A Tutorial on Model Selection

A Tutorial on Model Selection

Enes Makalic, Daniel Francis Schmidt and Abd-Krim Seghouane

October 9, 2012

Contents

0.1 Introduction

A common and widespread problem in science and engineering is the determination of a suitable model to describe or characterize an experimental data set. This determination consists of two tasks: (i) the choice of an appropriate model structure, and (ii) the estimation of the associated model parameters. Given the structure or dimension of the model, the task of parameter estimation is generally done by maximum likelihood or least squares procedures. The choice of the dimension of a model is often facilitated by the use of a model selection criterion. The key idea underlying most model selection criteria is the parsimonity principle which says that there should be a tradeoff between data fit and complexity. This chapter examines three broad approaches to model selection commonly applied in the literature: (i) methods that attempt to estimate the distance between the fitted model and the unknown distribution that generated the data (see Section 0.2); (ii) Bayesian approaches which use the posterior distribution of the parameters to make probabilistic statements about the plausibility of the fitted models (see Section 0.3), and (iii) information theoretic model selection procedures that measure the quality of the fitted models by how well these models compress the data (see Section 0.4).

Formally, we observe a sample of n data points $\mathbf{y} = (y_1, \ldots, y_n)'$ from an unknown probability distribution $p^*(\mathbf{y})$. The model selection problem is to learn a good approximation to $p^*(\cdot)$ using only the observed data $\mathbf{y}$. This problem is intractable unless some assumptions are made about the unknown distribution $p^*(\cdot)$. The assumption made in this chapter is that $p^*(\cdot)$ can be adequately approximated by one of the distributions from a set of parametric models $p(\mathbf{y}|\boldsymbol{\theta}; \gamma)$ where $\gamma \in \Gamma \subseteq \mathbb{N}$ defines the model structure and the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)' \in \Theta_\gamma \subset \mathbb{R}^k$ indexes a particular distribution from the model γ. The dimensionality of $\boldsymbol{\theta}$ will always be clear from the context. As an example, if we are interested in inferring the order of a polynomial in a linear regression problem, the set Γ may include first-order, second-order and third-order polynomials. The parameter vector $\boldsymbol{\theta}$ then represents the coefficients for a particular polynomial; for example, the intercept and linear term for the first-order model. The model selection problem is to infer an appropriate model structure γ and parameter vector $\boldsymbol{\theta}$ that provide a good approximation to the data generating distribution $p^*(\cdot)$.

This chapter largely focuses on the inference of the model structure γ. A common approach to estimating the parameters $\boldsymbol{\theta}$ for a given model γ is the method of maximum likelihood. Here, the parameter vector is chosen such that the probability of the observed data $\mathbf{y}$ is maximised, that is

$$\hat{\boldsymbol{\theta}}(\mathbf{y}; \gamma) = \arg\max_{\boldsymbol{\theta} \in \Theta_\gamma} p(\mathbf{y}|\boldsymbol{\theta}; \gamma) \tag{1}$$

where $p(\mathbf{y}|\boldsymbol{\theta}; \gamma)$ denotes the likelihood of the data under the distribution indexed by $\boldsymbol{\theta}$. This is a powerful approach to parameter estimation that possesses many attractive statistical properties [1]. However, it is not without its flaws, and as part of this chapter we will examine an alternative method of parameter estimation that should sometimes be preferred to maximum likelihood (see Section 0.4.2). The statistical theory underlying model selection is often abstract and difficult to understand in isolation, and it is helpful to consider these concepts within the context of a practical example. To this end, we have chosen to illustrate the application of the model selection criteria discussed in this chapter to the important and commonly used linear regression model.

0.1.1 Nested and Non-Nested Sets of Candidate Models

The structure of the set of candidate models Γ can have a significant effect on the performance of model selection criteria. There are two general types of candidate model sets: *nested* sets of candidate models, and *non-nested* sets of candidate models. Nested sets of candidate models have special properties that make

model selection easier and this structure is often assumed in the derivation of model selection criteria, such as in the case of the distance based methods of Section 0.2. Let us partition the model set $\Gamma = \{\Gamma_0, \Gamma_1, \Gamma_2, \ldots\}$, where the subsets Γ_k are the set of all candidate models with k free parameters. A candidate set Γ is considered nested if: (1) there is only one candidate model with k parameters ($|\Gamma_k| = 1$ for all k), and (2) models $\gamma \in \Gamma_k$ with k parameters can exactly represent all distributions indexed by models with less than k parameters. That is, if $\gamma_k \in \Gamma$ is a model with k free parameters, and model $\gamma_l \in \Gamma$ is a model with l free parameters, where $l < k$, then

$$\forall \boldsymbol{\theta}_l \in \Theta_l, \exists \boldsymbol{\theta}_k \in \Theta_k : p(\mathbf{y}|\boldsymbol{\theta}_l; \gamma_l) = p(\mathbf{y}|\boldsymbol{\theta}_k; \gamma_k).$$

A canonical example of a nested model structure is the polynomial order selection problem. Here, a second-order model can exactly represent any first-order model by setting the quadratic term to zero. Similarly, a third-order model can exactly represent any first-order or second-order model by setting the appropriate parameters to zero, and so on. In the non-nested case, there is not necessarily only a single model with k parameters, and there is no requirement that models with more free parameters can exactly represent all simpler models. This can introduce difficulties for some model selection criteria, and it is thus important to be aware of the structure of the set Γ before choosing a particular criterion to apply. The concept of nested and non-nested model sets will be made clearer in the following linear regression example.

0.1.2 Example: The Linear Model

Linear regression is of great importance in engineering and signal processing as it appears in a wide range of problems such as smoothing a signal with polynomials, estimating the number of sinusoids in noisy data and modeling linear dynamic systems. The linear regression model for explaining data $\mathbf{y}$ assumes that the means of the observations are a linear combination of $q(\geq 0)$ measured variables, or covariates, that is,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_q)$ is the *full* design matrix, $\boldsymbol{\beta} \in \mathbb{R}^q$ are the parameter coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ are independent and identically distributed Gaussian variates $\varepsilon_i \sim N(0, \tau)$. Model selection in the linear regression context arises because it is common to assume that only a subset of the covariates is associated with the observations; that is, only a subset of the parameter vector is non-zero. Let $\gamma \subseteq \{1, \ldots, q\}$ denote an index set determining which of the covariates comprise the design submatrix $\mathbf{X}_\gamma$; the set of all candidate subsets is denoted by Γ. The linear model indexed by γ is then

$$\mathbf{y} = \mathbf{X}_\gamma\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

where $\boldsymbol{\beta}$ is the parameter vector of size $|\gamma|$ and

$$\mathbf{X}_\gamma = (\mathbf{x}_{\gamma_1}, \ldots, \mathbf{x}_{\gamma_k}).$$

The total number of unknown parameters to be estimated for model γ, including the noise variance parameter, is $k = |\gamma| + 1$. The negative log-likelihood for the data $\mathbf{y}$ is

$$-\log p(\mathbf{y}|\mathbf{X}_\gamma, \boldsymbol{\beta}, \tau; \gamma) = \frac{n}{2}\log 2\pi + \frac{n}{2}\log\tau + \frac{1}{2\tau}(\mathbf{y} - \mathbf{X}_\gamma\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}_\gamma\boldsymbol{\beta}). \tag{4}$$

The maximum likelihood estimates for the coefficient $\boldsymbol{\beta}$ and the noise variance τ are

$$\hat{\boldsymbol{\beta}}(\mathbf{y}; \gamma) = \left(\mathbf{X}'_\gamma\mathbf{X}_\gamma\right)^{-1}\mathbf{X}'_\gamma\mathbf{y}, \tag{5}$$

$$\hat{\tau}(\mathbf{y}; \gamma) = \frac{1}{n}(\mathbf{y} - \mathbf{X}_\gamma\hat{\boldsymbol{\beta}}(\mathbf{y}; \gamma))'(\mathbf{y} - \mathbf{X}_\gamma\hat{\boldsymbol{\beta}}(\mathbf{y}; \gamma)), \tag{6}$$

which coincide with the estimates obtained by minimising the sum of squared errors (the least squares estimates). The negative log-likelihood evaluated at the maximum likelihood estimates is given by

$$-\log p(\mathbf{y}|\mathbf{X}_\gamma, \hat{\boldsymbol{\beta}}(\mathbf{y};\gamma), \hat{\tau}(\mathbf{y};\gamma);\gamma) = \frac{n}{2}\log 2\pi + \frac{n}{2}\log \hat{\tau}(\mathbf{y};\gamma) + \frac{n}{2}. \tag{7}$$

The structure of the set Γ in the context of the linear regression model depends on the nature and interpretation of the covariates. In the case that the covariates are polynomials of increasing order, or sinusoids of increasing frequency, it is often convenient to assume a nested structure on Γ. For example, let us assume that the $q = 4$ covariates are polynomials such that $\mathbf{x}_1$ is the zero-order term (constant), $\mathbf{x}_2$ the linear term, $\mathbf{x}_3$ the quadratic term and $\mathbf{x}_4$ the cubic term. To use a model selection criterion to select an appropriate order of polynomial, we can form the set of candidate models Γ as

$$\Gamma = \{\Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4\},$$

where

$$\Gamma_1 = \{1\},\ \Gamma_2 = \{1,2\},\ \Gamma_3 = \{1,2,3\},\ \Gamma_4 = \{1,2,3,4\}.$$

From this structure it is obvious that there is only one model with k free parameters, for all k, and that models with more parameters can exactly represent all models with less parameters by setting the appropriate coefficients to zero. In contrast, consider the situation in which the covariates have no natural ordering, or that we do not wish to impose an ordering; for example, in the case of polynomial regression, we may wish to determine which, if any, of the individual polynomial terms are associated with the observations. In this case there is no clear way of imposing a nested structure, as we cannot *a priori* rule out any particular combination of covariates being important. This is usually called the *all-subsets* regression problem, and the candidate model set then has the following structure

$$\begin{aligned}
\Gamma_0 &= \{\emptyset\},\\
\Gamma_1 &= \{\{1\},\ \{2\},\ \{3\},\ \{4\}\},\\
\Gamma_2 &= \{\{1,2\},\ \{1,3\},\ \{1,4\},\ \{2,3\},\ \{2,4\},\ \{3,4\}\},\\
\Gamma_3 &= \{\{1,2,3\},\ \{1,2,4\},\ \{1,3,4\},\ \{2,3,4\}\},\\
\Gamma_4 &= \{1,2,3,4\},
\end{aligned}$$

where $\emptyset$ denotes the empty set, that is, none of the covariates are to be used. It is clear that models with more parameters can not necessarily represent all models with less parameters; for example, the model $\{1,3\}$ cannot represent the model $\{2\}$ as it is lacking covariate $\mathbf{x}_2$. Further, there is now no longer just a single model with k free parameters; in the above example, there are four models with one free parameter, and six models with two free parameters. In general, if there are q covariates in total, the number of models with k free parameters is given by $\binom{q}{k}$, which may be substantially greater than one for moderate q.

Throughout the remainder of this chapter, the linear model example will be used to demonstrate the practical application of the model selection criteria that will be discussed.

0.2 Minimum Distance Estimation Criteria

Intuitively, model selection criteria can be derived by quantifying "how close" the probability density of the fitted model is to the unknown, generating distribution. Many measures of distance between two probability distributions exist in the literature. Due to several theoretical properties, the Kullback-Leibler divergence [2], and its variants, is perhaps the most frequently used information theoretic measure of distance.

As the distribution $p^*(\cdot)$ that generated the observations $\mathbf{y}$ is unknown, the basic idea underlying the minimum distance estimation criteria is to construct an *estimate* of how close the fitted distributions are to the truth. In general, these estimators are constructed to satisfy the property of unbiasedness, at least for large sample sizes n. In particular, the distance-based methods we examine are the well known Akaike information criterion (AIC) (see Section 0.2.1) and symmetric Kullback information criterion (KIC) (see Section 0.2.2), as well as several corrected variants that improve the performance of these criteria when the sample size is small.

0.2.1 The Akaike Information Criterion

Recall the model selection problem as introduced in Section 0.1. Suppose a collection of n observations $\mathbf{y} = (y_1, \ldots, y_n)'$ has been sampled from an unknown distribution $P^*(\mathbf{y})$ having density function $p^*(\mathbf{y})$. Estimation of $p^*(\mathbf{y})$ is done within a set of nested candidate models $\Gamma \subseteq \mathbb{N}$, characterized by probability densities $p(\mathbf{y}|\boldsymbol{\theta};\gamma)$, $\gamma \in \Gamma$, where $\boldsymbol{\theta} \in \Theta_\gamma \subseteq \mathbb{R}^k$ and $k = \dim(\Theta_\gamma)$ is the number of free parameters possessed by model γ. Let $\hat{\boldsymbol{\theta}}(\mathbf{y};\gamma)$ denote the vector of estimated parameters obtained by maximizing the likelihood $p(\mathbf{y}|\boldsymbol{\theta};\gamma)$ over Θ_γ, that is,

$$\hat{\boldsymbol{\theta}}(\mathbf{y};\gamma) = \arg\max_{\boldsymbol{\theta}\in\Theta_\gamma} \{p(\mathbf{y}|\boldsymbol{\theta};\gamma)\}$$

and let $p(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y};\gamma);\gamma)$ denote the corresponding fitted model. To simplify notation, we introduce the shorthand notation $\hat{\boldsymbol{\theta}}_\gamma \equiv \hat{\boldsymbol{\theta}}(\mathbf{y};\gamma)$ to denote the maximum likelihood estimator, and $p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) \equiv p(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y};\gamma);\gamma)$ to denote the maximised likelihood for model γ.

To determine which candidate density model best approximates the true unknown model $p^*(\mathbf{y})$, we require a measure which provides a suitable reflection of the separation between $p^*(\mathbf{y})$ and an approximating model $p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma)$. The Kullback–Leibler divergence, also known as the cross-entropy, is one such measure. For the two probability densities $p^* \equiv p^*(\mathbf{x})$ and $p_{\boldsymbol{\theta}_\gamma} \equiv p(\mathbf{x}|\boldsymbol{\theta};\gamma)$, the Kullback–Leibler divergence, $\Delta_n(\cdot)$, between $p^*(\mathbf{x})$ and $p(\mathbf{x}|\boldsymbol{\theta};\gamma)$ with respect to $p^*(\mathbf{x})$ is defined as

$$\begin{aligned} 2\Delta_n(p^*, p_{\boldsymbol{\theta}_\gamma}) &= \mathrm{E}_{\mathbf{x}\sim p^*}\left[2\log\frac{p^*(\mathbf{x})}{p(\mathbf{x}|\boldsymbol{\theta};\gamma)}\right] \\ &= \mathrm{E}_{\mathbf{x}\sim p^*}\left[-2\log p(\mathbf{x}|\boldsymbol{\theta};\gamma)\right] - \mathrm{E}_{\mathbf{x}\sim p^*}\left[-2\log p^*(\mathbf{x})\right] \\ &= d_n(p^*, p_{\boldsymbol{\theta}_\gamma}) - d_n(p^*, p^*) \end{aligned}$$

where

$$d_n(p^*, p_{\boldsymbol{\theta}_\gamma}) = \mathrm{E}_{\mathbf{x}\sim p^*}\left[-2\log p(\mathbf{x}|\boldsymbol{\theta};\gamma)\right], \tag{8}$$

and $\mathrm{E}_{\mathbf{x}\sim p^*}[\cdot]$ represents the expectation with respect to $p^*(\mathbf{x})$, that is, $\mathbf{x} \sim p^*(\cdot)$. Ideally, the selection of the best model from the set of candidate models Γ would be done by choosing the model with the minimum KL divergence from the data generating distribution $p^*(\cdot)$, that is

$$\hat{\gamma} = \arg\min_{\gamma\in\Gamma}\left\{\Delta_n\left(p^*, p_{\hat{\boldsymbol{\theta}}_\gamma}\right)\right\},$$

where $\hat{\boldsymbol{\theta}}_\gamma$ is the maximum likelihood estimator of the parameters of model γ based on the observed data $\mathbf{y}$. Since $d_n(p^*, p^*)$ does not depend on the fitted model $\hat{\boldsymbol{\theta}}_\gamma$, any ranking of the candidate models according to $\Delta_n(p^*, p_{\hat{\boldsymbol{\theta}}_\gamma})$ is equivalent to ranking the models according to $d_n(p^*, p_{\hat{\boldsymbol{\theta}}_\gamma})$. Unfortunately, evaluating $d_n(p^*, p_\gamma)$ is not possible since doing so requires the knowledge of $p^*(\cdot)$ which is the aim of model selection in the first place.

As noted by Takeuchi [**?**], twice the negative maximised log-likelihood, $-2\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma)$, acts as a downwardly biased estimator of $d_n(p^*, p_{\hat{\theta}_\gamma})$ and is therefore not suitable for model comparison by itself. An unbiased estimate of the KL divergence is clearly of interest and could be used as a tool for model selection. It can be shown that the bias in estimation is given by

$$\mathrm{E}_{\mathbf{y}\sim p^*}\left[\mathrm{E}_{\mathbf{x}\sim p^*}\left[-2\log p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{y};\gamma);\gamma)\right]\right] - \mathrm{E}_{\mathbf{y}\sim p^*}\left[-2\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y};\gamma);\gamma)\right], \tag{9}$$

which, under suitable regularity conditions can be asymptotically estimated by [**?** 3]

$$2\,\mathrm{Tr}(\boldsymbol{\Omega}^{-1}(\boldsymbol{\theta}_0;\gamma)\,\boldsymbol{\Sigma}(\boldsymbol{\theta}_0;\gamma)). \tag{10}$$

where

$$\boldsymbol{\Omega}(\boldsymbol{\theta}_0;\gamma) = -\mathrm{E}_{\mathbf{x}\sim p^*}\left[\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\theta};\gamma)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \tag{11}$$

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}_0;\gamma) = \mathrm{E}_{\mathbf{x}\sim p^*}\left[\left(\frac{\partial \log p(\mathbf{x}|\boldsymbol{\theta};\gamma)}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \log p(\mathbf{x}|\boldsymbol{\theta};\gamma)}{\partial\boldsymbol{\theta}}\right)'\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \tag{12}$$

and $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. The parameter vector $\boldsymbol{\theta}_0$ defined by

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta_\gamma}\{\Delta_n(p^*, p_{\boldsymbol{\theta}})\} \tag{13}$$

indexes the distribution in the model γ that is closest to the data generating distribution $p^*(\cdot)$ in terms of KL divergence. Unfortunately, the parameter vector $\boldsymbol{\theta}_0$ is unknown and so one cannot compute the required bias adjustment. However, Takeuchi [**?**] noted that under suitable regularity conditions, the maximum likelihood estimate can be used in place of $\boldsymbol{\theta}_0$ to derive an approximation to (10), leading to the Takeuchi Information Criterion (TIC)

$$\mathrm{TIC}(\gamma;\mathbf{y}) = -2\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) + 2\,\mathrm{Tr}(\boldsymbol{\Omega}^{-1}(\mathbf{y};\gamma)\,\boldsymbol{\Sigma}(\mathbf{y};\gamma)), \tag{14}$$

where

$$\boldsymbol{\Omega}(\mathbf{y};\gamma) = -\frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta};\gamma)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\gamma}, \tag{15}$$

$$\boldsymbol{\Sigma}(\mathbf{y};\gamma) = \sum_{i=1}^{n}\left(\frac{\partial \log p(y_i|\boldsymbol{\theta};\gamma)}{\partial\boldsymbol{\theta}}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\gamma}\right)\left(\frac{\partial \log p(y_i|\boldsymbol{\theta};\gamma)}{\partial\boldsymbol{\theta}}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\gamma}\right)'. \tag{16}$$

are empirical estimates of the matrices (11) and (12), respectively. To use the TIC for model selection, we choose the model $\gamma \in \Gamma$ with the smallest TIC score, that is

$$\hat{\gamma}_{\mathrm{TIC}}(\mathbf{y}) = \arg\min_{\gamma\in\Gamma}\{\mathrm{TIC}(\gamma;\mathbf{y})\}.$$

The model with the smallest TIC score corresponds to the model we believe to be the closest in terms of KL divergence to the data generating distribution $p^*(\cdot)$. Under suitable regularity conditions, the TIC is an unbiased estimate of the KL divergence between the fitted model $p_{\hat{\theta}_\gamma}$ and the data generating model $p^*(\cdot)$, that is

$$\mathrm{E}_{\mathbf{y}\sim p^*}\left[\mathrm{TIC}(\gamma;\mathbf{y})\right] = \mathrm{E}_{\mathbf{y}\sim p^*}\left[d_n(p^*, p_{\hat{\theta}_\gamma})\right] + o(1).$$

where $o(1)$ is a quantity that vanishes as $n \to \infty$. The empirical matrices (15) and (16) are a good approximation to (11) and (12) only if the sample size n is very large. As this is often not the case in practice, the TIC is rarely used.

The dependence of the TIC on the empirical matrices can be avoided if one assumes that the data generating distribution $p^*(\cdot)$ is a distribution contained in the model γ. In this case, the the matrices (11) and (12) coincide. Thus, $\text{Tr}(\boldsymbol{\Omega}^{-1}(\boldsymbol{\theta}_0;\gamma)\,\boldsymbol{\Sigma}(\boldsymbol{\theta}_0;\gamma))$ reduces to the number of parameters k, and we obtain the widely known and extensively used Akaike's Information Criterion (AIC) [4]

$$\text{AIC}(\gamma;\mathbf{y}) = -2\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) + 2k. \tag{17}$$

An interesting way to view the AIC (17) is as a penalised likelihood criterion. For a given model γ, the negative maximised log-likelihood measures how well the model fits the data, while the "$2k$" penalty measures the complexity of the model. Clearly, the complexity is an increasing function of the number of free parameters, and this acts to naturally balance the complexity of a model against its fit. Practically, this means that a complex model must fit the data well to be preferred to a simpler model with a similar quality of fit.

Similar to the TIC, the AIC is an asymptotically unbiased estimator of the KL divergence between the generating distribution $p^*(\cdot)$ and the fitted distribution $p_{\hat{\boldsymbol{\theta}}_\gamma}$, that is

$$\text{E}_{\mathbf{y}\sim p^*}\left[\text{AIC}(\gamma;\mathbf{y})\right] = \text{E}_{\mathbf{y}\sim p^*}\left[d_n(p^*, p_{\hat{\boldsymbol{\theta}}_\gamma})\right] + o(1).$$

When the sample size is large and the number of free parameters is small relative to the sample size, the $o(1)$ term is approximately zero, and the AIC offers an excellent estimate of the Kullback–Leibler divergence. However, in the case that n is small, or k is large relative to n, the $o(1)$ term may be quite large, and the AIC does not adequately correct the bias, making it unsuitable for model selection.

To address this issue, Hurvich and Tsai [5] proposed a small sample correction to the AIC in the linear regression setting. This small sample AIC, referred to as AIC_c, is given by

$$\text{AIC}_c(\gamma;\mathbf{y}) = -2\log p(\mathbf{y},\hat{\boldsymbol{\theta}}_\gamma) + \frac{2kn}{n-k-1}. \tag{18}$$

for a model γ with k free parameters, and has subsequently been applied to models other than linear regressions. From (18) it is clear that the penalty term is substantially larger than $2k$ when the sample size n is not significantly larger than k, and that as $n \to \infty$ the AIC_c is equivalent to the regular AIC (17). Practically, a large body of empirical evidence strongly suggests that the AIC_c performs significantly better as a model selection tool than the AIC, largely irrespective of the problem. The approach taken by Hurvich and Tsai to derive their AIC_c criterion is not the only way of deriving small sample minimum distance estimators based on the KL divergence; for example, the work in [6] derives alternative AIC-like criteria based on variants of the KL divergence, while [7] offers an alternative small sample criterion obtained through bootstrap approximations.

For the reader interested in the theoretical details and complete derivations of the AIC, including all regularity conditions, we highly recommend the excellent text by Linhart and Zucchini [3]. We also recommend the text by McQuarrie and Tsai [8] for a more practical treatment of AIC and AIC_c.

0.2.2 The Kullback Information Criterion

Since the Kullback-Leibler divergence is an asymmetric measure, an alternative directed divergence can be obtained by reversing the roles of the two models in the definition of the measure. A undirected measure of model dissimilarity can be obtained from the sum of the two directed divergences. This measure is known

as Kullback's symmetric divergence, or J-divergence [9]. Since the J-divergence combines information about model dissimilarity through two distinct measures, it functions as a gauge of model disparity which is arguably more sensitive than either of its individual components. The J-divergence, $J_n(\cdot)$, between the true generating distribution $p^*(\cdot)$ and a distribution $p_{\boldsymbol{\theta}_\gamma}$ is given by

$$\begin{aligned} 2\,J_n(p^*, p_{\boldsymbol{\theta}_\gamma}) &= 2\,\Delta_n(p^*, p_{\boldsymbol{\theta}_\gamma}) + 2\,\Delta_n(p_{\boldsymbol{\theta}_\gamma}, p^*) \\ &= d_n(p^*, p_{\boldsymbol{\theta}_\gamma}) - d_n(p^*, p^*) + d_n(p_{\boldsymbol{\theta}_\gamma}, p^*) - d_n(p_{\boldsymbol{\theta}_\gamma}, p_{\boldsymbol{\theta}_\gamma}) \end{aligned}$$

where

$$d_n(p_{\boldsymbol{\theta}_\gamma}, p_{\boldsymbol{\theta}_\gamma}) = \mathrm{E}_{\mathbf{x}\sim p_{\boldsymbol{\theta}_\gamma}}\left[-2\log p(\mathbf{x}|\boldsymbol{\theta};\gamma)\right], \tag{19}$$

$$d_n(p_{\boldsymbol{\theta}_\gamma}, p^*) = \mathrm{E}_{\mathbf{x}\sim p_{\boldsymbol{\theta}_\gamma}}\left[-2\log p^*(\mathbf{x})\right]. \tag{20}$$

As in Section (0.2.1), we note that the term $d_n(p^*, p^*)$ does not depend on the fitted model $\boldsymbol{\theta}_\gamma$ and may be dropped when ranking models. The quantity

$$K_n(p^*, p_\gamma) = d_n(p^*, p_{\boldsymbol{\theta}_\gamma}) + d_n(p_{\boldsymbol{\theta}_\gamma}, p^*) - d_n(p_{\boldsymbol{\theta}_\gamma}, p_{\boldsymbol{\theta}_\gamma})$$

may then be used as a surrogate for the J-divergence, and leads to an appealing measure of dissimilarity between the generating distribution and the fitted candidate model. In a similar fashion to the AIC, twice the negative maximised log-likelihood has been shown to be a downwardly biased estimate of the J-divergence. Cavanaugh [10] derives an estimate of this bias, and uses this correction to define a model selection criterion called KIC (symmetric Kullback information criterion), which is given by

$$\mathrm{KIC}(\gamma;\mathbf{y}) = -2\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) + 3k. \tag{21}$$

Comparing (21) to the AIC (17), we see that the KIC complexity penalty is slightly greater. This implies that the KIC tends to prefer simpler models (that is, those with less parameters) than those selected by minimising the AIC. Under suitable regularity conditions, the KIC satisfies

$$\mathrm{E}_{\mathbf{y}\sim p^*}\left[\mathrm{KIC}(\gamma;\mathbf{y})\right] = \mathrm{E}_{\mathbf{y}\sim p^*}\left[K_n(p^*, p_{\boldsymbol{\theta}_\gamma})\right] + o(1).$$

Empirically, the KIC has been shown to outperform AIC in large sample linear regression and autoregressive model selection, and tends to lead to less overfitting than AIC. Similarly to the AIC, when the sample size n is small the bias correction term $3k$ is too small, and the KIC performs poorly as a model selection tool. Seghouane and Bekara [11] derived a corrected KIC, called KIC_c, in the context of linear regression that is suitable for use when the sample size is small relative to the total number of parameters k. The KIC_c is

$$\mathrm{KIC}_c(\gamma;\mathbf{y}) = -2\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) + \frac{2kn}{n-k-1} - n\psi\left(\frac{n-k-1}{2}\right) + n\log\frac{n}{2}, \tag{22}$$

where $\psi(\cdot)$ denotes the digamma function. The digamma function can be difficult to compute, and approximating the digamma function by a second-order Taylor series expansion yields the so called AKIC_c, which is given by

$$\mathrm{AKIC}_c(\gamma;\mathbf{y}) = -2\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) + \frac{k(3n-k-1)}{n-k-1} + \frac{k-1}{n-k-1}. \tag{23}$$

The second term in (22) appears as the complexity penalty in the AIC_c (18), which is not surprising given that the J-divergence is a sum of two KL divergence functions. Empirically, the KIC_c has been shown to outperform the KIC as a model selection tool when the sample size is small.

0.2.3 Example Application: Linear Regression

We now demonstrate the application of the minimum distance estimation criteria to the problem of linear regression introduced in Section 0.1.2. Recall that in this setting, γ specifies which covariates from the full design matrix $\mathbf{X}$ should be used in the fitted model. The total number of parameters is therefore $k = |\gamma| + 1$, where the additional parameter accounts for the fact that we must estimate the noise variance. The various model selection criteria are listed below:

$$\begin{aligned}
\mathrm{AIC}(\gamma; \mathbf{y}) &= n \log\left(2\pi\hat{\tau}(\mathbf{y}; \gamma)\right) + n + 2k, && (24)\\
\mathrm{AIC}_c(\gamma; \mathbf{y}) &= n \log\left(2\pi\hat{\tau}(\mathbf{y}; \gamma)\right) + n + \frac{2kn}{n-k-1}, && (25)\\
\mathrm{KIC}(\gamma; \mathbf{y}) &= n \log\left(2\pi\hat{\tau}(\mathbf{y}; \gamma)\right) + n + 3k, && (26)\\
\mathrm{KIC}_c(\gamma; \mathbf{y}) &= n \log\left(2\pi\hat{\tau}(\mathbf{y}; \gamma)\right) + n + \frac{2kn}{n-k-1} - n\psi\left(\frac{n-k-1}{2}\right). && (27)
\end{aligned}$$

It is important to stress that all these criteria are derived within the context of nested set of candidate models Γ (see Section 0.1.1). In the next section we will briefly examine the problems that can arise if these criteria are used in situations where the candidates are not nested.

0.2.4 Consistency and Efficiency

Let γ^* denote the model that generated the observed data $\mathbf{y}$ and assume that γ^* is a member of the set of candidate models Γ which is fixed and does not grow with sample size n; recall that the model γ^* contains the data generating distribution $p^*(\cdot)$ as a particular distribution. Furthermore, of all the models that contain $p^*(\cdot)$, the model γ^* has the least number of free parameters. A model selection criterion is consistent if the probability of selecting the true model γ^* approaches one as the sample size $n \to \infty$. None of the distance based criteria examined in this chapter are consistent model selection procedures. This means that even for extremely large sample sizes, there is a non-zero probability that these distance based criteria will overfit and select a model with more free parameters than γ^*. Consequently, if the aim of the experiment is to discover the true model, the aforementioned distance based criteria are inappropriate. For example, in the linear regression setting, if the primary aim of the model selection exercise is to determine only those covariates that are associated with the observations, one should not use the AIC or KIC (and their variants).

In contrast, if the true model is of infinite order, which in many settings may be deemed more realistic in practice, then the true model lies outside the class of candidate models. In this case, we cannot hope to discover the true model, and instead would like our model selection criterion to at least provide good predictions about future data arising from the same source. A criterion that minimizes the one-step mean squared error between the fitted model and the generating distribution $p^*(\cdot)$ with high probability as $n \to \infty$ is termed *efficient*. The AIC and KIC, and their corrected variants are all asymptotically efficient [**?**]. This suggests that for large sample sizes, model selection by distance based criteria will lead to adequate prediction errors even if the true model lies outside the set of candidate models Γ.

0.2.5 The AIC and KIC for Non-Nested Sets of Candidate Models

The derivations of the AIC and KIC, and their variants, depend on the assumption that the candidate model set Γ is nested. We conclude this section by briefly examining the behaviour of these criteria when they are used to select a model from a non-nested candidate model set. First, consider the case that Γ forms a nested set of models, and let $\gamma^* \in \Gamma$ be the true model, that is, the model that generated the data; this means

that γ^* is the model in Γ with the least number of free parameters, say k^*, that contains the data generating distribution $p^*(\cdot)$ as a particular distribution. If n is sufficiently large, the probability that the model selected by minimising the AIC will have $k^* + 1$ parameters is at least 16%, with the equivalent probability for KIC being approximately 8% [8].

Consider now the case in which Γ forms a non-nested set of candidate models. A recent paper by Schmidt and Makalic [12] has demonstrated that in this case, the AIC is a downwardly biased estimator of the Kullback–Leibler divergence even asymptotically in n, and this downward bias is determined by the size of the sets Γ_k. This implies, that in the case of all-subsets regression, the probability of overfitting is an increasing function of the number of covariates under consideration, q, and this probability cannot be reduced even by increasing the sample size.

As an example of how great this probability may be, we consider an all-subsets regression in which the generating distribution is the null model; that is, none of the measured covariates $\mathbf{x}_i$ are associated with the observations $\mathbf{y}$. Even if there is only as few as $q = 10$ covariates available for selection, the probability of erroneously preferring a non-null model to the null model using the AIC is approximately 81%, and by the time $q = 50$, this probability rises to 99.9%. Although [12] examines only the AIC, similar arguments follow easily for the KIC. It is therefore recommended that while both criteria are appropriate for nested model selection, neither of these distance based criteria should be used for model selection in a non-nested situation, such as the all subsets regression problem.

0.2.6 Applications

Minimum distance estimation criteria have been widely applied in the literature, and we present below an inexhaustive list of successful applications:

- Univariate linear regression models [5, 11],
- Multivariate linear regression models with arbitrary noise covariance matrices [13, 14],
- Autoregressive model selection [5, 10],
- Selection of smoothing parameters in nonparametric regression [15],
- Signal denoising [16, 17],
- Selection of the size of a multilayer perceptron network [18].

For details of many of these applications the reader is referred to [8].

0.3 Bayesian Approaches to Model Selection

Section 0.2 has described model selection criteria that are designed to explicitly minimise the distance between the estimated models and the true model that generated the observed data $\mathbf{y}$. An alternative approach, based on Bayesian statistics [19], is now discussed. The Bayesian approach to statistical analysis differs from the "classical" approach through the introduction of a *prior distribution* that is used to quantify an experimenter's *a priori* (that is, before the observation of data) beliefs about the source of the observed data. A central quantity in Bayesian analysis is the so-called *posterior distribution*; given a prior distribution,

$\pi_\theta(\boldsymbol{\theta}|\gamma)$, over the parameter space $\boldsymbol{\theta} \in \Theta_\gamma$, and the likelihood function, $p(\mathbf{y}|\boldsymbol{\theta}, \gamma)$, the posterior distribution of the parameters given the data may be found by applying Bayes' theorem, yielding

$$p(\boldsymbol{\theta}|\mathbf{y}, \gamma) = \frac{p(\mathbf{y}|\boldsymbol{\theta}, \gamma)\pi_\theta(\boldsymbol{\theta}|\gamma)}{p(\mathbf{y}|\gamma)}, \tag{28}$$

where

$$p(\mathbf{y}|\gamma) = \int_{\Theta_\gamma} p(\mathbf{y}|\boldsymbol{\theta}, \gamma)\pi_\theta(\boldsymbol{\theta}|\gamma)d\boldsymbol{\theta} \tag{29}$$

is known as the marginal probability of $\mathbf{y}$ for model γ, and is a normalisation term required to render (28) a proper distribution. In general, the specification of the prior distribution may itself depend on further parameters, usually called *hyperparameters*, but for the purposes of the present discussion this possibility will not be considered. The main strength of the Bayesian paradigm is that it allows uncertainty about models and parameters to be defined directly in terms of probabilities; there is no need to appeal to more abstruse concepts, such as the "confidence interval" of classical statistics. However, there is in general no free lunch in statistics, and this strength comes at a price: the specification and origin of the prior distribution. There are essentially two main schools of thought on this topic: (i) subjective Bayesianism, where prior distributions are viewed as serious expressions of subjective belief about the process that generated the data, and (ii) objective Bayesianism [20], where one attempts to express prior ignorance through the use of uninformative, objective distributions, such as the Jeffreys' prior [9], reference priors [21] and intrinsic priors [22]. The specification of the prior distribution is the basis of most criticism leveled at the Bayesian school, and an extensive discussion on the merits and drawbacks of Bayesian procedures, and the relative merits of subjective and objective approaches, is beyond the scope of this tutorial (there are many interesting discussions on this topic in the literature, see for example, [23, 24]). However, in this section of the tutorial, an approach to the problem of prior distribution specification based on asymptotic arguments will be presented.

While (28) specifies a probability distribution over the parameter space Θ_γ, conditional on the observed data, it gives no indication of the likelihood of the model γ given the observed data. A common approach is based on the marginal distribution (29). If a further prior distribution, $\pi_\gamma(\gamma)$, over the set of candidate models $\gamma \in \Gamma$ is available, one may apply Bayes' theorem to form a posterior distribution over the models themselves, given by

$$p(\gamma|\mathbf{y}) = \frac{p(\mathbf{y}|\gamma)\pi_\gamma(\gamma)}{\sum_{\gamma\in\Gamma} p(\mathbf{y}|\gamma)\pi_\gamma(\gamma)}. \tag{30}$$

Model selection then proceeds by choosing the model $\hat{\gamma}(\mathbf{y})$ that maximises the probability (30), that is

$$\hat{\gamma}(\mathbf{y}) = \arg\max_{\gamma\in\Gamma} \left\{ p(\mathbf{y}|\gamma)\pi_\gamma(\gamma) \right\}, \tag{31}$$

where the normalising constant may be safely ignored. An interesting consequence of the posterior distribution (30) is that the posterior-odds in favour of some model, γ_1, over another model γ_0, usually called the Bayes factor [25, 26], can be found from the ratio of posterior probabilities, that is,

$$\mathrm{BF}(\gamma_1, \gamma_0) = \frac{p(\mathbf{y}|\gamma_1)\pi_\gamma(\gamma_1)}{p(\mathbf{y}|\gamma_0)\pi_\gamma(\gamma_0)}. \tag{32}$$

This allows for a straightforward and highly interpretable quantification of the uncertainty in the choice of any particular model. The most obvious weakness in the Bayesian approach, ignoring the origin of the prior distributions, is computational. As a general rule, the integral in the definition of the marginal distribution (29), does not admit a closed-form solution and one must resort to numerical approaches or

approximations. The next section will discuss a criterion based on asymptotic arguments that circumvents both the requirement to specify an appropriate prior distribution as well as the issue of integration in the computation of the marginal distribution.

0.3.1 The Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC), also sometimes known as the Schwarz Information Criterion (SIC), was proposed by G. Schwarz in 1978 [27]. However, the resulting criterion is not unique and was also derived from information theoretic considerations by J. Rissanen in 1978 [28], as well as being implicit in the earlier work of C. Wallace and D. Boulton [29]; the information theoretic arguments for model selection are discussed in detail in the next section of this chapter. The BIC is based on the Laplace approximation for high dimensional integration [30]. Essentially, under certain regularity conditions, the Laplace approximation works by replacing the distribution to be integrated by a suitable multivariate normal distribution, which results in a straightforward integral. Making the assumptions that the prior distribution is such that its effects are "swamped" by the evidence in the data, and that the maximum likelihood estimator converges on the posterior mode as $n \to \infty$, one may apply the Laplace approximation to (29) yielding

$$\pi_\gamma(\gamma) \int_{\Theta_\gamma} p(\mathbf{y}|\boldsymbol{\theta}, \gamma)\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\gamma)d\boldsymbol{\theta} \simeq \frac{(2\pi)^{k/2} p(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y};\gamma),\gamma)\pi_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_\gamma(\mathbf{y};\gamma)|\gamma)\pi_\gamma(\gamma)}{|\mathbf{J}_n(\hat{\boldsymbol{\theta}}(\mathbf{y};\gamma);\gamma)|^{1/2}}, \tag{33}$$

where $\simeq$ denotes that the ratio of the left and right hand side approaches one as the sample size $n \to \infty$. In (33), $k = \dim(\Theta_\gamma)$ is the total number of continuous, free parameters for model γ, $\hat{\boldsymbol{\theta}}_\gamma(\mathbf{y};\gamma)$ is the maximum likelihood estimate, and $\mathbf{J}(\cdot)$ is the $(k \times k)$ expected Fisher information matrix for n data points, given by

$$\mathbf{J}_n(\boldsymbol{\theta}_0;\gamma) = -\mathrm{E}_{\boldsymbol{\theta}_0}\left[\left.\frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta},\gamma)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]. \tag{34}$$

The technical conditions under which (33) holds are detailed in [27]; in general, if the maximum likelihood estimates are consistent, that is, they converge on the true, generating value of $\boldsymbol{\theta}$ as $n \to \infty$, and also satisfy the central limit theorem, the approximation will be satisfactory, at least for large sample sizes. To find the BIC, begin by taking negative logarithms of the right hand side of (33)

$$-\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y};\gamma),\gamma) + \frac{1}{2}\log|\mathbf{J}_n(\hat{\boldsymbol{\theta}}(\mathbf{y};\gamma);\gamma)| - \log\pi_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}(\mathbf{y};\gamma)|\gamma) - \frac{k}{2}\log 2\pi - \log\pi_\gamma(\gamma). \tag{35}$$

A crucial assumption made is that the Fisher information satisfies

$$\mathbf{J}_1(\boldsymbol{\theta};\gamma) = \lim_{n\to\infty}\left\{\frac{\mathbf{J}_n(\boldsymbol{\theta};\gamma)}{n}\right\}, \tag{36}$$

where $\mathbf{J}_1(\cdot)$ is the asymptotic *per sample* Fisher information matrix satisfying $|\mathbf{J}_1(\cdot)| > 0$, and is free of dependency on n. This assumption is met by a large range of models, including linear regression models, autoregressive moving-average (ARMA) models and generalised linear models (GLMs) to name a few. This allows the determinant of the Fisher information matrix for n samples to be rewritten as

$$\begin{aligned} |\mathbf{J}_n(\boldsymbol{\theta};\gamma)| &= n^k \cdot |\mathbf{J}_1(\boldsymbol{\theta};\gamma)|, \\ &= n^k \cdot O(1), \end{aligned}$$

where $O(1)$ denotes a quantity that is constant in n. Using this result, (35) may be dramatically simplified by assuming that n is large and discarding terms that are $O(1)$ with respect to the sample size, yielding the BIC

$$\text{BIC}(\gamma; \mathbf{y}) = -\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y}; \gamma), \gamma) + \frac{k}{2} \log n. \tag{37}$$

Model selection using BIC is then done by finding the $\gamma \in \Gamma$ that minimises the BIC score (37), that is,

$$\hat{\gamma}_{\text{BIC}}(\mathbf{y}) = \arg\min_{\gamma \in \Gamma} \{\text{BIC}(\gamma; \mathbf{y})\}. \tag{38}$$

It is immediately obvious from (37) that a happy consequence of the approximations that are employed is the removal of the dependence on the prior distribution $\pi_{\boldsymbol{\theta}}(\cdot)$. However, as usual, this comes at a price. The BIC satisfies

$$-\log \int_{\Theta_\gamma} p(\mathbf{y}|\boldsymbol{\theta}, \gamma) \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\gamma) d\boldsymbol{\theta} = \text{BIC}(\gamma; \mathbf{y}) + O(1), \tag{39}$$

where the $O(1)$ term is only guaranteed to be negligible for large n. For any finite n, this term may be arbitrarily large, and may have considerable effect on the behaviour of BIC as a model selection tool. Thus, in general, BIC is only appropriate for use when the sample size is quite large relative to the number of parameters k. However, despite this drawback, when used correctly the BIC has several pleasant, and important, properties that are now discussed. The important point to bear in mind is that like all approximations, it can only be employed with confidence in situations that meet the conditions under which the approximation was derived.

0.3.2 Properties of BIC

Consistency of BIC

A particularly important, and defining, property of BIC is the consistency of $\hat{\gamma}_{\text{BIC}}(\mathbf{y})$ as $n \to \infty$. Let $\gamma^* \in \Gamma$ be the true model, that is, the model that generated the data; this means that γ^* is the model in Γ with the least number of free parameters that contains the data generating distribution $p^*(\cdot)$ as a particular distribution. Further, let the set of all models Γ be independent of the sample size n; this means that the number of candidate models that are being considered does not increase with increasing sample size. Then, if all models in Γ satisfy the regularity conditions under which BIC was derived, and in particular (36), the BIC estimate satisfies

$$\Pr(\hat{\gamma}_{\text{BIC}}(\mathbf{y}) = \gamma^*) \to 1 \tag{40}$$

as $n \to \infty$ [31, 32]. In words, this says that the BIC estimate of γ converges on the true, generating model with probability one in the limit as the sample size $n \to \infty$, and this property holds irrespective of whether Γ forms a nested or non-nested set of candidate models. Practically, this means that as the sample size grows the probability that the model selected by the minimising the BIC score overfits or underfits the true, generating model, tends to zero. This is an important property if the discovery of relevant parameters is of more interest than making good predictions, and is not shared by any of the distance-based criteria discussed in Section 0.2. An argument against the importance of model selection consistency stems from the fact that the statistical models under consideration are abstractions of reality, and that the true, generating model is not be a member of the set Γ. Regardless of this criticism, empirical studies suggest that the BIC tends to perform well in comparison to asymptotically efficient criteria such as AIC and KIC when the underlying generating process may be described by a small number of strong effects [8], and thus occupies a useful niche in the gallery of model selection techniques.

Bayesian Interpretations of BIC

Due to the fact that the BIC is an approximation of the marginal probability of data $\mathbf{y}$ (29), it admits the full range of Bayesian interpretations. For example, by computing the BIC scores for two models γ_0 and γ_1, one may compute an approximate negative log-Bayes factor of the form

$$-\log \mathrm{BF}(\gamma_1, \gamma_0) \approx \mathrm{BIC}(\gamma_0; \mathbf{y}) - \mathrm{BIC}(\gamma_1; \mathbf{y}).$$

Therefore, the exponential of the negative difference in BIC scores for γ_1 and γ_0 may be interpreted as an approximate posterior-odds in favour of γ_1. Further, the BIC scores may be used to perform model averaging. Selection of a single best model from a candidate set is known to be statistically unstable [33], particularly if the sample size is small and many models have similar criterion scores. Instability in this setting is defined as the variability in the choice of the single best model under minor perturbations of the observed sample. It is clear that if many models are given similar criterion scores, then changing a single data point in the sample can lead to significant changes in the ranking of the models. In contrast to model selection, model averaging aims to improve predictions by using a weighted mixture of all the estimated models, the weights being proportional to the strength of evidence for the particular model. By noting that the BIC score is approximately equal to the negative log-posterior probability of the model, a Bayesian predictive density for future data $\mathbf{y}_1$, conditional on an observed sample $\mathbf{y}_0$, may be found by marginalising out the model class indicator γ, that is,

$$p(\mathbf{y}_1|\mathbf{y}_0) \approx \frac{\sum_{\gamma\in\Gamma} p(\mathbf{y}_1|\hat{\boldsymbol{\theta}}(\mathbf{y}_0;\gamma),\gamma)e^{-\mathrm{BIC}(\gamma;\mathbf{y}_0)}}{\sum_{\gamma\in\Gamma} e^{-\mathrm{BIC}(\gamma;\mathbf{y}_0)}}. \tag{41}$$

Although such a mixture often outperforms a single best model when used to make predictions about future data [34], there are several drawbacks to the model averaging approach. The first is that the resulting density lacks the property of parsimonity, as no model is excluded from the mixture, and therefore all model parameters contribute to the mixture. In the case of linear regression, for example, this means the ability to decide whether particular covariates are "relevant" is lost. The second drawback is that the predictive density (41) is in general not of the same form as any its component models, which can lead to interpretability issues.

0.3.3 Example Application: Linear Regression

We conclude our examination of the BIC with an example of an application to linear regression models. Recalling that in this context the model indicator $\gamma \subseteq \{1, \ldots, q\}$ specifies which covariates, if any, should be used from the full design matrix $\mathbf{X} \in \mathbb{R}^{(n\times q)}$ to explain the observed samples $\mathbf{y}$. The BIC score for a particular covariate set $\gamma \in \Gamma$ is given by

$$\mathrm{BIC}(\gamma; \mathbf{y}) = \frac{n}{2}\log\left(2\pi\hat{\tau}(\mathbf{y};\gamma)\right) + \frac{n}{2} + \left(\frac{k+1}{2}\right)\log n, \tag{42}$$

where $\hat{\tau}(\cdot;\gamma)$ is the maximum likelihood estimator of the noise variance for model γ, and the $(k+1)$ accounts for the fact that the variance must be estimated from the data along with the k regression parameters. The BIC score (42) contains the terms $(1/2)(n + \log n)$ which are constant with respect to γ and may be omitted if the set of candidate models only consists of linear regressions; if the set Γ also contains models from alternative classes, such as artificial neural networks, then these terms are required to fully characterise the marginal probability in comparison to this alternative class of models. The consistency property of BIC is particularly useful in this setting, as it guarantees that if the data were generated from (42) then as $n \to \infty$, minimising the BIC score will recover the true model, and thus determine exactly which covariates are associated with $\mathbf{y}$.

Given the assumption of Gaussian noise, the Bayesian mixture distribution is a weighted mixture of Gaussian distributions and thus has a particularly simple conditional mean. Let $\boldsymbol{\alpha}(\gamma)$ denote a $(q \times 1)$ vector with entries

$$\alpha_i(\gamma) = \begin{cases} \hat{\boldsymbol{\beta}}_i(\mathbf{y}; \gamma) & i \in \gamma \\ 0 & \text{otherwise} \end{cases},$$

that is, the vector $\boldsymbol{\alpha}(\gamma)$ contains the maximum likelihood estimates for the covariates in γ, and zeros for those covariates that are not in γ. Then, the conditional mean of the predictive density for future data with design matrix $\mathbf{X}_1$, conditional on an observed sample $\mathbf{y}_0$, is simply a linear combination of the $\boldsymbol{\alpha}(\gamma)$

$$\mathrm{E}\left[\mathbf{y}_1 | \mathbf{X}_1, \mathbf{y}_0\right] = \mathbf{X}_1 \left(\sum_{\gamma \in \Gamma} \boldsymbol{\alpha}(\gamma) e^{-\mathrm{BIC}(\gamma; \mathbf{y}_0)} \right),$$

which is essentially a linear regression with a special coefficient vector. While this coefficient vector will be dense, in the sense that none of its entries will be exactly zero, it retains the high degree of interpretibility that makes linear regression models so attractive.

0.3.4 Priors for the model structure γ

One of the primary assumptions made in the derivation of the BIC is that the effects of the prior distribution for the model, $\pi_\gamma(\cdot)$, can be safely neglected. However, in some cases the number of models contained in Γ is very large relative to the sample size, or may even grow with growing sample size, and this assumption may no longer be valid. An important example is that of regression models in the context of genetic datasets; in this case, the number of covariates is generally several orders of magnitude greater than the number of samples. In this case, it is possible to use a modified BIC of the form

$$\mathrm{BIC}(\gamma; \mathbf{y}) = -\log p(\mathbf{y} | \hat{\boldsymbol{\theta}}(\mathbf{y}; \gamma), \gamma) + \frac{k}{2} \log n - \log \pi_\gamma(\gamma), \tag{43}$$

which requires specification of the prior distribution $\pi_\gamma(\cdot)$. In the setting of regression models, there are two general flavours of prior distributions for γ depending on the structure represented by Γ. The first case is that of *nested* model selection. A nested model class has the important property that all models with k parameters contain all models with less than k parameters as special cases. A standard example of this is polynomial regression in which the model selection criterion must choose the order of the polynomial. A uniform prior over Γ is an appropriate, and standard, choice of prior for nested model classes. Use of such a prior in (43) inflates the BIC score by a term of $\log |\Gamma|$; this term is constant for all γ, and may consequently be ignored when using the BIC scores to rank the models.

In contrast, the second general case, known as *all-subsets regression*, is less straightforward. In this setting, there are q covariates, and Γ contains all possible subsets of $\{1, \ldots, q\}$. This implies that the model class is no longer nested. It is tempting to assume a uniform prior over the members of Γ, as this would appear to be an uninformative choice. Unfortunately, such a prior is heavily biased towards subsets containing approximately half of the covariates. To see this, note that Γ contains a total of $\binom{q}{k}$ subsets that include k covariates. This number may be be extremely large when k is close to $q/2$, even for moderate q, and thus these subsets will be given a large proportion of the prior probability. For example, if $q = 20$ then $|\Gamma| \approx 10^6$ and $\binom{20}{10} \approx 1.8 \times 10^5$. Practically, this means that subsets with $k = 10$ covariates are given approximately one-fifth of the total prior probability; in contrast, subsets with a single covariate receive less than one percent

of the prior probability. To address this issue, an alternative approach is to use a prior which assigns equal prior probability to each set of subsets of k covariates, for all k, that is,

$$-\log \pi_\gamma(\gamma) = \log \binom{q}{|\gamma|} + \log(q+1). \tag{44}$$

This prior (or ones very similar) also arise from both empirical Bayes [35] and information theoretic arguments [23, 36], and have been extensively studied by J. Scott and J. Berger [37]. This work has demonstrated that priors of the form (44) help Bayesian procedures overcome the difficulties associated with all-subsets model selection that adversely affect distance based criteria such as AIC and KIC [12].

In fact, J. Chen and Z. Chen [38] have proposed a generalization of this prior as part of what they call the "extended BIC". An important (specific) result of this work is the proof that using the prior (44) relaxes conditions required for the consistency of the BIC. Essentially, the total number of covariates may now be allowed to depend on the sample size n in the sense that $q = O(n^\kappa)$, where κ is a constant satisfying $0 < \kappa < \infty$, implying that $|\Gamma|$ grows with increasing n. Then, under certain identifiability conditions on the complete design matrix $\mathbf{X}$ detailed in [38], and assuming that $\gamma^* \in \Gamma$, the BIC given by (43) using the prior (44) almost surely selects the true model γ^* as $n \to \infty$.

0.3.5 Markov-Chain Monte-Carlo Bayesian Methods

We conclude this section by briefly discussing several alternative approaches to Bayesian model selection based on random sampling from the posterior distribution, $p(\boldsymbol{\theta}|\mathbf{y})$, which fall under the general umbrella of Markov Chain Monte Carlo (MCMC) methods [39]. The basic idea is to draw a sample of parameter values from the posterior distribution, using a technique such as the Metropolis–Hastings algorithm [40, 41] or the Gibbs sampler [42, 43]. These techniques are in general substantially more complex than the information criteria based approaches, both computationally and in terms of implementation, and this complexity generally brings with it greater flexibility in the specification of prior distributions as well as less stringent operating assumptions.

The most obvious way to apply MCMC methods to Bayesian model selection is through direct evaluation of the marginal (29). Unfortunately, this is a very difficult problem in general, and the most obvious approach, the so called harmonic mean estimator, suffers from statistical instability and convergence problems and should be avoided. In the particular case that Gibbs sampling is possible, and that the posterior is unimodal, Chib [44] has provided an algorithm to compute the approximate marginal probability from posterior samples.

An alternative to attempting to directly compute the marginal probability is to view the model indicator γ as the parameter of interest and randomly sample from the posterior $p(\gamma|\mathbf{y})$. This allows for the space of models Γ to be explored by simulation, and approximate posterior probabilities for each of the models to be determined by the frequency at which a particular model appears in the sample. There have been a large number of papers published that discuss this type of approach, and important contributions include the reversible jump MCMC algorithm of Green [45], the stochastic search variable selection algorithm [46], the shotgun stochastic search algorithm [47], and an interesting paper from Casella and Moreno [24] that combines objective Bayesian priors with a stochastic search procedure for regression models.

Finally, there has been a large amount of recent interest in using regularisation and "sparsity inducing" priors to combine Bayesian model selection with parameter estimation. In this approach, special priors over the model parameters are chosen that concentrate prior probability on "sparse" parameter vectors, that is, parameter vectors in which some of the elements are exactly zero. These methods have been motivated by the Bayesian connections with non-Bayesian penalized regression procedures such as the non-negative

garotte [48] and the LASSO [49]. A significant advantage of this approach is that one needs only to sample from, or maximise over, the posterior of a single model containing all parameters of interest, and there is no requirement to compute marginal probabilities or to explore discrete model spaces. This is a rapidly growing area of research, and some important contributions of note include the relevance vector machine [50], Bayesian artificial neural networks [51] and the Bayesian LASSO [52, 53].

0.4 Model selection by compression

This section examines the minimum message length (MML) (see Section 0.4.1) and minimum description length (MDL) (see Section 0.4.6) principles of model selection. Unlike the aforementioned distance based criteria, the MML and MDL principles are based on information theory and data compression. Informally, given a dataset and a set of candidate models, both MML and MDL advocate choosing the model that yields the briefest encoding of a hypothetical message comprising the model and the data. This optimal model is the simplest explanation of the data amongst the set of competing models. In this fashion, MML and MDL treat codelengths of compressed data as a measure of model complexity and, therefore, a yardstick for evaluating the explanatory power of competing models. Since data compression is equivalent to statistical learning of regularity in the data, this is intuitively a sensible procedure. It is important to note that one does not need to encode the models or the data in order to do MML and MDL inference. All that is required is to be able to calculate codelengths which can then be used for model comparison. Before discussing how MML and MDL compute codelengths of models and data, a brief discussion of information theory is necessary. For a detailed treatment of information theory, the interested reader is recommended [54].

To aid in understanding the fundamental information theory concepts, it is helpful to consider the following hypothetical scenario. There exists a sender who is transmitting a data set (message) to a receiver over a transmission channel. The receiver does not know the content of the message. The message, $\mathcal{M}$, is recorded using some alphabet $\mathcal{X}$, such as a subset of the English alphabet $\mathcal{X} = \{a, b, c\}$, and comprises a sequence of symbols formed from $\mathcal{X}$; for example, the message using $\mathcal{X} = \{a, b, c\}$ may be $\mathcal{M} = \{bcca\}$. The transmission channel is noiseless and does not modify transmissions in any fashion. Without loss of generality, we assume the coding alphabet used on the transmission channel is the binary alphabet $\mathcal{Y} = \{0, 1\}$. Prior to sending the message, the sender and the receiver agree on a suitable codebook which will be used to transmit all messages on this channel. The codebook is a mapping $\mathrm{C} : \mathcal{X} \to \mathcal{Y}^*$ from the source alphabet $\mathcal{X}$ to the target alphabet $\mathcal{Y}^*$, where $\mathcal{Y}^*$ is the set of all strings obtained by concatenating elements of $\mathcal{Y}$. Some possible codebooks are:

$$\mathrm{C}_1 = \{a \to 0, b \to 10, c \to 11\}, \tag{45}$$

$$\mathrm{C}_2 = \{a \to 00, b \to 01, c \to 10\}, \tag{46}$$

$$\mathrm{C}_3 = \{a \to 0, b \to 01, c \to 001\}. \tag{47}$$

For example, if the codebook C_1 (45) is used by the sender, the message $\{bcca\}$ is mapped to the target code $\{1011110\}$. If the sender uses codebook C_2 instead, the message is encoded as $\{01101000\}$. The task of the sender and receiver is to choose a codebook such that the transmitted message is uniquely decodable by the receiver and as brief as possible.

If the codebook C possess the prefix property, a message transmitted using C is uniquely decodable requiring no delimiters between successive symbols. The prefix property states that no codeword is a prefix of any other codeword. In the above example, codebooks C_1 and C_2 possess the prefix property, while codebook C_3 does not as the code 0 for source symbol a is a prefix to both codes 01 and 001. Since there exist infinitely many prefix codes, the sender and the receiver may choose the prefix code that results in

the briefest possible encoding of the message $\mathcal{M}$. Intuitively, symbols in the message that appear more frequently should be assigned smaller codewords, while symbols that are very rare can be given longer codewords in order to reduce the average length of the encoding. This optimisation problem is now formally explored. Note, that in the rest of the section we only consider codebooks which are uniquely decodable and possess the prefix property.

Let X denote a discrete random variable defined over the support (source alphabet) $\mathcal{X}$ with probability distribution function $p_X(\cdot)$. The random variable X represents the sender of the message. Furthermore, let $l : \mathcal{X} \to \mathbb{R}_+$ denote the codelength function which gives the codeword length for each symbol in the source alphabet $\mathcal{X}$. For example, $l(a) = 1$ bit for the codebook C_1, while $l(a) = 2$ bits, for the codebook C_2. The average codelength per symbol from source alphabet $\mathcal{X}$ with probability distribution function $p_X(\cdot)$ is defined as

$$\mathrm{E}\left[l(X)\right] = \sum_{x \in \mathcal{X}} p_X(x) l(x) \tag{48}$$

where E denotes the expected value of a random variable. We wish to choose the function $l(\cdot)$ such that the codelength for data from the random variable X is on average the shortest possible; that is we seek a function $l(\cdot)$ that minimises

$$\arg \min_{l(x), x \in \mathcal{X}} \{\mathrm{E}\left[l(X)\right]\} \tag{49}$$

The solution to this optimisation problem is

$$l(X) = \log 1/p_X(x) \tag{50}$$

which is the Shannon information of a random variable [55]. The unit of information is derived from the base of the logarithm; commonly binary and natural logarithms are used corresponding to the bit and the nit (nat) units respectively. In the rest of this section, all logarithms are assumed to be natural logarithms, unless stated otherwise.

The Shannon information agrees with the previous intuition that high probability symbols should be assigned shorter codewords, while low probability symbols should be given longer code words. At first, it may seem troubling that Shannon information allows for non-integer length codewords. After all, how does one encode and transmit symbols of length, say, 0.2 of a nit? However, this is not an issue in practice as there exist sophisticated coding algorithms, such as arithmetic codes [56], that allow for efficient encoding in the presence of non-integer codewords. It is important to re-iterate here that MML and MDL model selection does not require any actual encoding of data to be performed; all that is required is a function to compute codelengths for data and models. Model selection is then performed by computing codelengths for all candidate models and choosing the one with the smallest codelength as optimal.

The minimum possible average codelength per symbol for a discrete random variable X is known as the Shannon entropy of X and is defined as

$$H(X) = \mathrm{E}[\log 1/p_X(X)] = \sum_{x \in \mathcal{X}} p_X(x) \log 1/p_X(x) \tag{51}$$

with the convention that $0 \log 0 = 0$; since $x \log x \to 0$ as $x \to 0$, this is reasonable. The entropy of a random variable is always non-negative, $H(X) \geq 0$, and is a measure of uncertainty in the random variable. The maximum entropy is attained when there is maximum uncertainty in X, that is, when X is uniformly distributed over the support $\mathcal{X}$. For example, the entropy of a random variable X with support $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ and probability distribution function

$$p_X(j) = \frac{1}{6} \quad (j = 1, \ldots, 6) \tag{52}$$

is approximately 2.585 bits or 1.792 nits. As an another example, the entropy of random variable X following a Geometric distribution ($\mathcal{X} = \{0, 1, 2, \ldots\}$) with distribution function

$$p_X(X|\lambda) = (1-\lambda)^x \lambda \quad (0 < \lambda \leq 1) \tag{53}$$

is given by

$$H(X|\lambda) = -\frac{1}{\lambda}\left[(1-\lambda)\log(1-\lambda) + \lambda \log \lambda\right], \tag{54}$$

where $H(X|\lambda) = 0$ for $\lambda = 1$, and $H(X|\lambda) \to \infty$ as $\lambda \to 0$.

The entropy of a random variable is the minimum average codelength per symbol with the length of each codeword X given by $l(X) = \log 1/p_X(\cdot)$. By the law of large numbers, it can also be shown that using codelengths of the form $l(X)$ results in a code that is optimal not only on average, but also optimal for the actually generated sequences of data. It is of interest to quantify the inefficiency resulting from encoding symbols generated by $p_X(\cdot)$ using an alternative distribution function $q_X(X)$, where $p_X(\cdot) \neq q_X(\cdot)$ for at least one $X \in \mathcal{X}$. The extra number of nits (or bits) required per symbol on average if $q_X(\cdot)$ is used instead of $p_X(\cdot)$ is given by the Kullback-Leibler (KL) divergence (see Section 0.2). The KL divergence is always non-negative, and is only equal to zero when $p_X(\cdot) = q_X(\cdot)$, for all $X \in \mathcal{X}$. In words, the briefest encoding for data from source $p_X(\cdot)$ is achieved using $p_X(\cdot)$ as the coding distribution and any other distribution, say $q_X(\cdot)$, will necessarily result in messages with a longer average codelength.

Recall that both MML and MDL seek models that result in the shortest codelength of a hypothetical message comprising the observed data and the model. That is, both MML and MDL seek models that result in the best compression of the observed data. Formally, this is a sensible procedure since the set of compressible strings from any data source is in fact quite small in comparison to incompressible strings from the same source. This is quantified by the following *no hypercompression inequality* (p. 136 [57]). Given a sample of n data points $\mathbf{y} = (y_1, \ldots, y_n)$ generated from a probability distribution $p_X(\cdot)$, the probability of compressing $\mathbf{y}$ by *any* code is

$$p\left(\log 1/p_X(y_1, \ldots, y_n) \geq \log 1/q_X(y_1, \ldots, y_n) + K\right) \leq 2^{-K} \tag{55}$$

In words, the inequality states that the probability of compressing $\mathbf{y}$ by K bits using any code is small and decreases exponentially with increasing K. Clearly, the sender and the receiver would like K to be as large as possible leading to the briefest encoding of the observed data. For any moderate K, it is highly unlikely for a random model to result in a compression of K-bits as the number of models that can compress the data in such a manner is vanishingly small with K. In the next section, we examine model selection with the minimum message principle which is based on information theory and data compression.

0.4.1 Minimum Message Length (MML)

The Minimum Message Length (MML) principle for model selection was introduced by C. S. Wallace [23] and collaborators in the late 1960's [29], and has been continuously refined in the following years [58, 59]. The MML principle connects the notion of compression with statistical inference, and uses this connection to provide a unified framework for model selection and parameter estimation. Recalling the transmitter-receiver framework outlined in Section 0.4, consider the problem of efficiently transmitting some observed data $\mathbf{y}$ from the transmitter to the receiver. Under the MML principle, the model that results in the shortest encoding of a decodable message comprising the model and data is considered optimal. For this message to be decodable, the receiver and the transmitter must agree on a common language prior to any data being transmitted. In the MML framework, the common knowledge is a set of parametric models Γ,

each comprising a set of distributions $p(\mathbf{y}|\boldsymbol{\theta}, \gamma)$ indexed by $\boldsymbol{\theta}$, and a prior distribution $\pi(\boldsymbol{\theta}, \gamma) = \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\gamma)\pi_{\gamma}(\gamma)$. The use of a subjective prior distribution means that MML is an unashamedly Bayesian principle built on information theoretic foundations.

Given a model from γ and a prior distribution $\pi_{\boldsymbol{\theta}}(\cdot)$ over Θ_{γ} we may define an (implicit) prior probability distribution, $r(\mathbf{y}|\gamma)$, on the n-dimensional data space $\mathcal{Y}^n$. This distribution $r(\cdot)$, called the marginal distribution of the data, characterises the probability of observing any particular data set $\mathbf{y}$ given our choice of prior beliefs reflected by $\pi_{\boldsymbol{\theta}}(\cdot)$, and plays an important role in conventional Bayesian inference (see Section 0.3). The marginal distribution is

$$r(\mathbf{y}|\gamma) = \int_{\boldsymbol{\theta}\in\Theta_{\gamma}} p(\mathbf{y}|\boldsymbol{\theta}, \gamma)\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\gamma)d\boldsymbol{\theta}. \tag{56}$$

The receiver and the transmitter both have knowledge of the prior density $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\gamma)$ and the likelihood function $p(\mathbf{y}|\boldsymbol{\theta}, \gamma)$, and therefore the marginal distribution. Under the assumption that our prior beliefs accurately reflect the unknown process that generated the data $\mathbf{y}$, the average length of the shortest message that communicates the data $\mathbf{y}$ to the receiver using the model γ is the entropy of $r(\mathbf{y}|\gamma)$. The marginal distribution is then used to construct a codebook for the data resulting in a decodable message with codelength

$$I_0(\mathbf{y}) = -\log r(\mathbf{y}|\gamma) = -\log \int_{\boldsymbol{\theta}\in\Theta_{\gamma}} p(\mathbf{y}|\boldsymbol{\theta}, \gamma)\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\gamma)d\boldsymbol{\theta}. \tag{57}$$

This is the most efficient encoding of the data using the model γ under the assumed likelihood function and prior density. While the marginal distribution is commonly used in Bayesian inference for model selection (as discussed in Section 0.3), the fact that the parameter vector $\boldsymbol{\theta}$ is integrated out means that it may not be used to make inferences about the particular distribution in Θ_{γ} that generated the data. That is, the message does not convey anything about the quality of any particular distribution in Θ_{γ} in regards to encoding the observed data and cannot be used for point estimation. Following Wallace ([23], pp. 152–153), we shall refer to this code as the *non-explanation* code for the data.

In the MML framework, the message transmitting the observed data comprises two components, commonly named the *assertion* and the *detail*. The assertion is a codeword naming a particular distribution indexed by $\boldsymbol{\theta} \in \Theta_{\gamma}$ from the model γ, rather than a range of possible distributions in γ. The detail is a codeword that states the data $\mathbf{y}$ using the distribution $\boldsymbol{\theta}$ that was named in the assertion. The total, joint codelength of the message is then

$$I(\mathbf{y}, \boldsymbol{\theta}, \gamma) = I(\boldsymbol{\theta}, \gamma) + I(\mathbf{y}|\boldsymbol{\theta}, \gamma), \tag{58}$$

where $I(\boldsymbol{\theta}, \gamma)$ and $I(\mathbf{y}|\boldsymbol{\theta}, \gamma)$ denote the length of the assertion and detail respectively. The length of the assertion measures the complexity of a particular distribution $\boldsymbol{\theta}$ in model γ, while the detail length measures the quality of the fit of the distribution $\boldsymbol{\theta}$ to the data $\mathbf{y}$. Thus, the total, joint codelength automatically captures the tradeoff between model complexity and model capacity. Intuitively, a complex model with many free parameters can be tuned to fit a large range of data sets, leading to a short detail length. However, since all parameters need to be transmitted to the receiver, the length of the assertion will be long. In contrast, a simple model with few free parameters requires a short assertion, but may not fit the data well, potentially resulting in a long detail. This highlights the first defining feature of the MML principle: *the measure of model complexity and model fit are both expressed in the same unit, namely the codelength*. The minimum message length principle advocates choosing the model γ and distribution $\boldsymbol{\theta}$ that minimises this two-part message, that is,

$$\left\{\hat{\boldsymbol{\theta}}(\mathbf{y}; \gamma), \hat{\gamma}_{\mathrm{MML}}(\mathbf{y})\right\} = \arg \min_{\gamma\in\Gamma, \boldsymbol{\theta}\in\Theta_{\gamma}} \{I(\mathbf{y}, \boldsymbol{\theta}, \gamma)\}.$$

This expression highlights the second defining feature of the MML principle: *minimisation of the codelength is used to perform both selection of a model structure, $\gamma \in \Gamma$, as well as estimation of the model parameters,*

$\boldsymbol{\theta} \in \Theta$. This is in contrast to both the distance based criteria of Section 0.2, and the Bayesian information criterion introduced in Section 0.3, which only select a structure γ and do not provide an explicit framework for parameter estimation. It remains to determine how one computes the codelength (58).

In general, while the set of candidate models Γ will usually be countable[1], the parameter space Θ_γ for a model γ is commonly continuous. The continuity of Θ_γ creates a problem for the transmitter when designing a codebook for the members of Θ_γ needed for the assertion, as valid codebooks can only be constructed from discrete distributions. This problem may be circumvented by quantizing the continuum Θ_γ into a discrete, countable subset $\bar{\Theta}_\gamma = \{\bar{\theta}_1, \bar{\theta}_2, \ldots\} \subset \Theta_\gamma$. Given a subset $\bar{\Theta}_\gamma$ we can define a distribution over $\bar{\Theta}_\gamma$, say $h(\cdot|\gamma)$. This distribution implicitly defines a codelength for the members of $\bar{\Theta}_\gamma$. The assertion length for stating a particular $\boldsymbol{\theta} \in \bar{\Theta}_\gamma$ is then $I(\boldsymbol{\theta}, \gamma) = -\log h(\boldsymbol{\theta}|\gamma)\pi_\gamma(\gamma)$. The length of the detail, encoded using the distribution indexed by $\boldsymbol{\theta} \in \bar{\Theta}_\gamma$ is $I(\mathbf{y}|\boldsymbol{\theta}, \gamma) = -\log p(\mathbf{y}|\boldsymbol{\theta}, \gamma)$, which is the negative log-likelihood of the observed data $\mathbf{y}$. The question remains: how do we choose the quantisation $\bar{\Theta}_\gamma$ and the distribution $h(\cdot|\gamma)$? The minimum message length principle specifies that the quantisation $\bar{\Theta}_\gamma$ and the distribution $h(\cdot|\gamma)$ be chosen such that the *average* codelength of the resulting two-part message is minimum, the average being taken with respect to the marginal distribution. Formally, for a model γ, the MML two-part codebook solves the following minimisation problem:

$$\min_{\bar{\Theta},h} \left\{ \sum_{\mathbf{y} \in \mathcal{Y}^n} r(\mathbf{y}|\gamma) I(\mathbf{y}, \boldsymbol{\theta}, \gamma) \right\}. \tag{59}$$

In the literature, this procedure is referred to as strict minimum message length (SMML) [58]. For computational reasons, the SMML procedure commonly described in the literature solves an equivalent minimisation problem by partitioning the dataspace $\mathcal{Y}^n$, which implicitly defines the parameter space quantisation $\bar{\Theta}_\gamma$ and the distribution $h(\cdot|\gamma)$ (for a detailed exposition the reader is referred to Chapter 3 of [23]). Model selection and parameter estimation by SMML proceeds by choosing the model/parameter pair from $\bar{\Theta}_\gamma$ and Γ that results in the shortest two-part message, given $h(\cdot|\gamma)$. Interestingly, this is equivalent to performing *maximum a posteriori* (MAP) estimation over the quantised parameter space $\bar{\Theta}_\gamma$, treating $h(\cdot|\gamma)$ as a special "quantised prior" which is implicitly defined by both the quantisation $\bar{\Theta}_\gamma \subset \Theta_\gamma$ and the original continuous prior distribution $\pi_{\boldsymbol{\theta}}(\cdot)$ over Θ_γ.

Analysis of the SMML code shows that it is approximately $(1/2)(\log(k\pi) - 1)$ nits longer than the non explanation code (57), where k is the dimensionality of Θ_γ (pp. 237–238, [23]). This is not surprising given that the SMML code makes a statement about the particular distribution from Θ_γ which was used to code the data, rather than using a mixture of distributions as in the non-explanation code. This apparent "inefficiency" in SMML allows the codelength measure to be used not only for model selection (as in the case of the non-explanation code), but also provides a theoretically sound basis for parameter estimation.

While the SMML procedure possesses many attractive theoretical properties (see for example, pp. 187–195 of [23] and the "false oracle" property discussed in [60]), the minimisation problem (59) is in general NP-hard [61], and exact solutions are computationally tractable in only a few simple cases. The difficulties arise primarily due to the fact that a complete quantisation of the continuous parameter space must be found, which is a non-trivial problem that does not scale well with increasing dimensionality of $\mathcal{Y}^n$. In the next section we shall examine an approximation to the two-part message length that exploits the fact that we are only interested in the computing the *length* of the codes, rather than the codebooks. This allows us to circumvent the problematic issue of complete quantisation by using only local properties of the likelihood and prior distribution, resulting in an implementation of the MML principle that is applicable to many commonly used statistical models.

[1] A set A is countable if its members can be put into a one-to-one correspondence with the natural numbers, that is, $A \leftrightarrow \mathbb{N}$.

0.4.2 The Wallace–Freeman Message Length Approximation (MML87)

To circumvent the problem of determining a complete quantisation of the parameter space, Wallace and Freeman [59] presented a formula that gives an approximate length of the two-message for a particular $\boldsymbol{\theta} \in \Theta_\gamma$ and data set $\mathbf{y}$. Rather than finding an optimal quantisation for the entire parameter space Θ_γ, the key idea underlying their approach is to find an optimal *local* quantisation for the particular $\boldsymbol{\theta}$ named in the assertion. The important implication of this approach is that the quantisation, and therefore the resulting message length formula, depends only on the particular $\boldsymbol{\theta}$ that is of interest. Under certain regularity conditions, the Wallace–Freeman approximation states that the joint codelength of a distribution $\boldsymbol{\theta} \in \Theta_\gamma$ in model γ, and data $\mathbf{y}$, is

$$I_{87}(\mathbf{y}, \boldsymbol{\theta}, \gamma) = \underbrace{-\log \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\gamma)\pi_\gamma(\gamma) + \frac{1}{2}\log|\mathbf{J}_n(\boldsymbol{\theta};\gamma)| + \frac{k}{2}\log\kappa_k}_{I_{87}(\boldsymbol{\theta},\gamma)} + \underbrace{\frac{k}{2} - \log p(\mathbf{y}|\boldsymbol{\theta},\gamma)}_{I_{87}(\mathbf{y}|\boldsymbol{\theta},\gamma)}, \tag{60}$$

where $k = \dim(\Theta_\gamma)$ is the total number of continuous, free parameters for model γ and $\mathbf{J}(\cdot)$ is the Fisher information matrix for n data points given by

$$\mathbf{J}_n(\boldsymbol{\theta}_0;\gamma) = -\mathrm{E}_{\boldsymbol{\theta}_0}\left[\left.\frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta},\gamma)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]. \tag{61}$$

The quantity κ_k is the normalized second moment of an optimal quantising lattice in k-dimensions [62]. Following ([23], p. 237), the need to determine κ_k for arbitrary dimension k is circumvented by use of the following approximation

$$\frac{k}{2}(\log\kappa_k + 1) \approx -\frac{k}{2}\log(2\pi) + \frac{1}{2}\log(k\pi) + \psi(1) = c(k),$$

where $\psi(\cdot)$ is the digamma function, and $\psi(1) \approx -0.5772$. Model selection, and parameter estimation, is performed by choosing the model/parameter pair that minimises the joint codelength, that is,

$$\left\{\hat{\gamma}_{87}(\mathbf{y}), \hat{\boldsymbol{\theta}}_{87}(\mathbf{y}; \hat{\gamma}_{87}(\mathbf{y}))\right\} = \arg\min_{\gamma\in\Gamma, \boldsymbol{\theta}\in\Theta_\gamma}\left\{I_{87}(\mathbf{y}, \boldsymbol{\theta}, \gamma)\right\}. \tag{62}$$

Unlike the SMML codelength, the MML87 codelength (60) is a continuous function of the continuous model parameters $\boldsymbol{\theta}$ which makes the minimisation problem (62) considerably easier. Again, it is important to re-iterate that unlike the distance based criteria (AIC, KIC, etc.) and the BIC, minimisation of the MML87 formula yields both estimates of model structure as well as model parameters. Further, the MML87 parameter estimates possess attractive properties in comparison with estimates based on other principles such as maximum likelihood or maximum posterior mode.

An important issue remains: under what conditions is the MML87 formula valid? In brief, the MML87 codelength is generally valid if: (i) the maximum likelihood estimates associated with the likelihood function obey the central limit theorem everywhere in Θ_γ, and (ii) the local curvature of the prior distribution $\pi_{\boldsymbol{\theta}}(\cdot)$ is negligible compared to the curvature of the negative log-likelihood everywhere in Θ_γ. An important point to consider is how close the Wallace–Freeman approximate codelength is to the exact SMML codelength given by (59). If the regularity conditions under which the MML87 formula was derived are satisfied, then the MML87 codelength is, on average, practically indistinguishable from the exact SMML codelength (pp. 230–231, [23]).

In the particular case that the prior $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\alpha},\gamma)$ is *conjugate* [2] with the likelihood $p(\mathbf{y}|\boldsymbol{\theta},\gamma)$, and depends on some parameters $\boldsymbol{\alpha}$ (usually referred to as "hyperparameters"), Wallace (pp. 236–237, [23]) suggested

[2] The use of a conjugate prior ensures that the posterior distribution of the parameters is of the same mathematical form as the conjugate prior distribution.

a clever modification of the MML87 formula that makes it valid even if the curvature of the prior is significantly greater than the curvature of the likelihood. The idea is to first propose some imaginary "prior data" $\mathbf{y}_\alpha$ whose properties depend only on the prior hyperparameters $\boldsymbol{\alpha}$. It is then possible to view the prior $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\alpha}, \gamma)$ as a *posterior* of the likelihood of this prior data $\mathbf{y}_\alpha$ and some initial uninformative prior $\pi_0(\boldsymbol{\theta})$ that does not depend on the hyperparameters $\boldsymbol{\alpha}$. Formally, we seek the decomposition

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\alpha}, \gamma) \propto \pi_0(\boldsymbol{\theta}) p(\mathbf{y}_\alpha|\boldsymbol{\theta}, \gamma) \tag{63}$$

where $p(\mathbf{y}_\alpha|\boldsymbol{\theta}, \gamma)$ is the likelihood of m imaginary prior samples and $\pi_0(\boldsymbol{\theta})$ is an uninformative prior. The new Fisher Information $\mathbf{J}_{n+m}(\boldsymbol{\theta}; \gamma)$ is then constructed from the new combined likelihood $p(\mathbf{y}, \mathbf{y}_\alpha|\boldsymbol{\theta}, \gamma) = p(\mathbf{y}|\boldsymbol{\theta}) p(\mathbf{y}_\alpha|\boldsymbol{\theta}, \gamma)$, and the "curved prior" MML87 approximation is simply

$$\begin{aligned} I^*_{87}(\mathbf{y}, \boldsymbol{\theta}, \gamma) &= -\log \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\alpha}, \gamma) \pi_\gamma(\gamma) + \frac{1}{2} \log |\mathbf{J}_{n+m}(\boldsymbol{\theta}; \gamma)| - \log p(\mathbf{y}|\boldsymbol{\theta}, \gamma) + c(k), & (64) \\ &= -\log \pi_0(\boldsymbol{\theta}) \pi_\gamma(\gamma) + \frac{1}{2} \log |\mathbf{J}_{n+m}(\boldsymbol{\theta}; \gamma)| - \log p(\mathbf{y}, \mathbf{y}_\alpha|\boldsymbol{\theta}, \gamma) + c(k). & (65) \end{aligned}$$

This new Fisher information matrix has the interesting interpretation of being the asymptotic lower bound of the inverse of the variance of the maximum likelihood estimator using the combined data $(\mathbf{y}, \mathbf{y}_\alpha)$ rather than for simply the observed data $\mathbf{y}$. It is even possible in this case to treat the hyperparameters as unknown, and use an extended message length formula to estimate both the model parameters $\boldsymbol{\theta}$ and the prior hyperparameters $\boldsymbol{\alpha}$ in this hierarchical structure [63, 64].

What is the advantage of using the MML87 estimates over the traditional maximum likelihood and MAP estimates? In addition to possessing important invariance and consistency properties that are discussed in the next sections, there is a large body of empirical evidence demonstrating the excellent performance of the MML87 estimator in comparison to traditional estimators such as maximum likelihood and the *maximum a posteriori* (MAP) estimator, particularly for small to moderate sample sizes. Examples include the normal distribution [23], factor analysis [65], estimation of multiple means [63], the von Mises and spherical von Mises-Fisher distribution [66, 67] (pp. 266–269, [23]) and autoregressive moving average models [68]. In the next sections we will discuss some important properties of the Wallace–Freeman approximation. For a detailed treatment of the MML87 approximation, we refer the interested reader to Chapter 5 of [23].

Model Selection Consistency

Recall from Section 0.3.2 that the BIC procedure yields a consistent estimate of the true, underlying model, say $\gamma^* \in \Gamma$, that generated the data. The estimate $\hat{\gamma}_{87}$ is also a consistent estimate of γ^* as $n \to \infty$ under certain conditions. Assuming that the Fisher information matrix satisfies

$$\mathbf{J}_1(\boldsymbol{\theta}; \gamma) = \lim_{n \to \infty} \left\{ \frac{\mathbf{J}_n(\boldsymbol{\theta}; \gamma)}{n} \right\}, \tag{66}$$

with $|\mathbf{J}_1(\cdot)| > 0$, the MML87 codelength approximation can be rewritten as

$$I_{87}(\mathbf{y}, \boldsymbol{\theta}, \gamma) = -\log p(\mathbf{y}|\boldsymbol{\theta}, \gamma) + \frac{k}{2} \log n + O(1). \tag{67}$$

This is clearly equivalent to the well known BIC discussed in Section 0.3.1, and thus the estimator $\hat{\gamma}_{87}$ is consistent as $n \to \infty$. For a more rigorous and general proof, the reader is directed to [69].

Invariance

There is in general no unique way to parameterise the distributions that comprise a statistical model. For example, the normal distribution is generally expressed in terms of mean and variance parameters, say (μ, τ). However, the alternate parameterisations $(\mu, \sqrt{\tau})$ and $(e^{-\mu}, \tau)$ are equally valid and there is no reason, beyond ease of interpretation, to prefer any particular parameterisation. A sensible estimation procedure should be invariant to the parameterisation of the model; that is, it should give the same answer irrespective of how the inference question is framed. While the commonly used maximum likelihood estimator possesses the invariance property, both the *maximum a posteriori* estimator and posterior mean estimator based on the Bayesian posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}, \gamma)$ given by (28) are *not invariant*. The SMML and MML87 *are invariant* under one-to-one reparameterisations, and thus provide an important alternative Bayesian estimation procedure.

Parameter Estimation Consistency

Consider the case that the data $\mathbf{y}$ is generated by a particular distribution $\boldsymbol{\theta}^*$ from the model γ, that is, $\boldsymbol{\theta}^* \in \Theta_\gamma$. A sequence of estimators, $\hat{\boldsymbol{\theta}}_n$, of parameter $\boldsymbol{\theta}$ is consistent if and only if

$$\Pr\left\{|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*| \geq \varepsilon\right\} \to 0$$

for any $\varepsilon > 0$ as $n \to \infty$. In words, this means that as we accumulate more and more data about the parameters the estimator converges to the true parameter value $\boldsymbol{\theta}^*$. If we consider the case that the dimensionality of Θ_γ does not depend on n, and the Fisher information matrix satisfies (36) then the maximum likelihood estimator is consistent under suitable regularity conditions [70]. The consistency of the MML87 estimator in this particular case follows by noting that minimising the asymptotic message length (67) is equivalent to maximising the likelihood; similar arguments can also be used to establish the consistency of the MAP and mean posterior estimators in this case.

The situation is strikingly different in the case that the dimensionality of the parameter space depends on the sample size. In this setting there is usually a small number of parameters of interest for which information grows with increasing sample size, and a large number of auxiliary "nuisance" parameters for which an increasing sample size brings no new information. Maximum likelihood estimation, as well as the MAP and mean posterior estimation, of the parameters of interest is in general inconsistent. In contrast, the MML87 estimator has been shown to provide consistent estimates in the presence of many nuisance parameters in several important statistical models, including the Neyman–Scott problem [71], factor analysis [65, 72] (also pp. 297–303 in [23]), multiple cutpoint estimation [73] and mixture modelling ([74] and pp. 275–297 in [23]). While a general proof of the consistency of the MML principle in the presence of nuisance parameters does not currently exist, *there have been no problems studied so far in which it has failed to yield consistent estimates.*

Similarities with Bayesian Inference

There are many similarities between the minimum message length principle and the more conventional Bayesian approaches such as those discussed in Section 0.3 (see [75] for a detailed discussion). The differences in codelengths (MML87 or otherwise) between two models may be used to compute an approximate Bayes factor, that is,

$$-\log \mathrm{BF}(\gamma_1, \gamma_0) \approx I(\mathbf{y}, \boldsymbol{\theta}, \gamma_0) - I(\mathbf{y}, \boldsymbol{\theta}, \gamma_1).$$

The usual interpretations applied to regular Bayes factors apply equally well to the approximate Bayes factors determined by codelength differences. Further, the codelength measure may be used to perform model averaging over the set of candidate models Γ in a similar fashion to BIC, as discussed in Section 0.3.1. In this case, we use (41), replacing the BIC score with the codelength, and (ideally) replacing the maximum likelihood estimate with the appropriate minimum message length estimate.

0.4.3 Other Message Length Approximations

Recently, Schmidt [68, 76] introduced a new codelength approximation that can be used for models for which the application of the Wallace–Freeman approximation is problematic. The new approximation, called MML08, is more computationally tractable than the strict MML procedure, but requires evaluation of sometimes difficult integrals. The joint MML08 codelength of data $\mathbf{y}$ and parameters $\boldsymbol{\theta}$, for model γ, is given by

$$I_{08}(\boldsymbol{\theta}, \mathbf{y}, \gamma) = \underbrace{-\log \pi_\gamma(\gamma) \int_{\Omega_\theta} \pi_{\boldsymbol{\theta}}(\boldsymbol{\phi}|\gamma) d\boldsymbol{\phi}}_{I_{08}(\boldsymbol{\theta},\gamma)} \underbrace{- \log p(\mathbf{y}|\boldsymbol{\theta}, \gamma) + \left(\frac{1}{\int_{\Omega_\theta} \pi(\boldsymbol{\phi}|\gamma) d\boldsymbol{\phi}} \right) \int_{\Omega_\theta} \pi_{\boldsymbol{\theta}}(\boldsymbol{\phi}|\gamma) \Delta(\boldsymbol{\theta}||\boldsymbol{\phi}) d\boldsymbol{\phi}}_{I_{08}(\mathbf{y}|\boldsymbol{\theta},\gamma)}, \quad (68)$$

where $\Delta(\cdot)$ is the Kullback-Leibler divergence and $\Omega_{\boldsymbol{\theta}} \subset \Theta_\gamma$ is chosen to minimise the codelength. The MML08 is a generalisation of the previously proposed MMLD codelength approximation [77, 78] and is invariant under transformations of the parameters. The MML08 approximation has been applied to the problem of order selection for autoregressive moving average models [68]. Efficient algorithms to compute approximate MMLD and MML08 codelengths are given in [79, 68, 80].

0.4.4 Example: MML Linear Regression

We continue our running example by applying the minimum message length principle to the selection of covariates in the linear regression model. By choosing different prior distributions for the coefficients $\boldsymbol{\beta}$ we can arrive at many different MML criteria (for example, [81], [82] and pp. 270–272, [23]); however, provided the chosen prior distribution is not unreasonable all codelength criteria will perform approximately equivalently, particularly for large sample sizes. Two recent examples of MML linear regression criteria, called "MML$_u$" and "MML$_g$", that do not require the specification of any subjective prior information are given in [64]. The MML$_u$ criterion exploits some properties of the linear model with Gaussian noise to derive a data driven, proper uniform prior for the regression coefficients. Recall that in the linear regression case, the model index γ specifies the covariates to be used from the full design matrix $\mathbf{X}$. For a given γ, the MML$_u$ codelength for a linear regression model is

$$\left(\frac{n-k}{2}\right) \log 2\pi + \left(\frac{n-k}{2}\right) (\log \hat{\tau}_{87}(\mathbf{y}; \gamma) + 1) + \frac{k}{2} \log (\pi \mathbf{y}'\mathbf{y}) - \log \Gamma\left(\frac{k}{2} + 1\right) + \frac{1}{2} \log(k+1) - \log \pi_\gamma(\gamma), \quad (69)$$

where the MML$_u$ parameter estimates are given by

$$\hat{\tau}_{87}(\mathbf{y}; \gamma) = \left(\frac{1}{n-k}\right) (\mathbf{y}'\mathbf{y} - R(\gamma)), \quad (70)$$

$$\hat{\boldsymbol{\beta}}_{87}(\mathbf{y}; \gamma) = \left(\mathbf{X}_\gamma'\mathbf{X}_\gamma\right)^{-1} \mathbf{X}_\gamma'\mathbf{y}. \quad (71)$$

Here, $k = |\gamma|$ denotes the number of covariates in the regression model γ, $\pi_\gamma(\gamma)$ is a suitable prior for the model index (see Section 0.3.4 for a discussion of suitable priors), and

$$R(\gamma) = \hat{\boldsymbol{\beta}}(\mathbf{y}; \gamma)'(\mathbf{X}_\gamma'\mathbf{X}_\gamma)\hat{\boldsymbol{\beta}}(\mathbf{y}; \gamma) \quad (72)$$

is the fitted sum-of-squares for the least-squares estimates (5). Due to the use of a uniform prior on the regression coefficients, and the nature of the Fisher information matrix, the MML87 estimates of the regression coefficients coincide with the usual maximum likelihood estimates (5). In contrast to the maximum

likelihood estimate of τ, given by (6), the MML87 estimate is exactly unbiased even for finite sample sizes n.

The MML_g criterion further demonstrates the differences in parameter estimation that are possible by using the MML principle. This criterion is based on a special hierarchical Gaussian prior, and uses some recent extensions to the MML87 codelength to estimate the parameters of the prior distribution in addition to the regression coefficients [63]. For a given γ, the MML_g codelength for a linear regression model is

$$\left(\frac{n-k+2}{2}\right)(\log \hat{\tau}_{87}(\mathbf{y};\gamma)+1)+\left(\frac{k-2}{2}\right)\log\left(\frac{R(\gamma)}{\delta}\right)+\frac{\delta}{2}+\frac{1}{2}\log(n-k)k^2-\log\pi_\gamma(\gamma) \tag{73}$$

where $\delta = \max(1, k-2)$. This formula is applicable only when $k > 0$, and $m(\gamma) > 0$, where

$$m(\gamma)=\left(\frac{R(\gamma)}{\delta}-\hat{\tau}_{87}(\mathbf{y};\gamma)\right)_+,$$

and $(\cdot)_+ = \max(0, \cdot)$ is the positive-part function. The codelength for a "null" model ($k = 0$) is given by

$$\left(\frac{n}{2}\right)(\log \hat{\tau}_{87}(\mathbf{y};\gamma)+1)+\frac{1}{2}\log(n-1)+\frac{1}{2}. \tag{74}$$

We note that (73) corrects a minor mistake in the simplified form of MML_g in [64], which erroneously excluded the additional "$\delta/2$" term. The MML_g estimates of $\boldsymbol{\beta}$ and τ are given by

$$\hat{\tau}_{87}(\mathbf{y};\gamma) = \left(\frac{1}{n-k+2}\right)(\mathbf{y}'\mathbf{y}-R(\gamma)), \tag{75}$$

$$\hat{\boldsymbol{\beta}}_{87}(\mathbf{y};\gamma) = \left(\frac{m(\gamma)}{m(\gamma)+\hat{\tau}_{87}(\mathbf{y};\gamma)}\right)\left(\mathbf{X}_\gamma'\mathbf{X}_\gamma\right)^{-1}\mathbf{X}_\gamma'\mathbf{y}, \tag{76}$$

which are closely related to the unbiased estimate (70) of τ used by the MML_u criterion. The MML_g estimates of the regression coefficients $\boldsymbol{\beta}$ are the least-squares estimates scaled by a quantity less than unity which is called a *shrinkage* factor [83]. This factor depends crucially on the estimated strength of the noise variance and signal strength, and if the noise variance is too large, $m(\gamma) = 0$ and the coefficients are completely shrunk to zero (that is, none of the covariates are deemed associated with the observations). Further, it turns out that the MML_g shrinkage factor is optimal in the sense that for $k \geq 3$, the shrunken estimates of $\boldsymbol{\beta}$ will, on average, be better at predicting future data arising from the same source than the corresponding least squares estimates [84].

0.4.5 Applications of MML

The MML principle has been applied to a large number statistical models. Below is an inexhaustive list of some of the many successful applications of MML; a more complete list may be found in [78].

- Mixture modelling [29, 74, 85]
- Decision trees/graphs [86, 87, 88]
- Casual Networks [89, 90]
- Artificial neural networks [91, 92]

- Linear Regression [81, 82, 93, 64]
- Autoregressive moving average (ARMA) models [94, 68, 95]
- Discrete time series [96, 73]
- Hierarchical Bayesian Models [63, 64]

Detailed derivations of the message length formulas for many of these models, along with discussions of their behaviour, can also be found in Chapters 6 and 7 of [23].

0.4.6 The Minimum Description Length (MDL) Principle

The minimum description length (MDL) principle [97, 98] was independently developed by Jorma Rissanen from the late 1970s [28, 99, 100], and embodies the same idea of statistical inference via data compression as the MML principle. While there are many subtle, and not so subtle, differences between MML and MDL, the most important is the way in which they view prior distributions (for a comparison of early MDL and MML, see [101]). The MDL principle views prior distributions purely as devices with which to construct codelengths, and the bulk of the MDL research has been focused on finding ways in which the specification of subjective prior distributions may be avoided. In the MML principle, the codebook is constructed to minimise the average codelength of data drawn from the marginal distribution, which captures our prior beliefs about the data generating source. The MDL principle circumvents the requirement to explicitly specify prior beliefs, and instead appeals to the concept of minimising the worst-case behaviour of the codebook. Formally, we require the notion of coding *regret*

$$R(\mathbf{y}) = I(\mathbf{y}, \gamma) + \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y}; \gamma), \gamma), \tag{77}$$

where $I(\mathbf{y}, \gamma)$ is the codelength of the data $\mathbf{y}$ using model γ, and $\hat{\boldsymbol{\theta}}(\mathbf{y}; \gamma) \in \Theta_\gamma$ is the maximum likelihood estimator. As the maximum likelihood estimator minimises the codelength of the data (without stating anything about the model), the negative log-likelihood evaluated at the maximum represents an ideal target codelength for the data. Unfortunately, this is only realisable if the sender and receiver have *a priori* knowledge of the maximum likelihood estimator, which in an inferential setting is clearly nonsensical.

Clearly, one would like to keep the regret as small as possible in some sense, to ensure that the chosen codebook is efficient. The MML approach minimises average excess codelength over the marginal distribution, which requires an explicit prior distribution. To avoid this requirement, Rissanen advocates choosing a codebook that minimises the maximum (worst-case) coding regret (77), that is finding the probability distribution f that solves the following problem:

$$\min_f \left\{ \max_{\mathbf{y} \in \mathcal{Y}^n} \left\{ -\log f(\mathbf{y}) + \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y}; \gamma), \gamma) \right\} \right\}. \tag{78}$$

The solution to this optimisation problem is known as the normalized maximum likelihood (NML) [102, 103] (or Shtarkov [104]) distribution, and the resulting codelength is given by

$$I_{\text{NML}}(\mathbf{y}, \gamma) = -\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y}; \gamma), \gamma) + \log \int_{\mathbf{x} \in \mathcal{Y}^n} p(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x}; \gamma), \gamma) d\mathbf{x}. \tag{79}$$

In MDL parlance, the second term on the right hand side of (79) is called the *parametric complexity* of the model γ. The parametric complexity has several interesting interpretations: (i) it is the logarithm of

the normalising constant required to render the infeasible maximum likelihood codes realisable; (ii) it is the (constant) regret obtained by the normalized maximum likelihood codes with respect to the unattainable maximum likelihood codes, and (iii) it is the logarithm of the number of distinguishable distributions contained in the model γ at sample size n [105]. In addition to the explicit lack of prior distributions, the NML code (79) differs from the MML two-part codes in the sense that it does not explicitly depend on any particular distribution in the model γ. This means that like the marginal distribution discussed in Section 0.4.1, the NML codes cannot be used for point estimation of the model parameters $\boldsymbol{\theta}_\gamma$.

Practically, the normalising integral in (79) is difficult to compute for many models. Rissanen has derived an asymptotic approximation to the NML distribution which is applicable to many commonly used statistical models (including for example, linear models with Gaussian noise and autoregressive moving average models) [102]. Under this approximation, the codelength formula is given by

$$I_{\text{ANML}}(\mathbf{y}, \gamma) = -\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y};\gamma), \gamma) + \frac{k}{2}\log\left(\frac{n}{2\pi}\right) + \log \int_{\boldsymbol{\theta}\in\Theta_\gamma} \sqrt{|\mathbf{J}_1(\boldsymbol{\theta};\gamma)|}d\boldsymbol{\theta}, \tag{80}$$

where k is the number of free parameters for the model γ, n is the sample size and $\mathbf{J}_1(\cdot)$ is the per sample Fisher information matrix given by (36). The approximation (80) swaps an integral over the dataspace for an often simpler integral over the parameter space. For models that satisfy the regularity conditions discussed in [102], the approximate NML codelength satisfies

$$I_{\text{NML}}(\mathbf{y}, \gamma) = I_{\text{ANML}}(\mathbf{y}, \gamma) + o(1)$$

where $o(1)$ denotes a term that disappears as $n \to \infty$.

The NML distribution is only one of the coding methods considered by Rissanen and co-workers in the MDL principle; the interested reader is referred to P. Grünwald's book [57] for a detailed discussion of the full spectrum of coding schemes developed within the context of MDL inference.

0.4.7 Problems with Divergent Parametric Complexity

The NML distribution offers a formula for computing codelengths that is free of the requirement to specify suitable prior distributions, and therefore appears to offer a universal approach to model selection by compression. Unfortunately, the formula (79) is not always applicable as the parametric complexity can diverge, even in commonly used models [106]; this also holds for the approximate parametric complexity in (80). To circumvent this problem, Rissanen [102] recommends bounding either the dataspace, for (79), or the parameter space, for (80), so that the parametric complexity is finite. In the case of linear regression models with Gaussian noise, Rissanen extends this idea even further by treating the hyperparameters that specify the bounding region as parameters, and re-applying the NML formula to arrive at a parameter-free criterion (see [107] for details). Other models with infinite parametric complexity which have been addressed in a similar manner include the Poisson and geometric distributions [106] and stationary autoregressive models [108].

Unfortunately, the choice of bounding region is in general arbitrary, and different choices will lead to different codelength formulae, and hence different behaviour. In this sense, the choice of bounding region is akin to the selection of a prior distribution in MML; however, advocates of the Bayesian approach argue that selection of a prior density is significantly more transparent, and interpretable, than the choice of a bounding region.

0.4.8 Sequential variants of MDL

One way to overcome the problems of infinite parametric complexity is to code the data sequentially. In this framework, one uses the first t datapoints, $(y_1, \ldots, y_t)$, to construct a predictive model that is used to code y_{t+1}. This process may be repeated until the entire dataset $(y_1, \ldots, y_n)$ has been "transmitted", and the resulting codelength may be used for model selection. This idea was first proposed as part of the predictive MDL procedure [109, 97].

More recently, the idea has been refined with the introduction of the sequentially normalized maximum likelihood (SNML) model [110], and the related conditional normalized maximum likelihood (CNML) model [111]. These exploit the fact that the parametric complexity, *conditional* on some observed data, is often finite, even if the unconditional parametric complexity diverges. Additionally, the CNML distributions can be used to make predictions about future data arising from the same source in a similar manner to the Bayesian predictive distribution. A recent paper comparing SNML and CNML against Bayesian and MML approaches in the particular case of exponential distributions found that the SNML predictive distribution has favourable properties in comparison to the predictive distribution formed by "plugging in" the maximum likelihood estimates [112]. Further information on variants of the MDL principle is available in [98] and [57].

0.4.9 Relation to MML and Bayesian Inference

It is of interest to compare the NML coding scheme to the codebooks advocated by the MML principle. This is most easily done by comparing the approximate NML codelength (80) to the Wallace–Freeman approximate codelength (60). Consider the so-called Jeffreys prior distribution:

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \gamma) = \frac{\sqrt{|\mathbf{J}_n(\boldsymbol{\theta}; \gamma)|}}{\int_{\boldsymbol{\alpha} \in \Theta_\gamma} \sqrt{|\mathbf{J}_n(\boldsymbol{\alpha}; \gamma)|} d\boldsymbol{\alpha}}. \tag{81}$$

The normalising term in the above equation is the approximate parametric complexity in (80). Substituting (81) into the Wallace–Freeman code (60) leads to the cancellation of the Fisher information term, and the resulting Wallace–Freeman codelength is within $O(\log k)$ of the approximate NML codelength. This difference in codelengths is attributed to the fact that MML codes state a specific member of the set Θ_γ which is used to transmit the data, while the NML distribution makes no such assertion. The close similarity between the MDL and MML codelengths was demonstrated in [112] in the context of coding data arising from exponential distributions. A further comparison of MDL, the NML distribution and MML can be found in [23], pp. 408–415.

The close equivalence of the approximate NML and Wallace–Freeman codes implies that like the BIC, the approximate NML criterion is a consistent estimator of the true model γ^* that generated the data, assuming $\gamma^* \in \Gamma$, and that the normalising integral is finite. In fact, the earliest incarnation of MDL introduced in 1978 [28] was exactly equivalent to the Bayesian information criterion, which was also introduced in 1978. This fact has caused some confusion in the engineering community, where the terms MDL and BIC are often used interchangeably to mean the "k-on-two log n" criterion given by (37) (this issue and its implications are discussed, for example, in [113]).

We conclude this section with a brief discussion of the differences and similarities between the MML and MDL principles. This is done to attempt to clear up some of the confusion and misunderstands surrounding these two principles. The MDL principle encompasses a range of compression schemes (such as the NML model, the Bayesian model, etc.), which are unified by the deeper concept of *universal models* for data compression. In recent times, the MDL community has focused on those universal models that attain

minimax regret with respect to the maximum likelihood codes to avoid the specification of subjective prior distributions. The two-part codebooks used in the MML principle are also types of universal models, and so advocates of the MDL school often suggest that MML is subsumed in the MDL principle. However, as the above coincidence of Wallace–Freeman and NML demonstrates, the MDL universal models can be implemented in the MML framework by the appropriate choice of prior distribution, and thus the MML school tend to suggest that MDL is essentially a special case of the MML theory, particularly given the time-line of developments in the two camps.

Practically, the most important difference between the two principles is in the choice of two-part (for MML) versus one-part (for MDL) coding schemes. The two-part codes chosen by MML allow for the explicit definition of new classes of parameter estimators, in addition to model selection criteria, which are not obtainable by the one-part MDL codes. As empirical, and theoretical evidence strongly supports the general improvement of the MML estimators over the maximum likelihood and Bayesian MAP estimators, this difference seems to be of perhaps the greatest importance.

0.4.10 Example: MDL Linear regression

An MDL criterion for the linear regression example is now examined. Given the importance of linear regression, there have been a range of MDL inspired criteria developed, for example, the predictive MDL criterion [97], g-MDL [114], Liang and Barron's approach [115], sequentially normalised least squares [116, 117] and normalized maximum likelihood based methods [107, 36, 118]. We choose to focus on the NML criterion introduced by Rissanen in 2000 [107] as it involves no user specified hyperparameters and does not depend on the ordering of the data, a problem which affects the predictive and sequential procedures. Recall that in the linear regression case, the model index γ specifies which covariates are to be used from the full design matrix $\mathbf{X}$. For a given γ, the NML codelength, up to constants, is

$$\left(\frac{n-k}{2}\right)\log\hat{\tau}(\mathbf{y};\gamma)+\frac{k}{2}\log\left(\frac{\hat{R}(\gamma)}{n}\right)-\log\Gamma\left(\frac{n-k}{2}\right)-\log\Gamma\left(\frac{k}{2}\right)-\log\pi_\gamma(\gamma) \tag{82}$$

where $\hat{\tau}(\mathbf{y};\gamma)$ is the maximum likelihood estimate of τ given by (6),

$$\hat{R}(\gamma)=\hat{\boldsymbol{\beta}}(\mathbf{y};\gamma)'\left(\mathbf{X}_\gamma'\mathbf{X}_\gamma\right)\hat{\boldsymbol{\beta}}(\mathbf{y};\gamma) \tag{83}$$

and $\hat{\boldsymbol{\beta}}(\mathbf{y};\gamma)$ are the least-squares estimates of the model coefficients $\boldsymbol{\beta}$ for subset γ given by (5). As in the case of the BIC given by (43), and the MML linear regression criteria given by (69) and (73,74), the codelength includes a prior $\pi_\gamma(\cdot)$ required to state the subset γ, and suitable choices are discussed in Section 0.3.4.

Interestingly, Hansen and Yu [119] have shown that NML is either asymptotically consistent or asymptotically efficient depending on the nature of the data generating distribution. In this way, the NML criterion can be viewed as a "bridge" between the AIC-based criteria and the BIC. Given the close similarity between the NML and MML linear regression criteria [64], this property is expected to also hold for the MML linear regression criteria.

0.4.11 Applications of MDL

Below is an incomplete list of successful applications of the MDL principle to commonly used statistical models:

- The multinomial distribution [120],

	Criterion	Sample Size							
		25	50	75	100	125	150	200	500
$\hat{p} = p$	AIC_c	86.2	79.5	76.9	76.0	74.7	74.5	74.0	72.3
	KIC_c	93.4	90.8	89.8	89.8	89.3	89.2	88.9	88.6
	BIC	77.5	91.4	94.1	95.7	96.0	97.0	97.4	98.7
	MML_u	93.5	96.4	97.2	97.7	97.7	98.5	98.5	99.2
	MML_g	95.6	97.9	98.4	98.6	98.6	99.1	99.0	99.5
	NML	94.4	96.8	97.4	97.9	97.9	98.7	98.6	99.3
$\hat{p} > p$	AIC_c	13.8	20.5	23.1	24.0	25.3	25.5	26.1	27.7
	KIC_c	6.60	9.30	10.2	10.2	10.7	10.8	11.1	11.4
	BIC	22.5	8.58	5.94	4.30	4.03	2.96	2.60	1.29
	MML_u	6.50	3.64	2.82	2.29	2.32	1.48	1.50	0.78
	MML_g	4.36	2.14	1.57	1.44	1.43	0.94	1.01	0.50
	NML	5.62	3.16	2.57	2.07	2.13	1.30	1.39	0.72
Error	AIC_c	0.52	0.25	0.17	0.13	0.10	0.08	0.06	0.03
	KIC_c	0.47	0.22	0.14	0.11	0.09	0.07	0.05	0.02
	BIC	0.59	0.22	0.14	0.10	0.08	0.07	0.05	0.02
	MML_u	0.47	0.21	0.13	0.10	0.08	0.06	0.05	0.02
	MML_g	0.46	0.21	0.13	0.10	0.08	0.06	0.05	0.02
	NML	0.46	0.21	0.13	0.10	0.08	0.06	0.05	0.02

Table 1: Polynomial order selected by the criteria (expressed as percentages) and squared error in estimated coefficients

- Causal models and Naïve Bayes classification [121, 122],
- Variable bin-width histograms [123, 124],
- Linear regression [107, 114, 118],
- Signal denoising [107, 36],
- Autoregressive models [108],
- Statistical genetics [125, 126].

Of particular interest is the multinomial distribution as it is one of the few models for which the exact parametric complexity is finite without any requirement for bounding, and may be efficiently computed in polynomial time using the clever algorithm discussed in [120]. This algorithm has subsequently been used to compute codelengths in histogram estimation and other similar models.

0.5 Simulation

We conclude with a brief simulation demonstrating the practical application of the distance based, Bayesian and information theoretic model selection criteria discussed in this chapter. The simulation compares the AIC_c (25), KIC_c (27), BIC (42), MML_u (69), MML_g (73,74) and NML (82) criteria on the problem of polynomial order selection with linear regression. Datasets of various sample sizes $25 \leq n \leq 500$ were generated from the polynomial model

$$y^* = x^3 - 0.5x^2 - 5x - 1.5, \quad x \in [-3, 3] \tag{84}$$

with design points x uniformly generated from the compact set $x \in [-3, 3]$ [11, 98]. The noise variance τ was chosen to yield a signal-to-noise ratio of ten. Recalling the linear regression example from Section 0.1.2, the complete design matrix consisted of polynomial bases x^i for $(i = 0, \ldots, q)$,

$$\mathbf{X} = (\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^q). \tag{85}$$

For each generated data set, each of the six criteria were used to select a nested polynomial model up to the maximum polynomial of order $q = 20$. The simulation was repeated 10^4 times and the zero-order polynomial model was not considered. The criteria were compared on two important metrics: (i) order selection, and (ii) squared prediction error. The results are presented in Table 1.

The simulation is highly illustrative of the general behaviours of the model selection criteria under consideration. For small sample sizes ($n \leq 50$), the KIC_c criterion performs adequately in terms of order selection, but like the AIC_c, this performance does not improve with increasing sample size. This is not surprising given that both AIC_c and KIC_c are not consistent model selection procedures. While BIC performs relatively poorly for the smallest sample sizes, its performance dramatically improves when the sample size is larger. The asymptotic approximation from which BIC is derived is inadequate for small sample sizes which explains its poor performance. In contrast, the MML and NML criteria perform well for all sample sizes in terms of both order selection and prediction error. In this particular example, the MML_u and NML criteria are virtually indistinguishable, while the MML_g criterion performs slightly better. For larger sample sizes ($n \geq 125$) all four consistent criteria performed essentially the same.

This brief simulation serves to demonstrate the general behaviour of the six criteria and should not be taken as indicative of their performance in other model selection problems. For a more detailed comparison of some recent regression methods, including MML_u, MML_g and NML, see [118]. The MML_u, MML_g and NML criteria were found to be amongst the best of the eight criteria considered in all experiments conducted by the authors [3].

[3] Unfortunately, through no fault of the authors, the formulae for MML_g given in [118] is missing the $\delta/2$ term present in the formula (73) given in this tutorial. However, this has not effected the experimental results as the simulation code uses the correct formula (see Section 0.4.4).

Bibliography

[1] E. L. Lehmann, G. Casella, Theory of Point Estimation, Springer Texts in Statistics, Springer, 4th edition, 2003.

[2] S. Kullback, R. A. Leibler, The Annals of Mathematical Statistics 22 (1951) 79–86.

[3] H. Linhart, W. Zucchini, Model Selection, Wiley, New York, 1986.

[4] H. Akaike, IEEE Transactions on Automatic Control 19 (1974) 716–723.

[5] C. M. Hurvich, C.-L. Tsai, Biometrika 76 (1989) 297–307.

[6] A.-K. Seghouane, S.-I. Amari, IEEE Transactions on Neu 18 (2007) 97–106.

[7] A.-K. Seghouane, Signal Processing 90 (2010) 217–224.

[8] A. D. R. McQuarrie, C.-L. Tsai, Regression and Time Series Model Selection, World Scientific, 1998.

[9] H. Jeffreys, Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences 186 (1946) 453–461.

[10] J. E. Cavanaugh, Statistics & Probability Letters 42 (1999) 333–343.

[11] A.-K. Seghouane, M. Bekara, IEEE Transactions on Signal Processing 52 (2004) 3314–3323.

[12] D. F. Schmidt, E. Makalic, Lecture Notes in Artificial Intelligence 6464 (2010) 223–232.

[13] A. K. Seghouane, Signal Processing 86 (2006) 2074–2084.

[14] A.-K. Seghouane, IEEE Transactions on Aerospace and Electronic Systems 47 (2011) 1154–1165.

[15] C. M. Hurvich, J. S. Simonoff, C.-L. Tsai, Journal of the Royal Statistical Society (Series B) 60 (1998) 271–293.

[16] M. Bekara, L. Knockaert, A.-K. Seghouane, G. Fleury, Signal Processing 86 (2006) 1400–1409.

[17] C. M. Hurvich, C.-L. Tsai, Biometrika 85 (1998) 701–710.

[18] N. Murata, S. Yoshizawa, S. Amari, IEEE Transactions on Neural Networks 5 (1994) 865–872.

[19] C. Robert, The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, Springer Texts in Statistics, Springer, 2001.

[20] J. Berger, L. Pericchi, in: P. Lahiri (Ed.), Model Selection, volume 38 of *Lecture Notes Monograph Series*, Hayward, CA:, pp. 135–207.

[21] J. M. Bernardo, Journal of the Royal Statistical Society (Series B) 41 (1979) 113–147.

[22] J. Berger, L. Pericchi, Journal of the American Statistical Association 91 (1996) 109–122.

[23] C. S. Wallace, Statistical and Inductive Inference by Minimum Message Length, Information Science and Statistics, Springer, first edition, 2005.

[24] G. Casella, E. Moreno, Journal of the American Statistical Association 101 (2006) 157–167.

[25] H. Jeffreys, The theory of probability, Oxford University Press, 3rd edition, 1961.

[26] R. E. Kass, A. E. Raftery, Journal of the American Statistical Association 90 (1995) 773–795.

[27] G. Schwarz, The Annals of Statistics 6 (1978) 461–464.

[28] J. Rissanen, Automatica 14 (1978) 465–471.

[29] C. S. Wallace, D. M. Boulton, Computer Journal 11 (1968) 185–194.

[30] O. E. Barndorff-Nielsen, D. R. Cox, Asymptotic Techniques for Use in Statistics, New York: Chapman & Hall, 1989.

[31] D. M. A. Haughton, The Annals of Statistics 16 (1988) 342–355.

[32] C. R. Rao, Y. Wu, Biometrika 76 (1989) 369–374.

[33] L. Breiman, The Annals of Statistics 24 (1996) 2350–2383.

[34] D. Madigan, A. E. Raftery, Journal of the American Statistical Association 89 (1994) 1536–1546.

[35] E. I. George, D. P. Foster, Biometrika 87 (2000) 731–747.

[36] T. Roos, P. Myllymäki, J. Rissanen, IEEE Transactions on Signal Processing 57 (2009) 3347–3360.

[37] J. G. Scott, J. O. Berger, The Annals of Statistics 38 (2010) 2587–2619.

[38] J. Chen, Z. Chen, Biometrika 95 (2008) 759–771.

[39] C. P. Robert, G. Casella, Monte Carlo Statistical Methods, Springer, 2004.

[40] A. Metropolis, M. Rosenbluth, A. Rosenbluth, E. Teller, Journal of Chemical Physics 21 (1953) 1087–1092.

[41] W. Hastings, Biometrika 57 (1970) 97–109.

[42] S. Geman, D. Geman, IEEE Trans. Pattern Anal. Mach. Int. 6 (1984) 721–741.

[43] G. Casella, E. I. George, The American Statistician 46 (1992) 167–174.

[44] S. Chib, Journal of the American Statistical Association 90 (1995) 1313–1321.

[45] P. J. Green, Biometrika 82 (1995) 711–732.

[46] E. I. George, R. E. McCulloch, The Journal of the American Statistical Assocation 88 (1993) 881–889.

[47] C. Hans, A. Dobra, M. West, Journal of the American Statistical Association 102 (2007) 507–516.

[48] L. Breiman, Technometrics 37 (1995) 373–384.

[49] R. Tibshirani, Journal of the Royal Statistical Society (Series B) 58 (1996) 267–288.

[50] M. E. Tipping, Journal of Machine Learning Research 1 (2001) 211–244.

[51] R. M. Neal, Bayesian learning for neural networks, Lecture Notes in Statistics, Springer Verlag, 1996.

[52] T. Park, G. Casella, Journal of the American Statistical Association 103 (2008) 681–686.

[53] C. Hans, Biometrika 96 (2009) 835–845.

[54] T. M. Cover, J. A. Thomas, Elements of Information Theory, Wiley-Interscience, second edition, 2006.

[55] C. E. Shannon, Bell System Technical Journal 27 (1948) 379–423 and 623–656.

[56] J. Rissanen, J. G. G. Langdon, IEEE Transactions on Information Theory IT-27 (1981) 12–23.

[57] P. D. Grünwald, The Minimum Description Length Principle, Adaptive Communication and Machine Learning, The MIT Press, 2007.

[58] C. Wallace, D. Boulton, Classification Society Bulletin 3 (1975) 11–34.

[59] C. S. Wallace, P. R. Freeman, Journal of the Royal Statistical Society (Series B) 49 (1987) 240–252.

[60] C. S. Wallace, in: Proceedings of the International Conference on Information, Statistics and Induction in Science, World Scientific, 1996, pp. 304–316.

[61] G. E. Farr, C. S. Wallace, Computer Journal 45 (2002) 285–292.

[62] J. H. Conway, N. J. A. Sloane, Sphere Packing, Lattices and Groups, Springer-Verlag, third edition, 1998.

[63] E. Makalic, D. F. Schmidt, Statistics & Probability Letters 79 (2009) 1155–1161.

[64] D. Schmidt, E. Makalic, in: The 22nd Australasian Joint Conference on Artificial Intelligence, Melbourne, Australia, pp. 312–321.

[65] C. S. Wallace, P. R. Freeman, Journal of the Royal Statistical Society (Series B) 54 (1992) 195–209.

[66] C. S. Wallace, D. Dowe, MML estimation of the von Mises concentration parameter, Technical report, Department of Computer Science, Monash University, 1993.

[67] D. L. Dowe, J. Oliver, C. Wallace, Lecture Notes in Artificial Intelligence 1160 (1996) 213–227.

[68] D. F. Schmidt, Minimum Message Length Inference of Autoregressive Moving Average Models, Ph.D. thesis, Clayton School of Information Technology, Monash University, 2008.

[69] A. R. Barron, T. M. Cover, IEEE Transactions on Information Theory 37 (1991) 1034–1054.

[70] L. LeCam, University of California Publications in Statistics 11 (1953).

[71] D. L. Dowe, C. S. Wallace, in: Proc. 28th Symposium on the interface, volume 28 of *Computing Science and Statistics*, Sydney, Australia, pp. 614–618.

[72] C. S. Wallace, in: Proceedings of the Fourteenth Biennial Statistical Conference, Queenland, Australia, p. 144.

[73] M. Viswanathan, C. S. Wallace, D. L. Dowe, K. B. Korb, Lecture Notes in Artificial Intelligence 1747 (1999) 405–416.

[74] C. S. Wallace, D. L. Dowe, Statistics and Computing 10 (2000) 73–83.

[75] J. J. Oliver, R. A. Baxter, MML and Bayesianism: Similarities and differences, Technical Report TR 206, Department of Computer Science, Monash University, 1994.

[76] D. F. Schmidt, in: Proc. 5th Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-11).

[77] L. J. Fitzgibbon, D. L. Dowe, L. Allison, in: Proceedings of the Nineteenth International Conference on Machine Learning (ICML'02), pp. 147–154.

[78] D. L. Dowe, The Computer Journal 51 (2008) 523–560.

[79] E. Makalic, Minimum Message Length Inference of Artificial Neural Networks, Ph.D. thesis, Clayton School of Information Technology, Monash University, 2007.

[80] E. Makalic, L. Allison, in: Proceedings of the 85th Solomonoff Memorial Conference, Melbourne, Australia.

[81] M. Viswanathan, C. Wallace, in: Proc. 7th Int. Workshop on Artif. Intelligence and Statistics, Ft. Lauderdale, Florida, U.S.A., pp. 169–177.

[82] G. W. Rumantir, C. S. Wallace, in: Proceedings of the 5th International Symposium on Intelligent Data Analysis (IDA 2003), volume 2810, Springer-Verlag, Berlin, Germany, pp. 486–496.

[83] W. James, C. M. Stein, in: Proceedings of the Fourth Berkeley Symposium, volume 1, University of California Press, pp. 361–379.

[84] S. L. Sclove, Journal of the American Statistical Association 63 (1968) 596–606.

[85] N. Bouguila, D. Ziou, IEEE Transactions on Knowledge and Data Engineering 18 (2006) 993–1009.

[86] C. S. Wallace, J. D. Patrick, Machine Learning 11 (1993) 7–22.

[87] P. Tan, D. Dowe, Lecture Notes in Artificial Intelligence 2903 (2003) 269–281.

[88] P. Tan, D. Dowe, Lecture Notes in Artificial Intelligence 4293 (2006) 593–603.

[89] C. S. Wallace, K. B. Korb, in: A. Gammerman (Ed.), Causal Models and Intelligent Data Management, Springer-Verlag, pp. 89–111.

[90] J. W. Comley, D. Dowe, Minimum Message Length and Generalized Bayesian Nets with Asymmetric Languages, M.I.T. Press (MIT Press), pp. 265–294.

[91] E. Makalic, L. Allison, D. L. Dowe, in: Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2003).

[92] E. Makalic, L. Allison, A. P. Paplinski, in: Proceedings of the 8th Brazillian Symposium on Neural Networks (SBRN 2004), Sao Luis, Maranhao, Brazil.

[93] D. F. Schmidt, E. Makalic, Shrinkage and Denoising by Minimum Message Length, Technical Report 2008/230, Monash University, 2008.

[94] L. J. Fitzgibbon, D. L. Dowe, F. Vahid, in: Proceedings of the International Conference on Intelligent Sensing and Information Processing (ICISIP), pp. 439–444.

[95] D. F. Schmidt, in: Proceedings of the 85th Solomonoff Memorial Conference, Melbourne, Australia.

[96] R. A. Baxter, J. J. Oliver, Lecture Notes in Artificial Intelligence 1160 (1996) 83–90.

[97] J. Rissanen, Stochastic Complexity in Statistical Inquiry, World Scientific, 1989.

[98] J. Rissanen, Information and Complexity in Statistical Modeling, Information Science and Statistics, Springer, first edition, 2007.

[99] J. Rissanen, Circuits, Systems, and Signal Processing 1 (1982) 395–396.

[100] J. Rissanen, The Annals of Statistics 11 (1983) 416–431.

[101] R. A. Baxter, J. Oliver, MDL and MML: Similarities and Differences, Technical Report TR 207, Department of Computer Science, Monash University, 1994.

[102] J. Rissanen, IEEE Transactions on Information Theory 42 (1996) 40–47.

[103] J. Rissanen, IEEE Transactions on Information Theory 47 (2001) 1712–1717.

[104] Y. M. Shtarkov, Probl. Inform. Transm. 23 (1987) 3–17.

[105] V. Balasubramanian, in: I. J. M. P. D. Grünwald, M. A. Pitt (Eds.), Advances in Minimum Description Length: Theory and Applications, MIT Press, pp. 81–99.

[106] S. de Rooij, P. Grünwald, Journal of Mathematical Psychology 50 (2006) 180–192.

[107] J. Rissanen, IEEE Transactions on Information Theory 46 (2000) 2537–2543.

[108] D. F. Schmidt, E. Makalic, IEEE Transactions on Signal Processing 59 (2011) 479–487.

[109] A. P. Dawid, Journal of the Royal Statistical Society (Series A) 147 (1984) 278–292.

[110] T. Roos, J. Rissanen, in: Proc. 1st Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08), Tampere International Center for Signal Processing. (Invited Paper).

[111] J. Rissanen, T. Roos, in: Proc. 2007 Information Theory and Applications Workshop (ITA-07),, IEEE Press, 2007, pp. 337–341. (Invited Paper).

[112] D. F. Schmidt, E. Makalic, IEEE Transactions on Information Theory 55 (2009) 3087–3090.

[113] D. F. Schmidt, E. Makalic, IEEE Transactions on Signal Processing 60 (2012) 1508–1510.

[114] M. H. Hansen, B. Yu, Journal of the American Statistical Association 96 (2001) 746–774.

[115] F. Liang, A. Barron, IEEE Transactions on Information Theory 50 (2004) 2708–2726.

[116] J. Rissanen, T. Roos, P. Myllymäki, Journal of Multivariate Analysis 101 (2010) 839–849.

[117] D. F. Schmidt, T. Roos, in: Proc. 3rd Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-10), Tampere International Center for Signal Processing. (invited paper).

[118] C. D. Giurcăneanu, S. A. Razavi, A. Liski, Signal Processing (2011).

[119] M. Hansen, B. Yu, in: Science and Statistics: A Festchrift for Terry Speed, volume 40 of *Lecture Notes - Monograph Series*, Institute of Mathematical Statistics, pp. 145–164.

[120] P. Kontkanen, P. Myllymäki, Information Processing Letters 103 (2007) 227–233.

[121] T. Mononen, P. Myllymaki, in: Proceedings of the 10th International Conference on Discovery Science, volume 4755 of *Lecture Notes in Computer Science*, Sendai, Japan, pp. 151–160.

[122] T. Silander, T. Roos, P. Myllymäki, in: Proc. 12th International Conference on Artificial Intelligence and Statistics (AISTATS-09).

[123] J. Rissanen, T. P. Speed, B. Yu, IEEE Transactions on Information Theory 38 (1992) 315–323.

[124] P. Kontkanen, P. Myllymäki, in: M. Meila, X. Shen (Eds.), Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS 2007), San Juan, Puerto Rico.

[125] Y. Yang, I. Tabus, in: IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS 2007), pp. 1–4.

[126] I. Tabus, J. Rissanen, J. Astola, Signal Processing 83 (2003) 713–727.