if (1) is too complex.

Let $L(x|\theta,m) = \ln p_l(x|\theta,m)$, $P(\theta|m) = \ln p_p(\theta|m)$, $H(x,\theta|m) = L(x|\theta,m) + P(\theta|m)$. We assume that the MAP estimate $\hat{\theta}(x) = \arg\max_\theta H(x,\theta|m)$ is unique. Suppose the model m has dimension (number of free parameters) $d = d(m)$.

Bayesian-Laplace approximation, direct: If the posterior density is highly peaked, we can approximate the integrand via Laplace's method (Polya & Szego, 1972, p. 96). Laplace's approach employs a Taylor series expansion of H around the MAP estimate $\hat{\theta}$, taken to second order

$$H(x,\theta|m) \approx H(x,\hat{\theta}|m) - \frac{1}{2}(\theta-\hat{\theta})^T I_H(x:\hat{\theta}|m)(\theta-\hat{\theta}) \tag{2}$$

where

$$I_H(x:\hat{\theta}|m) = \left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}H(x,\theta|m)\right]\Bigg|_{\theta=\hat{\theta}}. \tag{3}$$

Throughout this paper, we will assume that the necessary partial derivatives exist and that I_H, as well as other related matrices of second partial derivatives, are positive definite. Note that the first derivative term in (2) vanishes since we are evaluating at a maximizer.

Substituting this approximation for $H(x,\theta|m)$ in the Bayesian integral (1) yields

$$\begin{aligned} p(x,m) &= p_m(m)\int_\theta \exp[H(x,\theta|m)]d\theta \\ &\approx p_m(m)\exp[H(x,\hat{\theta}|m)]\times \\ &\quad \int \exp[-\frac{1}{2}(\theta-\hat{\theta})^T I_H(x:\hat{\theta}|m)(\theta-\hat{\theta})]dx \\ &\overset{(a)}{=} p_m(m)\exp[H(x,\hat{\theta}|m)]\frac{(2\pi)^{d/2}}{\sqrt{\det I_H(x:\hat{\theta}|m)}}. \end{aligned} \tag{4}$$

Equality (a) follows readily from recognizing the integrand as the quadratic form of a Gaussian density.

In terms of logarithms, we have

$$\begin{aligned} \ln p(x,m) &\approx \ln p_m(m) + H(x,\hat{\theta}|m) \\ &\quad +\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\det I_H(x:\hat{\theta}|m). \end{aligned} \tag{5}$$

Choosing the m which maximizes (5) can easily been seen as equivalent to choosing the m which maximizes the *a posteriori* probability. The accuracy of approximations like (5) is considered by Clarke & Barron (1994); Balasubramanian (1997) undertakes an even sharper analysis. Some rather technical regularity conditions (see, for instance, Conditions 1-3 of Clarke & Barron (1994)) are needed for the approximation to be useful. These conditions hold for many problems of interest, and in the sequel we will assume that they hold without further comment. However, as one of the referees pointed out, users of model selection criteria often have a tendency to jump ahead and apply the sorts of "plug-in" formulas given throughout this paper without checking any of

the regularity conditions, so it behooves a potential user to give them some thought. See the tree classifier of Section 7.2 of Rissanen (1989) for an example where (5) is not helpful.

This equation is analogous to O'Hagan's message length formula (37), which will be presented in Section 5.1. Lanterman (1998a; 1998b; 2000) used this formula (5) for estimating the dimension of a model used to represent the thermodynamic state of vehicles for infrared target recognition.

Bayesian-Laplace Approximation, Oliver & Baxter version: Suppose that the prior $p_p(\theta|m)$ is sufficiently flat around the MAP estimator $\hat{\theta}$ that, for purposes of computing the integral, $p_p(\theta|m) \approx p_p(\hat{\theta}|m)$ may be extracted from the integrand as a constant (Oliver & Baxter, 1994). Applying Laplace's method to the remaining likelihood yields

$$\begin{aligned} p(x|m) &\approx p_p(\hat{\theta}) \int_\theta \exp[L(x|\theta,m)]d\theta \\ &\approx p_p(\hat{\theta}|m) \exp[L(x,\hat{\theta}|m)] \frac{(2\pi)^{d/2}}{\sqrt{\det I_L(x:\hat{\theta}|m)}} \\ &= \exp[H(x,\hat{\theta})] \frac{(2\pi)^{d/2}}{\sqrt{\det I_L(x:\hat{\theta}|m)}}, \end{aligned} \tag{6}$$

where

$$I_L(x:\hat{\theta}|m) = \left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} L(x|\theta,m) \right] \Big|_{\theta=\hat{\theta}} \tag{7}$$

is the *empirical* or *observed* Fisher information matrix. In terms of logarithms, we have

$$\begin{aligned} \ln p(x,m) &= \ln p_m(m) + H(x,\hat{\theta}|m) \\ &\quad + \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\det I_L(x:\hat{\theta}|m). \end{aligned} \tag{8}$$

This equation is analogous to Wallace & Freeman's message length formula (31), which we will present in Section 5.1.

Large sample-size asymptotics: Consider the Oliver & Baxter approximation presented above. Suppose we have N i.i.d. samples $x = x_1, x_2, \ldots, x_N$, so that the loglikelihood is

$$L(x|\theta,m) = \sum_{n=1}^{N} L(x_n|\theta,m) \tag{9}$$

and the empirical Fisher likelihood information is

$$I_L(x:\hat{\theta}|m) = \sum_{n=1}^{N} I_L(x_n:\hat{\theta}(x)|m). \tag{10}$$

Let $\hat{\theta}_{lim} = \lim_{N\to\infty} \hat{\theta}(x)$ be the limiting maximum-likelihood estimate under the class m, which is well-defined by the assumed regularity conditions. Let $p(x|\hat{\theta}_{lim},m)$ be the data-generating model assuming class m and the limiting ML estimate; recall x denotes the full, observed sample. If model m is the correct data-generating model, then $\hat{\theta}_{lim}$ can be considered the "true" parameter value by the asymptotic consistency of ML estimation.

Using the law of large numbers and uniform convergence, one can show that

$$\frac{E_{p(x|\hat{\theta}_{lim},m)}[I_L(x_1:\hat{\theta}(x)|m)]}{(1/N)I_L(x:\hat{\theta}(x)|m)} \to 1 \tag{11}$$

with probably one.

Hence, the rightmost term in (8) may be asymptotically approximated by

$$\begin{aligned} & -\frac{1}{2}\ln\det I_L(x:\hat{\theta}|m) \\ \approx\ & -\frac{1}{2}\ln\det\{NE_{p(x|\hat{\theta}_{lim},m)}[I_L(x_1:\hat{\theta}(x)|m)]\} \\ =\ & -\frac{1}{2}\ln\det NI_d - \frac{1}{2}\ln\det E_{p(x|\hat{\theta}_{lim},m)}[I_L(x_1:\hat{\theta}(x)|m)] \\ =\ & -\frac{d}{2}\ln N - \frac{1}{2}\ln\det E_{p(x|\hat{\theta}_{lim},m)}[I_L(x_1:\hat{\theta}(x)|m)], \end{aligned} \tag{12}$$

where I_d is the $d \times d$ identity matrix and the equalities are due to basic properties of determinants.

The log-probability of the model and the data can then be asymptotically approximated by

$$\begin{aligned} \ln p(x,m) \approx\ & \ln p_m(m) + \sum_{n=1}^{N} L(x_n|\hat{\theta}(x),m) + P(\hat{\theta}|m) \\ & + \frac{d}{2}\ln 2\pi - \frac{d}{2}\ln N \\ & - \frac{1}{2}\ln\det E_{p(x|\hat{\theta}_{lim},m)}[I_L(x_1:\hat{\theta}(x)|m)]. \end{aligned} \tag{13}$$

As $N \to \infty$, the terms which are functions of N begin to dominate the others, yielding the classic approximation

$$\ln p(x,m) \approx L(x|\hat{\theta},m) - \frac{d}{2}\ln N. \tag{14}$$

We choose the model m which maximizes (14). Often called the BIC (Bayesian Information Criterion) or occasionally SIC (Schwarz Information Criterion), Schwarz's method is most frequently stated in this asymptotic manner. Both the BIC and the AIC (Akaike Information Criterion) have been compared in the estimation of finite mixtures (Liang et al., 1992) (with application to medical image processing) and direction finding via narrowband sensor arrays (Wax & Kailath, 1985). BIC has also found use in estimating appropriate neighborhood sizes in texture modeling via Markov random fields (Smith & Miller, 1990). Note that certain conditions, in particular, the consistency of maximum-likelihood estimators with increasing N, need to hold for (14) to be meaningful. As one of the referees noted, investigators sometimes mistakenly employ criteria like (14) in situations where the steps leading up to it are not valid.

In the literature, the $-\frac{d}{2}\ln N$ term is sometimes interpreted as a logprior on d. This interpretation is somewhat misleading; although the resulting "prior" is summable

$$\sum_{d=0}^{\infty}\exp\left(-\frac{d}{2}\ln N\right) = \sum_{d=0}^{\infty} N^{-d/2} = \frac{1}{1-(1/\sqrt{N})}, \tag{15}$$

the result came from assuming some other prior $p_m(m)$ on the model class. Note that the result (14) depends only on d, not $p_m(m)$.

A less frequently used approximation, proposed by Draper (1995), retains the 2π term:

$$\ln p(x,m) \approx L(x|\hat{\theta},m) - \frac{d}{2}\ln\frac{N}{2\pi}. \tag{16}$$

As Draper suggests, the best way to know when N is sufficiently large to justify the use of these asymptotic approximations is to compute the nonasymptotic and asymptotic versions and compare

them (see Draper, 1995, p. 57). Contrary to intuition, (16) may not be superior to (14) in general. In the discussion section of Draper (1995), Raftery (pp. 78-79) and Kass and Wasserman (pp. 84-85) give examples in which Schwarz's original approximation (14) is better; in Draper's response (p. 91), he gives examples where retaining the 2π term (16) is better. This question is probably best answered for each individual problem. Such an investigation for problems in epidemiology was conducted by Hook *et al.* (1995).

Related work: Clarifications and extensions of Schwarz's technique are investigated by Poskitt (1987). O'Hagan (1995) considers the use of such Schwarz-Laplace expansions for approximating Bayes factors for hypothesis testing. Stone (1979) and Zhang (1993) compare the work of Akaike and Schwarz; Rahman & King (1999) propose a Joint Information Criterion (JIC) which represents a compromise between the AIC and the BIC. Laplace's method has also found application in the study of the information-theoretic asymptotics of Bayes methods (Clarke & Barron, 1990) and in computing probabilities of detection in low-noise automatic target recognition scenarios (Grenander et al., 1998). There are variations of Laplace's method which are more accurate for large numbers of parameters (Shun & McCullagh, 1995) or small sample sizes (McQuarrie, 1999).

A remark on random sampling: In some applications, the regularity conditions needed for Laplace's method to make sense do not hold, and thus we must turn to other techniques. Many authors, including Green and Richardson (1995; 1997), Raftery (1996), and Carlin & Chib (1995), have proposed using Markov chain Monte Carlo (MCMC) methods to compute the Bayesian integral (1). Well-crafted MCMC algorithms can efficiently traverse the space with wide enough breadth to find good approximations to (1).

4 Minimum Description Length and Related Principles

To introduce the minimum description length principle, suppose the data y and parameters x live in *discrete* spaces. We will later extend the discussion to continuous spaces. The overall idea is to choose a representation of the data which permits us to express them with the shortest possible message via a postulated set of models. Traditionally, the "message" or "description" length is measured in bits, which arise from using base-2 logarithms. Other units based on other bases can be used; however, to maintain consistency with the natural logarithms used above, we will measure information using "nats." (One can easily convert from "nats" to "bits" by dividing by $\ln 2$.) See Sections 2.2 and 2.3 of Rissanen (1989) or Sections 2.1 and 2.2 of Hansen & Yu (1998) for a review of the underlying coding-theoretic concepts.

In theory, the shortest computer program which generates the data y provides the most efficient description of the data. A profound theorem, proved independently by Kolmogorov, Chaitin, and Solomonoff (Li & Vitányi, 1997), asserts that there is no algorithm which can find the shortest computer program to reproduce a particular set of data, or even find the length of such a shortest program. (Cover & Thomas (1991 Chapter 7) offer a refreshing introduction to this important notion and related ideas; Vitányi & Li (1996; 2000) and Wallace & Dowe (1999) offer deep discussions on the relation between MDL/MML and Kolmogorov complexity.) Hence, any attempt to find the length of the absolutely shortest possible program would be futile. To avoid such a morass, we consider minimizing the description length over a set of candidate models $\mathcal{M}$. We do not even need to suppose that the data are the result of a realization of one of the models; as we develop more models, we are free to add them to $\mathcal{M}$.

One can construct practical MDL schemes (Chapter 3, Rissanen 1989) by coding models explicitly (two-part coding), by developing codes based on model averaging (stochastic complexity), or by so-called "predictive coding." Here we will focus on the simplest, the two-part code scheme. In Section 8, we will explore Rissanen's more advanced stochastic complexity framework. Although this paper will not explore the predictive version of MDL in depth, some discussion is provided in Section 9.

4.1 Minimizing Worst-Case vs. Expected Description Lengths

In devising coding schemes, there is no code which gives a short message length to *all* sequences of data. Given a model m and an instance of data x, the shortest codelength for x among all the available distributions in m is given by

$$-\log p(x|\hat{\theta}(x)) \stackrel{\text{df}}{=} -L(x|\hat{\theta}(x)), \tag{17}$$

where $\hat{\theta}(x)$ is the ML estimator for the class m. Unfortunately, no code exists which attains the codelength given by (17) for *every* instance of x. Therefore, some additional choices have to be made, and these choices unleash one of the main sources of contention between Rissanen's and Wallace's approach to MDL/MML style interference.

Wallace finds no trouble with minimizing expected code lengths, where expectations are taken over some assumed "true" distribution or a "subjective" Bayesian distribution. By contrast, Rissanen avoids assuming that the data is actually generated by any of the models under consideration, and seeks a code with lengths which are close to (17) regardless of what the data or data generating mechanisms are. Hence, he takes a worst-case approach which seeks the code with lengths len which minimize the *maximum excess* length

$$\max_x \{len(x) - [-\log p(x|\hat{\theta}(x))]\} \tag{18}$$

or the worst-case *expected* code length,

$$\max_{g \in \mathcal{G}} E_g\{len(x) - [-\log p(x|\hat{\theta}(x))]\} \tag{19}$$

where the maximum is taken over the set of all possible distributions, not just those in the classes under consideration. Solving these problems leads to Rissanen's *stochastic complexity*. Section 8 we will discuss these problems in depth. One way of approximating such minimax codes is to use a multi-part encoding scheme as described in the next section. In these schemes, parameter estimates are explicitly encoded. Readers interested in exploring the original works should be warned that the worst-case minimax interpretation of such multi-part codes, well-described by Rissanen in his more recent work (1996), was not clear in his early papers (1978; 1983).

Wallace's MML, which focuses on parameter estimation in addition to model selection, also explicitly encodes parameter estimates. In Rissanen's view, the encoding of parameter estimates is just a means to the end of model selection; in Wallace's view, the coded parameter estimates are important in their own right.

4.2 MDL with Multi-Part Codes

The goal here is to encode the data x with a three-part message. The first part indicates the model m, the second encodes parameters θ for that model, and the third encodes the data x given the model m and parameters θ. The total message length is

$$len(x, \theta, m) = len(x|\theta, m) + len(\theta|m) + len(m). \tag{20}$$

If the number of models $|\mathcal{M}|$ is finite, then from the discussion in the previous section, assuming that $len(m)$ is constant minimizes the *worst-case* number of bits needed to specify the model. In the literature, then, the explicit dependence on m is sometimes dropped and (20) is written as

$$len(x, \theta) = len(x|\theta) + len(\theta). \tag{21}$$

Hence, this procedure is most often referred to as a two-part coding scheme. (One exception to this is the intriguing study of ancient stone circle geometries by Patrick and Wallace (1982), who

in fact characterize models by the lengths of programs written in an ALGOL-like language. This is kindred to Chaitin's (1987; 1997) vivid approach to algorithmic information theory, which defines the complexity of expressions via the lengths of LISP programs.)

The "minimum description length" or "minimum message length" principle of model selection and parameter estimation chooses the θ and m which minimizes (20) for the collected data x. As Rissanen (1983) observes, "quite a bit is needed to convert this commonsensical principle to an explicit formula, directly applicable to a variety of estimation problems." In applying the principle, at certain points assumptions and approximations need to be made to obtain implementable procedures; different sets of choices will yield different expressions.

Shannon's theory tells us that for a given model m and parameter θ, we can construct a code for x with codewords of length $len(x|\theta, m) = \lceil -L(x|\theta, m) \rceil$. (For convenience, we will drop the notation for "next largest integer" in the remaining discussion.) Fortunately, it is not actually necessary to construct such codes; we only need to know that it is possible and to have expressions for codeword lengths. (A referee pointed out that many researchers feel the MDL principle cannot be used in their context because they mistakenly think they would need to actually construct the codes, which is often computationally infeasible.) If θ could somehow be transmitted cost-free, then we would choose the maximum-likelihood estimate, the θ which minimizes $len(x|\theta, m)$. However, we must also encode and transmit the parameter θ. In this data transmission viewpoint, the code for θ must be a prefix (also called "self-punctuating") code. This means that the stream representing x given θ may follow the stream representing θ without an additional "comma" symbol. This implies the code for θ must satisfy the Kraft inequality (Cover & Thomas, 1991, Section 5.2, p. 82)

$$\sum_{\theta} e^{-len(\theta)} \leq 1. \tag{22}$$

Hence, $p_p(\theta) \propto e^{-len(\theta)}$ gives us a proper prior distribution on x. Similarly, if we have $p_p(\theta)$ in hand, we can use it to find the code lengths $len(\theta)$. In this sense, MDL and Bayesian MAP estimation are equivalent.

When we extend the MDL principle to continuous parameter spaces, several complications arise, and this relationship no longer holds exactly; the often-stated equivalence between MAP and MDL is then a potentially misleading oversimplification (see the first section of Grünwald (1998) for discussion of this "folklore"). The issues are subtle and often not clearly explained (or not addressed at all) in the literature; hence, we will spend some time exposing them here.

The main complication is that to apply coding theory, continuous parameters and data must be truncated to finite precision. As we will see, truncation of the data presents no serious hurdles. Truncation of the parameters, however, is a rather sticky issue. Sometimes a useful truncation level seems evident in the nature of the problem; for instance, in describing the location of image points for classifying edges and junctions, Lindeberg and Li (1997) feel "it is natural to set this parameter to a value on the same order as the distance between adjacent pixels." However, the optimum precision for parameters is in fact determined by the data themselves, which makes the interpretation of $len(\theta|m)$ as a prior break down; the $len(\theta|m)$ is now implicitly a function of the data x, and the notion of a prior which depends on the data is an uncomfortable one at best. Sometimes this quandary is handled via averaging over the data; sometimes it seems to go entirely unaddressed.

Inference via the minimization of (20) has been studied most in depth by Rissanen, who uses term "minimum description length," and Wallace, who prefers the name "minimum message length." There are several important differences in their viewpoints, some philosophical and some practical. Rissanen's work tends to focus on the selection of models; the estimation of parameters within those models is only emphasized in as much as such estimates are needed to solve the model selection problem. With this in mind, Rissanen most often employs maximum-likelihood

estimates for parameters within models, even though technically, the MDL parameter estimate might differ from the ML estimate.

Rissanen writes (beginning of Section 4, 1983): "In many cases of interest the first [likelihood] term in [the description length] is dominant, and for each number of parameters the minimizing parameter values are *close* to the ML estimates." [Italics ours.] Thus his observation (in the introduction to the same paper) that the "Minimum Description Length (MDL) principle...turns out to degenerate to the more familiar Maximum Likelihood (ML) principle in case the number of parameters in the models is fixed, so that the description length of the parameters themselves can be ignored" only holds in this approximate sense. Although not obvious upon first reading of his papers, this becomes clear under deeper study. However, the belief that MDL estimation reduces *exactly* to ML estimation in fixed-order cases appears to persist in the applied literature.

Rissanen truncates ML estimates based on a worst-case analysis and tries to minimize the error associated with falling on the furthest edge of a truncation region. Wallace uses a different method of truncation in which his MML estimates may differ from ML and even MAP estimates; such MML estimates have certain invariance properties that make them attractive, *even in the fixed-model case when no model selection needs to be made.* See Wallace & Freeman (1987a, Sections 4.2 and 5.1) for further elucidation of this issue. Fixed-model MML estimators have been explored for binomial (Baxter, 1996a), Von Mises (Dowe et al., 1996b), and spherical Fisher (Dowe et al., 1996a) distributions. Takeuchi (1997) explores the relationships of Bayesian and Wallace-type estimators in the fixed-family case.

Remark: The "parameter truncation" is only needed in order to place our procedures in a coding context, so in practice it is not necessary to restrict ourselves to reporting strictly truncated estimates. However, Wallace & Freeman (1987a, Section 4) do discuss "strict" minimum message length estimators which yield truncated estimates. They note that such parameter estimators are discontinuous functions of the data and in general computationally intractable, so they spend the remainder of their paper exploring the less-strict MML estimators reviewed here.

4.3 A Bayesian Variant of Rissanen's MDL

This section considers employing Rissanen's method of parameter truncation in a Bayesian scenario. The overall approach is to describe the observed data x with a two-stage code in which we encode the MAP estimate $\hat{\theta}$ and then encode x under the model determined by $\hat{\theta}$. As introduced in the preceding section, Shannon's theory informs us that we can construct a code with length $\lceil -L(x|\theta, m) \rceil$. For real data x with likelihood density $p_l(x|\theta, m)$, we can truncate the data to a desired precision δ_x and replace the density with the probability $p_l(x|\theta, m)\delta_x^d$, yielding a code length $-L(x|\theta, m) - d \ln \delta_x$. Notice the $d \ln \delta_x$ term depends on neither the data x nor the parameter θ. Hence, the precision δ_x plays no role in the minimization and can safely be dropped. In the literature, $-L(x|\theta, m)$ is often referred to as a "code length" even when x is real-valued, with the understanding that any desired precision can be achieved. We adopt these conventions here.

Once we have encoded x given θ, we also need to encode the parameter θ. For discrete parameters, the code length for encoding the parameter θ is just $-P(\theta|m) = -\ln p_p(\theta|m)$. But if θ is real, it must be truncated as well. Unlike the truncation of the data, the truncation of the parameters is a significant issue as we will see below.

Assuming the model classes are sufficiently regular, expanding the loglikelihood as a function of the truncated value θ_{tr} in a Taylor series around the MAP estimator $\hat{\theta}$ (to second order) yields

$$\begin{aligned} -H(x, \theta_{tr}|m) &\approx -H(x, \hat{\theta}|m) \\ -\frac{1}{2}(\hat{\theta} - \theta_{tr})^T &\left[\frac{\partial^2}{\partial \theta_i \theta_j} H(x, \theta|m) \right] \Bigg|_{\theta=\hat{\theta}} (\hat{\theta} - \theta_{tr}). \end{aligned} \tag{23}$$

In his earliest work, Rissanen (1978) took the approach of Wallace & Boulton (1968), who treated the quantization error $\hat{\theta} - \theta_{tr}$ as a uniformly distributed random vector and chose the precision to

minimize the expected value of (23). (This is the approach taken by Michal (1993) for multiple target detection with radar antenna arrays.) In subsequent papers, Rissanen (1983) adopts an alternative worst-case minimax approach which picks the precision to minimize the maximum of (23) over the quantization region: "Instead of the maximum value we could calculate the mean increase by assuming some distribution [uniform in Wallace & Boulton (1968)] for the deviation of [the parameter] from the center of its enclosing rectangle. *Using the maximum value has the advantage, however, that it is independent of such distributions.*" [Italics added.]

Rissanen considers the problem of truncation in multiple dimensions as partitioning the parameter space into parallelepipeds, noting that the quadratic term reaches its maximum when θ falls at a corner of the parallelpiped. Consider a fixed maximum deviation

$$r = (\hat{\theta} - \theta_{tr})^T I_H(x : \hat{\theta}|m)(\hat{\theta} - \theta_{tr}). \tag{24}$$

Rissanen (1983) chooses the quantizing parallelpiped as the maximum-volume rectangle contained within the ellipsoid (24). This rectangle has edges parallel to the principal axes of the ellipsoid, with a maximum volume of

$$V(r) = \left(\frac{4r}{d}\right)^{(1/2)d} \frac{1}{\sqrt{\det I_H(x : \hat{\theta}|m)}}. \tag{25}$$

For a worst-case analysis, we seek the r which minimizes

$$\begin{aligned} & -\ln V(r) p_p(\theta_{tr}|m) - L(x|\theta_{tr}, m) \\ = \; & -\ln V(r) - H(x, \theta_{tr}|m) \\ \approx \; & -\ln V(r) - H(x, \hat{\theta}|m) + \frac{r}{2} \\ = \; & -\frac{d}{2}\ln 4r + \frac{d}{2}\ln d + \frac{1}{2}\ln\det I_H(x : \hat{\theta}|m) \\ & -H(x, \hat{\theta}|m) + \frac{r}{2}. \end{aligned} \tag{26}$$

Setting the first derivative w.r.t. r equal to zero easily reveals the minimizing r to be $r = d$. Substituting $r = d$ into (26) yields the description length

$$-H(x, \hat{\theta}|m) + \frac{1}{2}\ln\det I_H(x : \hat{\theta}|m) + \frac{d}{2}(1 - \ln 4). \tag{27}$$

We then seek the m which minimizes (27).

5 Wallace's Minimum Message Length

Since the multivariate application of Wallace's MML inference is somewhat complicated, we will illustrate the ideas by first working through the single variable case. In Section 5.2 we extend this to multiple dimensions. In their work, Wallace & Freeman (1987a) "take the orthodox Bayesian view of the existence of known, proper prior distributions for unknown quantities. It has been known since Jeffreys (1939) that using improper priors for model discrimination leads to indeterminate or nonsensical answers, so there can be no substitute for careful specification of whatever prior knowledge is available." This contrasts sharply with Rissanen's overriding non-Bayesian stance, as they note (1987a, p. 251): "Rissanen finds it meaningful to consider complete, or nearly complete, prior ignorance about a parameter and proceeds to construct a 'universal' prior distribution to be used in all such cases... We prefer the philosophical view that prior information always exists and there can be no easy substitute for thinking what it is and formulating it as well as possible. This

can, of course, only be undermined by any return to the notions of ignorance current in the 1960's and 70's."

Throughout this section, we assume that the model classes are sufficiently regular so that Taylor series approximations will be valid, as discussed in Section 3.

5.1 MML Inference: Univariate case

Wallace & Freeman, data-driven quantization: The prior probability that $\theta \in [\theta_{tr} - \frac{s}{2}, \theta_{tr} + \frac{s}{2}]$ is approximately $sp_p(\theta_{tr})$. Expanding the loglikelihood as a function of the truncated value in a Taylor series around θ (to second order) yields

$$\begin{aligned} -L(x|\theta_{tr}, m) &\approx -L(x|\theta, m) - (\theta - \theta_{tr})\frac{\partial}{\partial\theta}L(x|\theta, m) \\ &\quad -\frac{1}{2}(\theta - \theta_{tr})^2\frac{\partial^2}{\partial\theta^2}L(x|\theta, m). \end{aligned} \tag{28}$$

Wallace & Freeman suppose that the quantization error is uniformly distributed (so $E[\theta - \theta_{tr}] = 0$ and $E[(\theta - \theta_{tr})^2] = s^2/12$) and approximate $sp_p(\theta_{tr}) \approx sp_p(\theta)$, yielding

$$\begin{aligned} & E[-\ln sp_p(\theta_{tr}|m) - L(x|\theta_{tr}, m)] \\ \approx & -\ln s - P(\theta|m) - L(x|\theta, m) + \tfrac{s^2}{24} I_L(x:\theta|m). \end{aligned} \tag{29}$$

Recall that in (23), the first-order Taylor series term vanished since we used the MAP estimate. Here, the first order term vanishes from assumptions on the distribution of the quantization error.

Taking the derivative of (29) w.r.t. s and setting the result equal to zero yields the minimizing s:

$$\hat{s} = \sqrt{\frac{12}{I_L(x:\theta|m)}}. \tag{30}$$

Substituting (30) into (29) yields

$$\begin{aligned} & -P(\theta|m) - L(x|\theta, m) - \frac{1}{2}\ln 12 + \frac{1}{2}\ln I_L(x:\theta|m) + \frac{1}{2} \\ = & -H(x, \theta|m) - \frac{1}{2}\ln 12 + \frac{1}{2}\ln I_L(x:\theta|m) + \frac{1}{2}. \end{aligned} \tag{31}$$

The MML estimate, under this data-dependent quantization scheme, is the θ which minimizes (31). (Observe that if the MAP estimate was used instead of the MML estimate, then (31) would only differ from Oliver & Baxter's version of the Schwarz criterion (8) by a constant.)

Wallace & Freeman, data-independent quantization: From a strict data transmission standpoint, (31) is not enough, as the precision s must also be coded and transmitted. s itself must be then be quantized to some optimal precision s_1, which must in turn be quantized and transmitted with some optimal precision s_2, and so on. To avoid becoming ensnared in this quicksand, Wallace & Freeman (1987a, p. 245) give several arguments to try to illustrate that the effect is not significant. More generally, they resolve the issue by approximating the observed Fisher information with the expected Fisher information,

$$I_L(\theta|m) = \int p_l(x|\theta) I_L(x:\theta|m) dx, \tag{32}$$

yielding the approximation

$$-H(x, \theta|m) - \frac{1}{2}\ln 12 + \frac{1}{2}\ln I_L(\theta|m) + \frac{1}{2}. \tag{33}$$

The θ which minimizes (33) is the MML estimate.

O'Hagan's variation, data-driven quantization: O'Hagan (1987, p. 22) observed that Wallace and Freeman's approximation $sp_p(\theta_{tr}|m) \approx sp_p(\theta|m)$ amounts to using a zero-order Taylor approximation for the prior, while a second-order Taylor approximation is used for the posterior. As an alternative to approximating the prior as a constant and expanding the loglikelihood, he suggests expanding the full logposterior:

$$\begin{aligned} -H(x,\theta_{tr}|m) &= -H(x,\theta|m) - (\theta-\theta_{tr})\frac{\partial}{\partial\theta}H(x,\theta|m) \\ &\quad -\frac{1}{2}(\theta-\theta_{tr})^2\frac{\partial^2}{\partial\theta^2}H(x,\theta|m). \end{aligned} \tag{34}$$

Again supposing the quantization error is uniformly distributed so that $E[\theta-\theta_{tr}]=0$ and $E[(\theta-\theta_{tr})^2] = s^2/12$,

$$\begin{aligned} & E[-\log s - P(\theta_{tr}|m) - L(x|\theta_{tr},m)] \\ = \; & -\ln s - H(x,\theta|m) + \tfrac{s^2}{24} I_H(x:\theta|m). \end{aligned} \tag{35}$$

The minimizing s is given by

$$\hat{s} = \sqrt{\frac{12}{I_H(x:\theta|m)}}. \tag{36}$$

Substitution of (36) into (35) yields

$$-H(x,\theta|m) - \frac{1}{2}\ln 12 + \frac{1}{2}\ln I_H(x:\theta|m) + \frac{1}{2}. \tag{37}$$

O'Hagan's MML estimate is the θ which minimizes (37). (As in the preceding section, observe that if the MAP estimate is employed instead of the MML estimate, then (37) would only differ from our original formulation of the Schwarz criterion (5) by a constant.)

O'Hagan's variation, data-independent quantization: As in the original Wallace & Freeman formulation, approximating the observed Fisher information in the O'Hagan variation with the expected Fisher information yields

$$-H(x,\theta|m) - \frac{1}{2}\ln 12 + \frac{1}{2}\ln I_H(\theta|m) + \frac{1}{2}. \tag{38}$$

In their reply to O'Hagan's suggestion, Wallace & Freeman (1987b) give several reasons for their preference of their original formulation (33) over O'Hagan's proposal (37). They suggest that if the prior $p_p(\theta)$ has substantial second derivative, that the distribution of the quantization error can no longer be well approximated by a uniform density. They also note that (33) is invariant to nonlinear transformations of the parameters, while (37) is not. It is interesting to note, though, that Wallace and Freeman do adopt the expected Fisher information version of O'Hagan's variation (38), without making any reference to their previous objections, in their study of factor models of multivariate Gaussian distributions (Wallace & Freeman, 1992) where the variation of the logprior is not negligible compared with the variation of the loglikelihood.

Example: Consider the elementary problem of estimating the mean λ of independent, identically Poisson distributed data $x_1, x_2, \ldots x_N$. Let $X = \sum_i x_i$. The loglikelihood is $-N\lambda + X\ln\lambda$ and the observed Fisher information is X/λ^2. Since $E[X] = E[\sum_i x_i] = \sum_i E[x_i] = \sum_i \lambda = N\lambda$, the expected Fisher information is N/λ. Suppose the prior is sufficiently flat in a region around the ML estimate that it can be neglected. Then, the traditional ML estimator and the data-driven

and data-independent variations of the MML estimator are:

$$\begin{aligned}
\hat{\lambda}_{ML} &= \arg\min_{\lambda}[N\lambda - X\ln\lambda] = \frac{X}{N}, \\
\hat{\lambda}^{DD}_{MML} &= \arg\min_{\lambda}[N\lambda - X\ln\lambda + \frac{1}{2}\ln X - \ln\lambda] \\
&= \frac{X+1}{N}, \\
\hat{\lambda}^{DI}_{MML} &= \arg\min_{\lambda}[N\lambda - X\ln\lambda + \frac{1}{2}\ln N - \frac{1}{2}\ln\lambda] \\
&= \frac{X+\frac{1}{2}}{N}.
\end{aligned} \tag{39}$$

Note that although all three estimators are asymptotically equivalent as $N \to \infty$, they are slightly different for low sample values. Here, the MML approach may yield different answers than the ML estimate, even in this basic estimation problem where no model order selection needs to be made.

5.2 MML in Multiple Dimensions

Now consider extending (33) to the multivariate case. The original derivation by Wallace & Freeman (1987a, Section 5.2) is rather terse, so this section follows Oliver and Baxter's (1994, Section 5, Appendix 1) expanded derivation.

It will be convenient to make a change in coordinates $\tilde{\xi} = B^{-1}\tilde{\theta}$, where B is chosen so that $\tilde{\theta}^T I_L(\theta)\tilde{\theta} = \tilde{\xi}^T\tilde{\xi}$. To make this so, take $B = RS$, where R is a rotation matrix consisting of the eigenvectors of $I_L(\theta)$ arranged in columns, and S is a diagonal matrix consisting of the reciprocal of the square-root of the eigenvalues of $I_L(\theta)$. Let Λ denote the diagonal matrix consisting of the eigenvalues of $I_L(\theta)$. Then

$$\begin{aligned}
\tilde{\theta} I_L(\theta)\tilde{\theta} &= \tilde{\xi}^T B^T I_L(\theta) B\tilde{\xi} = \tilde{\xi}^T S^T R^T I_L R S\tilde{\xi} \\
&\overset{(a)}{=} \tilde{\xi}^T S^T R^T R\Lambda S\tilde{\xi} \overset{(b)}{=} \tilde{\xi}^T S^T I\Lambda S\tilde{\xi} \\
&= \tilde{\xi}^T S^T\Lambda S\tilde{\xi} \overset{(c)}{=} \tilde{\xi}^T I\tilde{\xi} = \tilde{\xi}^T\tilde{\xi},
\end{aligned} \tag{40}$$

where (a) arises from operating on eigenvectors of a matrix with that matrix, (b) is a property of rotation matrices, and (c) is by the definitions of S and Λ.

To avoid confusion, will use (ξ) superscripts on densities and logdensities which are expressed in terms of the new coordinates. In these new coordinates, the prior is

$$\begin{aligned}
p_p^{(\xi)}(\xi) &= \frac{p_p(\theta)}{Jacob(B^{-1})} = \frac{p_p(\theta)}{Jacob(S^{-1})Jacob(R^{-1})} \\
&= \frac{p_p(\theta)}{Jacob(S^{-1})} = \frac{p_p(\theta)}{\sqrt{\det I_L(\theta)}},
\end{aligned} \tag{41}$$

since

$$Jacob(S^{-1}) = \prod_{i=1}^{d}\sqrt{\lambda_i} = \sqrt{\prod_{i=1}^{d}\lambda_i} = \sqrt{\det I_L(\theta)}. \tag{42}$$

Performing the usual Taylor expansion of the loglikelihood, as in (28), in terms of the ξ-coordinates yields

$$-L^{(\xi)}(x|\xi_{tr}, m) \approx -L^{(\xi)}(x|\xi, m) + \frac{(\xi - \xi_{tr})^T(\xi - \xi_{tr})}{2}. \tag{43}$$

In our application of Rissanen's ideas to Bayesian inference in Section 4.3, we explored minimizing the maximum possible error, so no expectations were taken; the first partial derivative was zero in (23) since we were evaluating at the MAP estimate. Here, in Wallace's approach, the term involving the first partial derivative has zero expected value due to the assumption that the quantization error is uniformly distributed.

Instead of Rissanen's parallelpipeds, Wallace & Freeman (1987a, 1992) consider quantizing in multiple dimensions using optimal quantizing lattices. For instance, in two dimensions, the optimal quantizing lattice forms a hexagonal grid. In three dimensions, the optimal lattice is a *body-centered cubic* lattice (Conway & Sloane, 1993, p. 60), whose Voronoi regions (Conway & Sloane, 1993, p. 34) are *truncated octahedrons* (one of the Archimedean polyhedra). In higher dimensions, the optimal quantizing lattices are actually unknown.

Let s denote the volume of the quantization region. If we quantize according to an optimum regular lattice, then

$$E[(\xi - \xi_{tr})^T(\xi - \xi_{tr})] = d\kappa_d s^{2/d}, \tag{44}$$

where κ_d is a constant relating to the geometry of the d-dimensional lattice. The constant κ_d is not known for every d, although upper and lower bounds are available (Zador, 1982; Wallace & Freeman, 1987a). Upper and lower bounds for κ_d are given by

$$\frac{\Gamma(\frac{d}{2}+1)^{2/d}\Gamma(\frac{2}{d}+1)}{d\pi} > \kappa_d > \Gamma(\frac{d}{2}+1)^{2/d}(d+2)\pi. \tag{45}$$

As d grows, the upper and lower bounds converge to the same number (Conway & Sloane, Eq. 82, p. 58), yielding $\kappa_d \to \frac{1}{2\pi e} \approx 0.058550$.

Some examples of κ_d for the best known quantizing lattices, taken from Conway & Sloane (1993, Table 2.3, p. 61), are displayed in Table 5.2.

d	κ_d
1	0.083333
2	0.080188
3	0.078743
4	0.076603
5	0.075625
6	0.074244
7	0.073116
8	0.071682
12	0.070100
16	0.068299
24	0.065771

Table 1: Constants κ_d for the best known quantizing lattices in various dimensions d.

We need to find the s which minimizes the approximate expected message length

$$\begin{aligned} & E[-\ln s p_p^{(\xi)}(\xi_{tr}|m) - L^{(\xi)}(x|\xi_{tr}, m)] \\ \approx \quad & -\ln s - \ln p_p^{(\xi)}(\xi|m) - L^{(\xi)}(x|\xi, m) + \frac{d}{2}\kappa_d s^{2/d}. \end{aligned} \tag{46}$$

Straightforward calculus reveals that the minimizing $s = \kappa_d^{-d/2}$. Substituting this back into (46) yields

$$\frac{d}{2}\ln\kappa_d - \ln p_p^{(\xi)}(\xi|m) - L^{(\xi)}(x|\xi, m) + \frac{d}{2}. \tag{47}$$

Using (41) to translate back into the original θ coordinates gives us the message length

$$\begin{aligned} & -P(\theta|m) + \frac{1}{2}\ln\det I_L(\theta|m) - L(x|\theta,m) + \frac{d}{2}(1+\ln\kappa_d) \\ = \quad & -H(x,\theta|m) + \frac{1}{2}\ln\det I_L(\theta|m) + \frac{d}{2}(1+\ln\kappa_d). \end{aligned} \tag{48}$$

The MML procedure seeks the model m and parameters θ which minimize (48). Notice that $\kappa_1 = 1/12 \approx 0.0833$, so (48) reduces to (33) when $d = 1$.

As in the one-dimensional case discussed above, one could follow O'Hagan's reasoning and expand the prior as well, yielding a message length

$$-H(x,\theta|m) + \frac{1}{2}\ln\det I_H(\theta|m) + \frac{d}{2}(1+\ln\kappa_d). \tag{49}$$

We again seek the m and θ which minimize (49). As mentioned earlier, Wallace and Freeman (1992) employ (49) in their study of factor models of multivariate Gaussian distributions.

Wallace & Dowe (2000) employ MML for clustering mixtures of multi-state, Poisson, von Mises, and Gaussian data. Baxter & Oliver (2000) compare the performance of an MML classifier for Gaussian data with several other clustering techniques.

Takeuchi's Description Length: Taking a hybrid approach, in which the truncation error is viewed as a uniform random variable as in Wallace's work, but the regions are taken to be rectangular as in Rissanen's work, yields Takeuchi's description length (Takeuchi, 1997, first equation on p. 1172):

$$-H(x,\theta|m) + \frac{1}{2}\ln\det I_H(\theta|m) + \frac{d}{2}(1-\ln 12). \tag{50}$$

Takeuchi's procedure seeks the model m and parameters θ which minimize (50).

Notice that for $d = 1$, the Takeuchi (50) and Wallace-O'Hagan (49) description lengths are the same, whereas for $d > 1$, Takeuchi's description length is greater due to the nonoptimality of the rectangular quantizing lattice in higher dimensions. Takeuchi refers to estimates which minimize (50) as minimum description length estimators, although they might be more accurately called minimum message length estimators in reverence to Wallace's work. Unlike both Wallace and Rissanen, Takeuchi allows cases with improper priors; since he focuses on "the case of parameter estimation in a fixed family," and not on model selection, Wallace's objections to improper priors noted in the introduction to Section 5 are somewhat alleviated.

6 Penalized Loglikelihood Interpretation

As suggested, for example, by Green (1998), it is often fruitful to express model selection procedures via penalized loglikelihoods:

$$L_{pen}(\theta, x|m) = L(x|\theta,m) - C(\theta,x|m). \tag{51}$$

The penalties associated with Schwarz's approach (5), our Bayesian/Rissanen approach (from Section 4.3), and the Wallace-O'Hagan data-independent quantization approach (49) are given by

$$C_S(\theta,x|m) = -P(\theta|m) + \frac{1}{2}\ln\det I_H(x:\theta|m) - \frac{d}{2}(\ln 2\pi), \tag{52}$$

$$C_R(\theta,x|m) = -P(\theta|m) + \frac{1}{2}\ln\det I_H(x:\theta|m) - \frac{d}{2}(\ln 4 - 1), \tag{53}$$

$$C_W(\theta,x|m) = -P(\theta|m) + \frac{1}{2}\ln\det I_H(\theta|m) - \frac{d}{2}(\ln\frac{1}{\kappa_d} - 1). \tag{54}$$

For C_R and C_S, θ is the MAP estimate (in order to make the Laplace approximation and Rissanen's truncation techniques work), whereas for C_W, we take the θ which minimizes $C_W(\theta, x|m)$.

In the applied literature, C is occasionally interpreted as a logprior; this is somewhat misleading. We prefer to think of C as an adaptive complexity penalty.

Although they arrive from quite different viewpoints, Schwarz, Rissanen, and Wallace yield strikingly similar criteria. All three criteria have terms involving the posterior, a term which is one-half the log of the Fisher information (either the observed or the expected Fisher information, according to the particular strategy), and a term which is linear (for Rissanen and Schwarz) or almost linear (for Wallace) in the number of parameters. All three techniques involve a Taylor series expansion, so the appearance of the Fisher information terms is perhaps not too surprising.

Yang & Barron (1998) consider a wide class of penalized-likelihood selection criteria, including some penalties which do not necessarily have a description length interpretation.

7 MDL for Unbounded Nonrandom Parameters via a Universal Code for the Integers

In Section 4.3, we explored the MDL idea assuming that a prior on the parameters is available. In the early 80's, Rissanen (1983; 1987a) proposed a novel way of getting around the prior issue by representing parameters via a "universal" encoding of the integers.

Following his usual "worst-case" line of reasoning, Rissanen derives such a universal prior from a particular minimax problem. Suppose we must encode natural numbers generated by a probability distribution $p(\theta)$, where p is unknown, but is known to belong to some class $\mathcal{P}$. The objective is to find a prefix code with length $len(\theta)$ which minimizes the worst-case average redundancy

$$\min \sup_{p \in \mathcal{P}} \sum_{\theta \in \mathbf{N}} p(\theta) \frac{len(\theta)}{H(p)}, \tag{55}$$

where the minimization is taken over codes which satisfy the Kraft inequality and $H(p)$ is the entropy of p. If $\mathcal{P}$ is the set of non-singular distributions with finite support, then the minimax prior is the uniform prior over the region of support, which coincides with our intuitive notions of uninformative prior knowledge on bounded intervals. If $\mathcal{P}$ is the set of distributions with infinite support and infinite entropy that are nondecreasing in θ, then Rissanen (1983, Appendix B) proves that the minimax code has codewords of length $len(\theta) = \log^* \theta + \log(c)$ bits, where

$$\log^* \theta = \log \theta + \log \log \theta + \log \log \log \theta + \cdots \tag{56}$$

and $c \approx 2.865064$. The logarithms are base-2 and the sum only includes positive terms. See Rissanen (1983, p. 420) or Cover & Thomas (1991, p. 150) for an explanation of the log^* integer encoding scheme.

7.1 The Univariate Case with $\log^*$ MDL

If θ is a nonnegative real, Rissanen suggests converting it to an integer by dividing by a precision δ_θ, and encoding it using $\log^*(\theta/\delta_\theta) + \log(c)$ bits. This forms the basis for Rissanen's "objective" minimum description length criteria. This approach is most natural in nested models in which setting a parameter to zero is equivalent to using a lower-dimensional model; in these cases, there is an obvious canonical origin to the coordinate system.

Since our loglikelihoods are often conveniently expressed in terms of natural logarithms, it will be convenient to denote $\ln^* \theta = \ln(2) \log^* \theta$, so we can convert bits in the $\log^*$ formulation into nats. Again assuming appropriate regularity conditions, performing the now familiar second-order Taylor series expansion on the loglikelihood around the maximum-likelihood estimate $\hat{\theta}(x) =$

$\arg\max_\theta L(x:\theta)$ and substituting $\delta_\theta = \hat{\theta}(x) - \theta_{tr}$ for a worst-case analysis yields the total approximate description length

$$-L(x:\hat{\theta}(x)) + \frac{1}{2}\left(\frac{\delta_\theta}{2}\right)^2 I_L(x:\hat{\theta}(x)) + \ln^*\left(\frac{\hat{\theta}(x)}{\delta_\theta}\right) + \ln(c). \tag{57}$$

Finding the optimum δ_θ is difficult. If we approximate $\ln^* \approx \ln$, then straightforward differentiation yields an approximate optimum precision

$$\frac{\delta_\theta}{4} I_L(x:\hat{\theta}(x)) - \frac{1}{\delta_\theta} = 0 \;\Rightarrow\; \delta_\theta = \frac{2}{\sqrt{I_L(x:\hat{\theta}(x))}}. \tag{58}$$

Substituting the approximate optimum δ_θ back into (57) yields

$$-L(x:\hat{\theta}(x)) + \frac{1}{2} + \ln^*\left(\frac{\hat{\theta}(x)\sqrt{I_L(x:\hat{\theta}(x))}}{2}\right) + \ln(c). \tag{59}$$

If we have N independent snapshots, the usual asymptotics for $N \to \infty$ apply if we again approximate $\ln^* \approx \ln$:

$$\begin{aligned} & -L(x:\hat{\theta}(x)) + \frac{1}{2} + \ln\hat{\theta}(x) + \frac{1}{2}\ln I_L(x:\hat{\theta}(x)) - \ln 2 + \ln(c) \\ \to\quad & -L(x:\hat{\theta}(x)) + \frac{1}{2}\ln N. \end{aligned} \tag{60}$$

Notice that in Rissanen's formulation, the prior exists on the integers resulting from the truncated parameter, not the parameter itself. The prior is *not* used in computing the estimate of the parameter θ; it only appears in computing the description length. This represents a compromise in the Bayesian/non-Bayesian dispute, and contrasts with Wallace's MML principle (Section 5) in which the MML estimates may differ from both the ML and the MAP estimates.

Foster & Stine (1999) explore a variation of Rissanen's log^* scheme in which the parameter estimates are allowed to differ from the ML estimates. They discuss explicit coding schemes which yield estimators which seem to be nondifferentiable, although piecewise differentiable, functions of the data.

Related work: Moulin and Liu have employed $\log^*$ to construct priors on wavelet coefficients for Bayesian image denoising (Liu & Moulin, 1998; Moulin & Liu, 1999).

7.2 The Multivariate Case with $\log^*$ MDL

The multivariate case is substantially more complicated. A naive approach would be to quantize and represent each parameter separately. A better approach, as described in Section 4 of Rissanen (1983), is to divide the space into rectangles and employ a worst case analysis as in Section 4.3, except we encode the representative rectangle using the universal prior on the integers, where the rectangles are numbered in a spiral fashion, with index increasing with distance from the origin. The precise arguments are quite intricate; we refer the reader to Rissanen (1983) for the details, and merely cite the resulting description length:

$$-L(x:\hat{\theta}(x)) + \ln^*(vol(d)[\hat{\theta}(x)^T I_L(x:\hat{\theta}(x))\hat{\theta}(x)]^{d/2}), \tag{61}$$

where $vol(d)$ is the volume of the d-dimensional unit hypersphere

$$vol(d) = \frac{2^{\lfloor (k+1)/2 \rfloor}}{k(k-2)(k-4)\cdots}\pi^{\lfloor k/2 \rfloor}. \tag{62}$$

The MDL procedure seeks the model which minimizes (61).

8 Rissanen's Stochastic Complexity

Later in the 80's, Rissanen (1986; 1987b; 1989) introduced the notion of the *stochastic complexity* of a data set relative to a parametric model $p_l(x|\theta)$, defined as $\int p_l(x|\theta)w(\theta)d\theta$, where $w(\theta)$ is a weighting function. Rissanen prefers to interpret $w(\theta)$ as specifying a convex combination of models, and not as a prior probability. Stochastic complexity has been applied to the problem of nonparametric density estimation (Rissanen et al., 1992; Hall & Hannan, 1988). Later, Rissanen (1995a; 1995b; 1996) profoundly refined his definition of stochastic complexity by removing the dependence of its definition on the weighting function $w(\theta)$. This section explores this polished version of stochastic complexity.

8.1 Minimax Regret and Normalized Maximum-Likelihood

Suppose we have a parametric likelihood model $p_l(x|\theta)$, but that we do not have a prior $p_p(\theta)$. The θ leading to the shortest code length of x is given by the maximum-likelihood estimate $\hat{\theta}(x)$, yielding a code length $-\ln p(x|\hat{\theta}(x))$. In Section 7, the problem was addressed by encoding $\hat{x}$ with some precision using a "universal" prior on the integers. Here we explore an approach presented by Barron, Rissanen,[2] and Yu (1998) which does not explicitly encode θ. If we encoded the data using a density $q(x)$, then the extra number of nats needed, which Barron calls *regret*, in encoding x using $q(x)$ is

$$-\ln q(x) + \ln p_l(x|\hat{\theta}(x)) = \ln \frac{p_l(x|\hat{\theta}(x))}{q(x)}, \tag{63}$$

where we have explicitly noted that $\hat{\theta}$ is a function of x. Suppose we want to choose the $q(x)$ which minimizes the worst case regret:

$$\min_q \max_x \ln \frac{p_l(x|\hat{\theta}(x))}{q(x)}. \tag{64}$$

As cited in Barron et al. (1998), Shtarkov (1987) showed that the solution to this minimax problem is

$$q(x) = \frac{p_l(x|\hat{\theta}(x))}{\int_X p_l(\tilde{x}|\hat{\theta}(\tilde{x}))d\tilde{x}}. \tag{65}$$

Rissanen (1999a) calls this the *normalized maximum-likelihood* (NML) density, and the length of the resulting code,

$$-L(x|\hat{\theta}(x)) + \ln \int_{\mathcal{X}} p_l(\tilde{x}|\hat{\theta}(\tilde{x}))d\tilde{x}, \tag{66}$$

the *stochastic complexity* of the data x under the parametric model p_l; models with lower stochastic complexity are considered preferable to models with higher stochastic complexity. The second term is referred to as the *parametric complexity*, since it indicates the cost of not knowing the parameter θ. Barron & Cover (1991, p. 1038) observe that $p_l(x|\hat{\theta}(x))$ itself cannot be used to create a code; the normalization in (65) is essential. In a few problems, such as linear regression with Gaussian residuals (Barron et al., 1998, Rissanen, 2000a), (66) can be computed via a direct attack. More often, this is exceptionally difficult, so Rissanen offers an approximation derived in the next two sections.

[2]We are indebted to Jorma Rissanen for answering several pertinent questions on stochastic complexity and sending us a preprint of this paper.

Interestingly, the NML density (65) also solves a related minimax problem involving the *expected* regret:

$$\min_q \max_{g \in \mathcal{G}} E_g \left[\ln \frac{p_l(x|\hat{\theta}(x))}{q(x)} \right]. \tag{67}$$

where $\mathcal{G}$ is a rather broad class of probability distributions which include the parametric models $p_l(x|\theta)$ as a subset (Rissanen, 2000b).

8.2 The Bernardo-Jeffreys Prior

The Bayesian version of Rissanen's MDL and Wallace and Freeman's MML both required a prior $p_p(\theta)$. Sometimes no clear choice of prior is available. Bernardo (1979) suggests a novel way of crafting an "objective" prior based on viewing the statistical estimation problem as a communication channel with source Θ and data X, and choosing the prior to maximize the mutual information $I(\Theta; X) = H(\Theta) - H(\Theta|X)$. Bernardo's idea is that this places the least emphasis on the prior knowledge, and puts the greatest importance on the data. Using arguments from differential geometry, Balasubramanian (1997) substantiates this idea by showing that this prior is a uniform prior on the "natural" space of probability distributions (which is *not* the same as the uniform prior on the space of parameter values). For discrete sources, the iterative Blahut-Arimoto (Blahut, 1972; Arimoto, 1972; O'Sullivan, 1998) algorithm reveals the maximizing prior. For a continuous, sufficiently regular source distribution, Bernardo uses Laplace's approximation, replacing the observed Fisher information with its expected version. We have

$$\begin{aligned}
& H(\Theta|X) \\
= & \int_{\mathcal{X}} p(x) \int_{\mathbf{\Theta}} p(\theta|x) \ln p(\theta|x) d\theta dx \\
= & \int_{\mathcal{X}} p(x) \int_{\mathbf{\Theta}} \frac{p(x|\theta)p(\theta)}{p(x)} \ln \frac{p(x|\theta)p(\theta)}{p(x)} d\theta dx \\
= & \int_{\mathcal{X}} \int_{\mathbf{\Theta}} p(x|\theta)p(\theta) \ln \frac{p(x|\theta)p(\theta)}{p(x)} d\theta dx \\
= & \int_{\mathcal{X}} \int_{\mathbf{\Theta}} p(x|\theta)p(\theta) \{\ln[p(x|\theta)p(\theta)] \\
& - \ln[\int_{\mathbf{\Theta}} p(x|\tilde{\theta})p(\tilde{\theta}) d\tilde{\theta}]\} d\theta dx \\
\approx & \int_{\mathcal{X}} \int_{\mathbf{\Theta}} p(x|\theta)p(\theta) \{\ln p(x|\theta) + \ln p(\theta) \\
& \underbrace{- \ln p(x|\theta) + \ln \sqrt{\det I_L(\theta)} - \frac{d}{2} \ln 2\pi - \ln p(\theta)\}}_{\text{from Laplace approximation of } p(x)} d\theta dx \\
= & \int_{\mathcal{X}} \int_{\mathbf{\Theta}} p(x|\theta)p(\theta) \{\ln \sqrt{\det I_L(\theta)} - \frac{d}{2} \ln 2\pi\} d\theta dx \\
= & \int_{\mathbf{\Theta}} p(\theta) \{\ln \sqrt{\det I_L(\theta)} - \frac{d}{2} \ln 2\pi\} d\theta \\
= & \int_{\mathbf{\Theta}} p(\theta) \ln \sqrt{\det I_L(\theta)} d\theta - \frac{d}{2} \ln 2\pi,
\end{aligned} \tag{68}$$

yielding

$$\begin{aligned} I(\Theta;X) &= \int_{\mathbf{\Theta}} p(\theta)\ln p(\theta) \\ &\quad -\int_{\mathbf{\Theta}} p(\theta)\ln\sqrt{\det I_L(\theta)}d\theta + \frac{d}{2}\ln 2\pi \\ &= \int_{\mathbf{\Theta}} p(\theta)\ln\frac{p(\theta)}{\sqrt{\det I_L(\theta)}}d\theta + \frac{d}{2}\ln 2\pi. \end{aligned} \tag{69}$$

In order for $p(\theta)$ to maximize $I(\Theta;X)$, we need $p(\theta) \propto \sqrt{\det I_L(\theta)}$, hence the Bernardo-Jeffreys "reference" prior[3] is

$$p_p(\theta) = \frac{\sqrt{\det I_L(\theta)}}{\int_{\mathbf{\Theta}}\sqrt{\det I_L(\tilde{\theta})}d\tilde{\theta}}. \tag{70}$$

Several decades before Bernardo, Jeffreys (1946; 1967) derived this prior from a quite different viewpoint involving invariance arguments. Of course, this is only valid if the integral in the denominator of (70) exists. For many interesting problems (for instance, the exponential and Gaussian models of Examples 2 and 3 in Section V of Rissanen (1996)), it does not, and the Bernardo-Jeffreys prior is not directly applicable. Necessary modifications for our application are discussed in the next section.

Usually, a prior is chosen to represent knowledge which exists before any data is taken, and in fact is independent of whatever sensors may be employed. This prior knowledge is then fused with the data through the likelihood to form the posterior distribution; adding additional sensors merely requires adding further loglikelihoods, while the prior may remain unmolested. The Bernardo-Jeffreys prior, on the other hand, is a function of the likelihood, and hence the prior will change as different sensors are added or removed. In this sense, the Bernardo-Jeffreys prior is a creature quite unlike traditional Bayesian priors and cannot be interpreted the same way. See Berger & Bernardo (1992), Bernardo (1997), and Ghosh & Mukerjee (1992) for further discussion on non-informative reference priors.

8.3 Rissanen's Formula for Stochastic Complexity

Suppose we employ the Bernardo-Jeffreys prior

$$p_p(\theta|m) = \frac{\sqrt{\det I_L(\theta|m)}}{\int_{\mathbf{\Theta}}\sqrt{\det I_L(\tilde{\theta}|m)}d\tilde{\theta}}, \tag{71}$$

assuming the integral in the denominator is defined.

Suppose the data is sufficiently strong that the MAP estimate $\hat{\theta}$ may be approximated with the ML estimate $\hat{\theta}_{ML}$. Employing Baxter & Oliver's application of Laplace's method (in which

[3] In the literature, the term "Jeffreys' prior" refers to one of two functions: the improper prior $1/\theta$ on the positive reals, or a prior constructed from a Fisher information matrix as described in this section. We add the name "Bernardo" to make clear that the second meaning is intended, and to emphasize Bernardo's information-theoretic view which motivates its use here.

the likelihood is expanded to second order, but only the constant term of the prior is used) yields

$$\begin{aligned} p(x|m) &= \int p_l(x|\theta)p_p(\theta)d\theta \\ &= \int_{\mathbf{\Theta}} p_l(x|\theta)\frac{\sqrt{\det I_L(\theta|m)}}{\int_{\mathbf{\Theta}}\sqrt{\det I_L(\tilde{\theta}|m)}d\tilde{\theta}} \\ &= \frac{1}{\int_{\mathbf{\Theta}}\sqrt{\det I_L(\tilde{\theta}|m)}d\tilde{\theta}}\int_{\mathbf{\Theta}} p_l(x|\mathbf{\Theta})\sqrt{\det I_L(\theta|m)}d\theta \\ &\approx \frac{1}{\int_{\mathbf{\Theta}}\sqrt{\det I_L(\tilde{\theta}|m)}d\tilde{\theta}}\underbrace{p_l(x|\hat{\theta}_{ML})\frac{(2\pi)^{d/2}}{\sqrt{\det I_L(\hat{\theta}|m)}}}_{\text{expansion of likelihood}} \\ &\quad \times \underbrace{\sqrt{\det I_L(\hat{\theta}|m)}}_{\text{expansion of prior}} \\ &= \frac{1}{\int_{\mathbf{\Theta}}\sqrt{\det I_L(\tilde{\theta}|m)}d\tilde{\theta}}p_l(x|\hat{\theta}_{ML})(2\pi)^{d/2}. \end{aligned} \tag{72}$$

Using some rather complicated coding theoretic arguments, Rissanen shows that the negative logarithm of this approximation to $p(x|m)$ using the Bernardo-Jeffreys prior is in fact an approximate expression for the priorless stochastic complexity defined by (66):

$$-\ln p(x|m) \approx -L(x|\hat{\theta}_{ML}) - \frac{d}{2}\ln 2\pi + \ln\int_{\mathbf{\Theta}}\sqrt{\det I_L(\tilde{\theta}|m)}d\tilde{\theta}. \tag{73}$$

We choose the model m which minimizes (73). The reader is referred to Rissanen (1996) for details. Asymptotically, (73) holds exactly with P-probability one under the assumption that the data are generated by some i.i.d. distribution P in $\mathcal{M}$. If this assumption is dropped, then (66) and (73) and can differ by a constant (Balasubramanian, 1997). Kontkanen *et al.* (2000) discuss the role of stochastic complexity and Fisher information in the context of predictive distributions and Bayesian networks.

Note, as was the case in using the log^* prior in Section 7, that Rissanen does not use the Bernardo-Jeffreys prior to compute a Bayesian estimate θ; the prior only appears in computing the stochastic complexity of data under the model.

If the square root of the determinant of the Fisher information matrix is not integrable, we can integrate over subsets of the full space; Rissanen (1996) suggests using log^* to denote which subset. Several examples of this are given by Rissanen (1996, Section V); this is presented in a general form by Qian & Kunsch (1998, Eq. 7). We choose an increasing sequence $\mathbf{\Theta}(1) \subset \mathbf{\Theta}(2) \subset \cdots$ of bounded open subsets converging to $\mathbf{\Theta}$, and take the stochastic complexity to be

$$\begin{aligned} &-\ln p(x|m) \approx -L(x|\hat{\theta}_{ML}) - \frac{d}{2}\ln 2\pi \\ &+\ln\int_{\mathbf{\Theta}(\hat{a})}\sqrt{\det I_L(\tilde{\theta}|m)}d\tilde{\theta} + \ln^*(\hat{a}) + \ln(c), \end{aligned} \tag{74}$$

where $\hat{a}$ is the smallest a such that $\hat{\theta}_{ML} \in \mathbf{\Theta}(a)$.

It is not obvious at present how to choose the best sequences of subsets to use in (74). Indeed, different sequence will lead to different expressions for the stochastic complexity. Rissanen (1996)

chooses them to ease the computation of the integral in (74). Addressing this issue rigorously in a general context should be a subject of future research.

In applying stochastic complexity to the least-squares linear regression problem mentioned in Section 8.1, Rissanen's most recent work (2000a) drops the enumerated subset approach he takes in Section V of Rissanen (1996) in favor of making the limits of the integration a hyperparameter and applying NML to this hyperparameter. Alas, another unbounded normalizing integral is encountered; making the limits of this normalizing integral a hyper-hyperparameter and applying NML a third time finally reveals an exact closed-form (albeit rather cumbersome) formula for the stochastic complexity. It is not clear at present whether such a tactic of repeatedly limiting integrals and applying NML to the limits will yield fruit in other problems.

Remark: Using different assumptions and approximations than Rissanen, Qian & Kunsch (1998) derive a different approximation for stochastic complexity. We mention it for completeness and will not consider it further here.

9 Conclusions

This paper has detailed a wide variety of model selection methods, and attempted to alleviate some common misconceptions about them. Some of the methods, such as those presented in Sections 3, 4.3, and 5, are fundamentally Bayesian in nature.

Section 7 presented Rissanen's log^*-based MDL and Section 8.3 discussed his recent formula for stochastic complexity. Both approaches are manifestly non-Bayesian and attempt to provide "objective" inference in cases where little prior knowledge is available. They are also not based on the "classical frequentist" philosophy which underlies AIC and cross-validation (Smyth 2000). The original early-80's MDL approach employs a "universal" prior on the integers; the stochastic complexity tactic uses no prior at all, but can be seen to be almost equivalent to using a Bernardo-Jeffreys prior. These technical priors are constructed so as to inject as little prior knowledge into the inference as possible, and are solely used to determine model class, and not in estimating the parameters within that class.

We conclude by suggesting three areas of exploration which we have not addressed. One is Barron's work on minimum complexity density estimation (Barron & Cover, 1991; Yang & Barron, 1998), in which he shows not only that such estimators converge to the correct model, but that the rate of convergence can be bound by an *index of resolvability.* Some intriguing results on minimum complexity estimators are presented by Chi & Geman (1998). Another area is predictive inference via multiple models, as discussed by Rissanen (1984) and Gelman *et al.* (1996), where the goal is to infer future samples from previous samples. MDL and stochastic complexity, even for fixed length data, have a compelling predictive interpretation (Rissanen 1986; 1987b). Grünwald *et al.* (1998) compare MDL and MML from a predictive point of view. The predictive interpretation of MDL has close analogies with Dawid's *prequential principle* (Dawid 1984; 1992) and a version of cross-validation called *forward validation* (Rissanen 1989, Section 3.3). In some cases, it is not necessary to pick a specific model order; it may be better, in fact, to employ all the models in a weighted fashion (Wasserman, 2000; Hoeting et al., 1999; Chickering & Heckerman, 2000), as Singer & Feder (1999) illustrate for linear least-squares prediction problems.

The final avenue we would like to mention is the field of *rate-distortion theory*, which addresses lossy source coding subject to a fidelity constraint. Liu & Moulin (1997) explore some connections between MDL and operational rate-distortion curves in the application of image denoising via wavelets. Although it seems that the quantization questions confronted in Sections 4, 5, and 7 would mesh naturally with a rate-distortion framework, little further work seems to have been done in employing rate-distortion results in MDL and MML contexts. Such an approach seems compatible with the information theoretic viewpoint of Bernardo (Section 8.2), as well as work by Shusterman, Miller, & Rimoldi (1997) on designing target libraries for automatic target recognition

via rate-distortion techniques.

10 Acknowledgements

We are grateful to Dr. Pierre Moulin of the Univ. of Illinois at Urbana-Champaign and Dr. Natalia Schmit, Dr. Joseph O'Sullivan, Dr. Donald Snyder, and Dr. Daniel Fuhrmann of Washington Univ. in St. Louis for their detailed readings of various drafts of the manuscript and their helpful suggestions. We also thank Dr. Andrew Barron of Yale Univ. for a wonderful discussion on these topics in the lobby of the Hotel Loretto in Sante Fe, New Mexico during the 1999 IEEE Information Theory Workshop on Detection, Estimation, Classification and Imaging. The anonymous referees offered numerous deep, challenging, and encouraging comments and suggestions which greatly contributed to the shape of this paper.

This research was supported by ARO DAAH04-95-0494, ONR N00014-94-1-0859, ONR/AASERT N00014-94-1-1135, ONR N00014-95-0095, and ARO/AASERT DAAH04-94-G-0209, and in part by DARPA F49620-98-1-0498 administered by AFOSR.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. & Cáski, F., editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akademiai Kaidó. Reprinted in *Breakthroughs in Statistics*, Ed. S. Kotz & N.L. Johnson, (1992), volume I, pp. 599-624. New York: Springer.

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

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika*, 66(2):237–242.

Arimoto, S. (1972). An algorithm for computing the capacity of an arbitrary discrete memoryless channel. *IEEE Trans. on Information Theory*, 18:14–20.

Atkinson, A. (1978). Posterior probabilities for choosing a regression model. *Biometrika*, 65(1):39–48.

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

Barron, A. & Cover, T. (1991). Minimum complexity density estimation. *IEEE Trans. on Information Theory*, 37(4):1034–1054.

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Trans. on Information Theory*, 44(6):2743–2760.

Baxter, R. (1996a). The likelihood principle and MML estimators. In *Proceedings of the International Conference in Information, Statistics, and Induction in Science*, pages 292–303. World Scientific. Available at www.cs.monash.edu.au/˜rohan/PAPERS/lp.ps.

Baxter, R. (December 1996b). *Minimum Message Length Inductive Inference: Theory and Applications*. Ph.D. Dissertation, Department of Computer Science, Monash University, Clayton, Australia.

Baxter, R. & Oliver, J. (1994). MDL and MML: Similarities and differences (introduction to Minimum Encoding Inference - part III). Technical Report 207, Department of Computer Science, Monash University. available at www.cs.monash.edu.au/~rohan/PAPERS/intro.3.ps.

Baxter, R.A. & Oliver, J.J. (2000). Finding overlapping components with MML. *Statistics and Computing*, 10(1):5–16.

Berger, J.O. & Bernardo, J.M. (1992). On the development of reference priors. In Bernardo, J.M., Berger, J.O., Dawid, A.P., and Smith, A.F.M., editors, *Bayesian Statistics 4*, pages 35–60. Oxford Univ Press.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society B*, 41(2):113–147.

Bernardo, J.M. (1997). Non-informative priors do not exist: a dialogue with José M. Bernardo. *Journal of Statistical Planning and Inference*, 65:159–189.

Blahut, R. (1972). Computation of channel capacity and rate distortion functions. *IEEE Trans. on Information Theory*, 18:460–473.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in informational complexity *Journal of Mathematical Psychology*, Special Issue in Model Selection, in press.

Carlin, B. & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B*, 57(3):473.

Cavanaugh, J.E. & Shumway, R.H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference*, 67(1):45–65.

Chaitin, G. (1987). *Algorithmic Information Theory*. Cambridge University Press.

Chaitin, G. (1997). *The Limits of Mathematics: A Course on Information Theory and the Limits of Formal Reasoning*. Springer-Verlag.

Chi, Z. & Geman, S. (1998). On the consistency of minimum complexity nonparametric estimation. *IEEE Trans. on Information Theory*, 44(5):1968–1973.

Chickering, D.M. & Heckerman, D. (2000). A comparison of scientific and engineering criteria for Bayesian model selection. *Statistics and Computing*, 10(1):55–62.

Clarke, B.S. & Barron, A.R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. on Information Theory*, 36(3):453–471.

Clarke, B.S. & Barron, A.R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41:37–40.

Conway, J. & Sloane, N. (1993). *Sphere Packings, Lattices and Groups*. Springer-Verlag, New York, second edition.

Cover, T. & Thomas, J. (1991). *Elements of Information Theory*. John Wiley and Sons, New York.

Dawid, A.P. (1992). Prequential analysis, stochastic complexity, and Bayesian inference. In Bernardo, J.M., Berger, J.O., Dawid, A.P., and Smith, A.F.M., editors, *Bayesian Statistics 4*, pages 109-125. Oxford Univ Press.

Dawid, A.P. (1984). Present position and potential developments: some personal views, statistical theory, the prequential approach. *Journal of the Royal Statistical Society A*, 147(2):278-292.

Dowe, D., Oliver, J., & Wallace, C. (1996a). MML estimation of the parameters of the spherical Fisher distribution. In Arikawa, S. & Sharma, A., editors, *Seventh International Workshop on Algorithmic Learning Theory*, pages 213–227, Berlin. Springer-Verlag: Lecture Notes in Computer Science.

Dowe, D., Oliver, J., Baxter, R., & Wallace, C. (1996b). Bayesian estimation of the Von Mises concentration parameter. In Hanson, K. & Silver, R., editors, *Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods*, pages 51–59, Dordrecht. Kluwer Academic. Available at www.cs.monash.edu.au/~rohan/PAPERS/maxent.ps.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B*, 57(1):45–97.

Foster, D.P. & Stine, R.A. (1999). Local asymptotic coding and the minimum description length. *IEEE Trans. on Information Theory*, 45(4):1289-1293.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Bayesian posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–807.

Ghosh, J.K. & Mukerjee, R. (1992). Non-informative priors. In Bernardo, J.M., Berger, J.O., Dawid, A.P., and Smith, A.F.M., editors, *Bayesian Statistics 4*, pages 195–210. Oxford Univ Press.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):771–732.

Green, P. & Richardson, S. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society B*, 59:731–792.

Green, P. (1998) Penalized likelihood. In *Encyclopedia of Statistical Sciences,* Update Volume 2. John Wiley & Sons. Available at www.stats.bris.ac.uk/~peter/abstracts/penlik.html.

Grenander, U., Srivastava, A., & Miller, M. (1998). Performance analysis of Bayesian object recognition in computer vision. *IEEE Trans. on Information Theory*, accepted for publication.

Grünwald, P. (1998). *The Minimum Description Length Principle and Reasoning under Uncertainty*. PhD Thesis, Univ. of Amsterdam, The Netherlands. Available as ILLC Dissertation Series 1998-03; see robotics.stanford.edu/~grunwald.

Grünwald, P. (1998). Minimum encoding approaches for predictive modeling. In Cooper, G. & Moral, S., editors, *Proceedings of the Fourteenth International Conference on Uncertainty in Artificial Intelligence (UAI '98)*, pages 182–192. Morgan Kauffmann Publishers, San Francisco, CA.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, Special Issue in Model Selection, in press.

Hall, P. & Hannan, E. (1988). On stochastic complexity and nonparametric density estimation. *Biometrika*, 75(1):705–714.

Hansen, M. & Yu, B. (1998). Model selection and the principle of minimum description length. Submitted to the *Journal of the American Statistical Association*, available at cm.bell-labs.com/who/cocteau/papers/postscript/main.ps.gz.

Heckerman, D., Geiger, D., & Chickering, D.M. (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243.

Hoeting, J.A., Madigan, D., Raftery, A.E., & Volinsky, C.T. (1995). Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–417.

Hook, E., Redfearn, W., & Regal, R. (1995). Two pi or not two pi? Comparisons of various methods of model selection and adjustment for model uncertainty in simulations of capture-recapture estimates used in epidemiology. In *Workshop on Model Uncertainty and Model Robustness*, Bath, England. Abstract available at www.isds.duke.edu/conferences/bath/abstracts.html.

Hurvich, C., Simonoff, J., & Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike criterion. *Journal of the Royal Statistical Society B*, 60(2):271–293.

Jeffreys, H. (1939). *Theory of Probability*. Clarendon Press, Oxford, first edition.

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

Jeffreys, H. (1967). *Theory of Probability*. Clarendon Press, Oxford, third edition.

Kontkanen, P., Myllymäki, P., Silander, T., Tirri, H., & Grünwald, P. (2000). On predictive distributions and Bayesian networks. *Statistics and Computing*, 10(1):39–54.

Lanterman, A.D. (1998a). Minimum description length understanding of infrared scenes. In Sadjadi, F., editor, *Automatic Target Recognition VIII*, SPIE Proc. 3371, Orlando, FL. 375–386.

Lanterman, A.D. (1998b). *Modeling Clutter and Target Signatures for Pattern-Theoretic Understanding of Infrared Scenes*. D.Sc. Dissertation, Dept. of Electrical Engineering, Sever Institute of Technology, Washington Univ., St. Louis, MO.

Lanterman, A.D. (2000). Bayesian inference of thermodynamic state incorporating Schwarz-Rissanen complexity for infrared target recognition. *Optical Engineering*, 39(5):1282–1292.

Li, M. & Vitányi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications* Springer-Verlag, New York, second edition.

Liang, Z., Jaszczak, R., & Coleman, R. (1992). Parameter estimation of finite mixtures using the EM algorithm and information criteria with application to medical imaging processing. *IEEE Trans. on Medical Imaging*, 39(4):1126–1133.

Lindeberg, T. & Li, M.-X. (1997). Segmentation and classification of edges using minimum description length approximation and complementary junction cues. *Computer Vision and Image Understanding*, 67(1):88–98.

Linhart, H. & Zucchini, W. (1986). *Model Selection*. Wiley, New York.

Liu, J. & Moulin, P. (1997). Complexity-regularized image denoising. In *Proceedings of the 1997 IEEE International Conference on Image Processing*, Chicago, IL. Available at www.ifp.uiuc.edu/~moulin/paper.html.

Liu, J. & Moulin, P. (1998). Complexity-regularized image restoration. In *Proceedings of the 1998 IEEE International Conference on Image Processing*, Chicago, IL. Available at www.ifp.uiuc.edu/~moulin/paper.html.

Mark, K. E. & Miller, M. I. (1992). Bayesian model selection and minimum description length estimation of auditory-nerve discharge rates. *Journal of the Acoustical Society of America*, 91(2):989–1002.

A.D. McQuarrie (1999). A small-sample correction for the Schwarz SIC model selection criterion. *Statistics and Probability Letters*, 44(1):79–86.

Michal, D. (1993). *Multiple Target Detection for an Antenna Array.* D.Sc. Dissertation, Department of Electrical Engineering, School of Engineering and Applied Science, Washington University, St. Louis, MO.

Moulin, P. & Liu, J. (1999). Analysis of multiresolution image denoising schemes using generalized-Gaussian and complexity priors. *IEEE Trans. on Information Theory*, 45(3):909-919.

O'Hagan, A. (1987). Contribution to the discussion of the papers by Dr. Rissanen and Professors Wallace and Freeman. *Journal of the Royal Statistical Society B*, 49(3):256–257.

O'Hagan, A. (1995). Fraction Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society B*, 57(1):99–138.

Oliver, J. & Baxter, R. (1994). MML and Bayesianism: Similarities and differences (introduction to Minimum Encoding Inference - part II). Technical Report 206, Department of Computer Science, Monash University. Available at www.cs.monash.edu.au/~jono/TechReports/intro.2.ps.

Oliver, J. & Hand, D. (1994). Introduction to minimum encoding inference. Technical Report 205, Department of Computer Science, Monash University. Available at www.cs.monash.edu.au/~jono/Open_Uni/TR4-94.ps.

O'Sullivan, J.A. (1998). Alternating minimization algorithms: From Blahut-Arimoto to expectation-maximization. In Vardy, A., editor, *Codes, Curves, and Signals: Common Threads in Communication*, chapter 10, pages 173–192. Kluwer Academic.

Papoulis, A. (1991). *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, Inc., New York, third edition.

Patrick, J. & Wallace, C. (1982). Stone circle geometries: An information theory approach. In Heggie, D., editor, *Archaeoastronomy in the Old World*, pages 231–264. Cambridge Univ. Press.

Polya, G. & Szego, G. (1972). *Problems and Theorems in Analysis I.* Springer-Verlag, New York.

Poskitt, D. (1987). Precision, complexity and Bayesian model determination. *Journal of the Royal Statistical Society B*, 49(2):199–208.

Qian, G. & Kunsch, H. (1998). Some notes on Rissanen's stochastic complexity. *IEEE Trans. on Information Theory*, 44(2):782–896.

Raftery, A. (1996). Hypothesis testing and model selection. In W. Gilks, S. Richardson, D. S., editor, *Markov Chain Monte Carlo in Practice*, chapter 10, pages 381–399. Chapman and Hall.

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

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431.

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Trans. on Information Theory*, 30:629–636.

Rissanen, J. (1986). Stochastic complexity and modeling. *The Annals of Statistics*, 14(3):1080–1100.

Rissanen, J. (1987a). *Minimum-Description-Length Principle*. In *Encyclopedia of Statistical Sciences*, volume 5, pages 523–527. John Wiley & Sons, New York.

Rissanen, J. (1987b). Stochastic complexity. *Journal of the Royal Statistical Society B*, 49(3):223–239.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Co., New York.

Rissanen, J. (1995a). Stochastic complexity and its applications. In *Workshop on Model Uncertainty and Model Robustness*. On-line proceedings only; paper available at www.isds.duke.edu/conferences/bath/rissanen.ps.

Rissanen, J. (1995b). Stochastic complexity in learning. In *Computational Learning Theory: Second European Conference, EuroCOLT '95, Barcelona, Spain*, pages 196–210. Springer-Verlag: Lecture Notes in Computer Science.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. on Information Theory*, 42(1):40–47.

Rissanen, J. (1999). Hypothesis selection and testing by the MDL principle. *The Computer Journal*, 42(4):261–269.

Rissanen, J. (1999). *Lectures on Statistical Modeling Theory*. Available at www.cs.tut.fi/~rissanen/papers/lectures2.ps.

Rissanen, J. (2000). MDL denoising. Submitted to *IEEE Trans. on Information Theory*. Available at www.cs.tut.fi/~rissanen/papers/denoise.ps.

Rissanen, J. (2000). Strong optimality of the normalized ML models as universal codes. Submitted to *IEEE Trans. on Information Theory*. Available at www.cs.tut.fi/~rissanen/papers/bound.ps.

Rissanen, J., Speed, T., & Yu, B. (1992). Density estimation by stochastic complexity. *IEEE Trans. on Information Theory*, 38(2):315–323.

Rohman, M.S. & King, M.L. (1999). Improved model selection criterion. *Communications in Statistics: Simulation and Computation*, 28(1):51–71.

Sakamoto, Y. (1992). *Categorical Data Analysis by AIC*. Kluwer Academic Publishers, Dordrecht. Translated from the Japanese.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. Kluwer Academic Publishers, Dordrecht. Translated from the Japanese.

E. Shusterman, Miller, M., & B. Rimoldi (1997). Rate-distortion theoretic codebook design for automatic object recognition. Monograph, Electronic Systems and Signals Research Laboratory, Washington University, St. Louis, Missouri.

Schwarz, G. (1978a). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Shibata, R. (1976). Selection of the order of an auto-regressive model by Akaike's information criterion. *Biometrika*, 63:117–126.

Shtarkov, Y. (1997). Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3–17.

Shun, Z. & McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society B*, 57(4):749–760.

Singer, A. & Feder, M. (1999). Universal linear prediction by model order weighting. *IEEE Trans. on Signal Processing*, 47(10):2685–2699.

Smith, K. & Miller, M. (April 1990). A Bayesian approach incorporating Rissanen complexity for learning Markov random field texture models. *Proceedings of the Intl. Conference on Acoustics, Speech, and Signal Processing*, 4:2317–2320.

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 9(1):63–72.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society B*, 39:44–47.

Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society B*, 41(2):276–278.

Takeuchi, J. (1997). Characterization of the Bayes Estimator and the MDL estimator for exponential families. *IEEE Trans. on Information Theory*, 43(4):1165–1174.

Taylor, C. (1987). Akaike's information criterion and the histogram. *Biometrika*, 74(3):636–639.

Wallace, C. & Boulton, D. (1968). An information measure for classification. *The Computer Journal*, 11(2):195–209.

Wallace, C.S. & Dowe, D.L. (1999). Minimum message length and Kolmogorov complexity. *The Computer Journal*, 42(4):270–283.

Wallace, C.S. & Dowe, D.L. (2000). MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing*, 10(1):73–83.

Wallace, C. & Freeman, P. (1987a). Estimation and inference by compact coding. *Journal of the Royal Statistical Society B*, 49(3):241–252.

Wallace, C. & Freeman, P. (1987b). Reply to the discussion of the papers by Dr. Rissanen and Professors Wallace and Freeman. *Journal of the Royal Statistical Society B*, 49(3):262–265.

Wallace, C. & Freeman, P. (1992). Single-factor analysis by minimum message length estimation. *Journal of the Royal Statistical Society B*, 54(1):195–209.

Wax, M. & Kailath, T. (1985). Detection of signals by information theoretic criteria. *IEEE Trans. on ASSP*, 33(2):387–392.

Vitányi, P. & Li, M. (1996). Ideal MDL and its relation to Bayesianism. *Proc. ISIS: Information, Statistics and Induction in Science*, World Scientific, Singapore, 282–291.

Vitányi, P. & Li, M. (2000). Minimum Description Length induction, Bayesianism, and Kolmogorov complexity. *IEEE Trans. on Information Theory*, 46(2):446–464.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, Special Issue in Model Selection, in press.

Yang, Y. & Barron, A. (1998). An asymptotic property of model selection criteria. *IEEE Trans. on Information Theory*, 44(1).

Zador, P. (1982). Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Trans. on Information Theory*, 28:139–149.

Zhang, P. (1993). On the convergence rate of model selection criteria. *Communications in Statistics*, 22:2765-2775.