The Behaviour of the Akaike Information Criterion When Applied to Non-nested Sequences of Models

Daniel Francis Schmidt and Enes Makalic

The University of Melbourne
Centre for MEGA Epidemiology
Carlton VIC 3053, Australia
{dschmidt,emakalic}@unimelb.edu.au

Abstract. A typical approach to the problem of selecting between models of differing complexity is to choose the model with the minimum Akaike Information Criterion (AIC) score. This paper examines a common scenario in which there is more than one candidate model with the same number of free parameters which violates the conditions under which AIC was derived. The main result of this paper is a novel upper bound that quantifies the poor performance of the AIC criterion when applied in this setting. Crucially, the upper-bound does not depend on the sample size and will not disappear even asymptotically. Additionally, an AIC-like criterion for sparse feature selection in regression models is derived, and simulation results in the case of denoising a signal by wavelet thresholding demonstrate the new AIC approach is competitive with *SureShrink* thresholding.

1 Introduction

Every day thousands of researchers use the celebrated Akaike Information Criterion (AIC) [1] as a guide for selecting features when building models from observed data. Perhaps the most canonical example is the use of AIC to determine which features (covariates) to include in a multiple regression, which forms, for example, the basis of epidemiological and medical statistics. The AIC was derived under the assumption that the set of models under consideration (the candidate models) forms a strictly nested sequence; that is, the more complex models completely contain all of the simpler models. If we measure a model's "complexity" by the number of free parameters it possesses, a necessary (but not sufficient) requirement for this assumption to hold is that each of the candidate models possesses a unique number of free parameters.

A classic example in which this assumption is violated is subset selection of regression models; if we include all possible subsets of q features in our set of candidate models, there will be $\binom{q}{k}$ different models with exactly k free parameters. It is clear that if the number of features, q, we are considering is large then the number of models with the same number of parameters in the candidate set can be enormous.

J. Li (Ed.): AI 2010, LNAI 6464, pp. 223–232, 2010.

While the poor performance of AIC when applied to non-nested sequences of models has been noted in the literature (see for example, [2]), there appears to have been no attempts to formally quantify just how badly the AIC may perform. The primary contribution of this paper is to remedy this situation by providing a novel asymptotic upper bound quantifying the extent to which AIC may deviate from the quantity it is attempting to estimate in the setting of non-nested sequences of models. The most interesting, and worrying, finding is that the upper bound depends crucially on the maximum number of models being considered, and in the limit as the sample size $n \to \infty$ the upper bound does not converge to the usual AIC score. This implies the following critical conclusion: *that the poor performance of AIC when applied to non-nested sequences of models cannot be overcome even by obtaining large amounts of data* – the problem is tied fundamentally to the confluence of models rather than sample size. We believe this is a very important discovery with profound effects on the way the AIC should be employed in the research community.

2 Akaike's Information Criterion

The problem that the Akaike Information Criterion aims to solve is the following: we have observed n samples $\mathbf{y} = (y_1, \ldots, y_n)$ and wish to learn something about the process that generated the data. In particular, we have a set of candidate models of differing complexity which we may fit to the data. If we choose too simple a model then the predictions of future data will be affected by the bias present due to the limitations of the model; in contrast, if we choose an overly complex model then the increased variance in the parameter estimates will lead to poor predictions. The AIC aims to select the model from the candidate set that best trades off these two sources of error to give good predictions.

2.1 Models and Nested Model Sequences

It is impossible to discuss the properties of AIC and its problems when applied to non-nested sequence of models without first defining some notation. We let $\gamma \in \Gamma$ denote a statistical model, with $\boldsymbol{\theta}_\gamma \in \Theta_\gamma$ denoting a parameter vector for the model γ and Γ denoting the set of all candidate models. A statistical model γ indexes a set of parametric probability distributions over the data space; denote this by $p(\mathbf{y}|\boldsymbol{\theta}_\gamma)$. The parameter vector $\boldsymbol{\theta}_\gamma \in \Theta_\gamma$ indexes a particular distribution within the model γ. The number of free parameters possessed by a model γ (or equivalently, the dimensionality of $\boldsymbol{\theta}_\gamma$) is denoted by k_γ.

Using this notation, we can now introduce the notion of a "true" model and a "true" distribution. The true distribution is the particular distribution in the true model that generated the observed data $\mathbf{y}$. Let γ^* denote the true model, and $\boldsymbol{\theta}_*$ denote the parameter vector that indexes the true distribution. Using the shorthand notation that $p_{\boldsymbol{\theta}_\gamma}$ denotes the distribution indexed by $\boldsymbol{\theta}_\gamma$ in the model γ, we can say that $\mathbf{y} \sim p_{\boldsymbol{\theta}^*_{\gamma^*}}$.

In the context of AIC the idea of a *nested sequence* of models is very important. If a set of models form a nested sequence then they possess the special

property that a model with k free parameters can represent all of the distributions contained in all models with less than k parameters; usually, this involves setting some of the parameters to zero, though this is neither universally the case, nor a requirement. The following are two important properties possessed by nested sequences of models.

Property 1. Each of the models in a nested sequence of models has a unique number of free parameters.

Property 2. If the true model γ^* is part of a nested sequence of models, then for all models γ with $k_\gamma > k_{\gamma^*}$ (i.e., with more free parameters) there is a parameter vector $\boldsymbol{\theta}_\gamma \in \Theta_\gamma$ that indexes the same distribution as the "true" distribution $p_{\boldsymbol{\theta}_*}$.

Let this parameter vector be denoted by the symbol $\boldsymbol{\theta}^*_\gamma$.

In words, this says that if the true distribution can be represented by the model in the nested sequence with k parameters, then it can also be exactly represented by all the models with more than k parameters. Thus, the "true" model is simply the model with the least number of parameters that can represent the true distribution. An example will illustrate the concepts presented in this section.

Example: Polynomial Models. Consider the class of normal regression models, where the mean is specified by a polynomial of degree k. If the maximum degree is q, the model class index $\gamma \in \{0, 1, \ldots, q\}$ denotes the degree of the polynomial; i.e., $\gamma = k$ specifies a polynomial of the form

$$y = a_0 + a_1 x + a_2 x^2 + \ldots a_k x^k + \varepsilon$$

with ε normally distributed with variance τ. The polynomial model indexed by γ has $k_\gamma = \gamma + 2$ free parameters (including the noise variance) given by $\boldsymbol{\theta}_\gamma = (a_0, \ldots, a_\gamma, \tau)$, with the parameter space $\Theta_\gamma = \mathbb{R}^{k+1} \times \mathbb{R}_+$. The models form a *nested sequence* as a polynomial of degree k can represent any polynomial of degree $j < k$ by setting $a_{j+1}, \ldots, a_k = 0$; for example, a quintic polynomial can represent a cubic polynomial by setting $a_4 = a_5 = 0$.

2.2 Model Fitting and Goodness of Fit

There are many ways of fitting a model γ to the observed data (often called "point estimation"); a powerful and general procedure is called *maximum likelihood* (ML), and it is this process that is integral to the derivation of the AIC. Maximum likelihood fitting simply advocates choosing the parameter vector $\boldsymbol{\theta}_\gamma$ for a chosen model γ such that the probability of observed data $\mathbf{y}$ is maximised

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

For a model selection criterion to be useful it must aim to select a model from the candidate set that is close, in some sense, to the truth. In order to measure

how close the fitted approximating model $\hat{\boldsymbol{\theta}}_\gamma$ is to the generating distribution $\boldsymbol{\theta}_*$, one requires a distance measure between probability densities. A commonly used measure of distance between two models, say $\boldsymbol{\theta}_*$ and $\hat{\boldsymbol{\theta}}_\gamma$ is the directed Kullback–Leibler (K–L) divergence [3], given by

$$\Delta(\boldsymbol{\theta}_* || \hat{\boldsymbol{\theta}}_\gamma) = \mathrm{E}_{\boldsymbol{\theta}_*} \left[\log \frac{p(\mathbf{y}|\boldsymbol{\theta}_*)}{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma)} \right] \tag{2}$$

where the expectation is taken with respect to $\mathbf{y} \sim p_{\boldsymbol{\theta}_*}$. The directed K–L divergence is non-symmetric and strictly positive for all $\hat{\boldsymbol{\theta}}_\gamma \neq \boldsymbol{\theta}_*$. Defining the function

$$d(\boldsymbol{\theta}_*, \hat{\boldsymbol{\theta}}_\gamma) = 2\mathrm{E}_{\boldsymbol{\theta}_*} \left[\log 1/p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) \right] \tag{3}$$

the K–L divergence may be written as

$$2\Delta(\boldsymbol{\theta}_* || \hat{\boldsymbol{\theta}}_\gamma) = d(\boldsymbol{\theta}_*, \hat{\boldsymbol{\theta}}_\gamma) - d(\boldsymbol{\theta}_*, \boldsymbol{\theta}_*) \tag{4}$$

The first term on the right hand side of (4) is generally known as the cross-entropy between $\boldsymbol{\theta}_*$ and $\hat{\boldsymbol{\theta}}_\gamma$, while the second is known as the entropy of $\boldsymbol{\theta}_*$. The use of the Kullback–Leibler divergence can be justified by both its invariance to the parameterisation of the models (as opposed to Euclidean distance, for example) as well as its connections to information theory.

2.3 Akaike's Information Criterion

Ideally, one would rank the candidate models in ascending order based on their K–L divergence from the truth, and select the model with the smallest K–L divergence as optimal. However, this procedure requires knowledge of the true model and is thus not feasible in practice. Even though the truth is not known, *one may attempt to construct an estimate of the K–L divergence based solely on the observed data*. This idea was first explored by Akaike in his groundbreaking paper [1] in the particular case of a nested sequence of candidate models. Akaike noted that the negative log-likelihood serves as a downwardly biased estimate of the average cross entropy (the cross-entropy risk), and subsequently derived an asymptotic bias correction. The resulting Akaike Information Criterion (AIC) advocates choosing a model, from a nested sequence of models, that minimises

$$\mathrm{AIC}(\gamma) = 2 \log 1/p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\gamma) + 2k_\gamma \tag{5}$$

where $\hat{\boldsymbol{\theta}}_\gamma$ is the maximum likelihood estimator for the model γ and the second term is the bias correction. Under suitable regularity conditions [4], and assuming that the fitted model γ is at least as complex as the truth (i.e., the true distribution is contained in the distributions indexed by the model γ), the AIC statistic can be shown to satisfy

$$\mathrm{E}_{\boldsymbol{\theta}^*} \left[\mathrm{AIC}(\gamma) \right] = \mathrm{E}_{\boldsymbol{\theta}^*} \left[d(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_\gamma) \right] + o_n(1) \tag{6}$$

where $o_n(1)$ denotes a term that vanishes as the sample size $n \to \infty$. In words, (6) states that the AIC statistic is, up to a constant, an unbiased estimator of twice the Kullback–Leibler risk (average Kullback–Leibler divergence from the truth) attained by a particular model γ; that is, for sufficiently large sample sizes, the AIC score is *on average* equal to the average cross–entropy between the truth and the maximum likelihood estimate for the fitted model γ. Although the AIC estimates the cross–entropy risk rather than the complete Kullback–Leibler risk, the omitted entropy term $d(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)$ does not depend on the fitted model γ and will thus have no effect on the ranking of models by their AIC scores. The selection of a candidate model using AIC is therefore equivalent to choosing one with the lowest estimated Kullback–Leibler risk.

In the case of non-nested model sequences, the number of candidate models with k parameters may be greater than one and the downward bias of the negative log-likelihood is greater than the AIC model structure penalty. Problematically, this extra source of additional bias remains even as the sample size $n \to \infty$. The next section derives a novel upper-bound on this additional bias under certain conditions.

3 The Bias in AIC for Multiple Selection

The main result of this paper is an expression for the additional downward bias that is introduced when $q_k > 1$. Let

$$\Gamma_k = \{\gamma \in \Gamma : k_\gamma = k\}$$

denote the set of all candidate models with k parameters, with $q_k = |\Gamma_k|$ being the number of candidate models with k parameters. In the case of a nested sequence of models, $q_k = 1$ for all k. Then, let

$$\hat{m}_k = \underset{m \in \Gamma_k}{\arg\min} \left\{ \log 1/p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m) \right\} \tag{7}$$

denote the candidate model with k parameters with the smallest negative log-likelihood. We can now recast the model selection problem as one of selecting between the *best* of the k parameter models, i.e. we limit our candidates to the new set of L fitted models

$$\Gamma' = \left\{ \hat{\boldsymbol{\theta}}_{\hat{m}_1}, \ldots, \hat{\boldsymbol{\theta}}_{\hat{m}_L} \right\} \tag{8}$$

Assuming the following holds

1. The true model γ^* has no free parameters
2. All candidate models $\gamma \in \Gamma$ contain the true distribution $p_{\boldsymbol{\theta}_*}$ as a particular element, i.e., *all candidate models are overfitting*. Let the parameter vector that indexes the true distribution for the model γ be denoted by $\boldsymbol{\theta}^*_\gamma$
3. The maximum likelihood estimator converges to the truth, $\hat{\boldsymbol{\theta}}_\gamma \to \boldsymbol{\theta}^*_\gamma$ as $n \to \infty$, and is asymptotically normally distributed, $\hat{\boldsymbol{\theta}}_\gamma \sim N(\boldsymbol{\theta}^*_\gamma, \mathbf{J}^{-1}(\boldsymbol{\theta}^*_\gamma))$

4. All candidate models of k parameters are independent; that is,

$$\log \frac{p(\mathbf{y}|\boldsymbol{\theta}_*)}{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m)}, \quad m \in \Gamma_k$$

are independent random variates.

Theorem 1: Under the above conditions we have

$$2\mathrm{E}_{\boldsymbol{\theta}^*}\left[\log 1/p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\hat{m}_k})\right] + 2\alpha(k, q_k) = \mathrm{E}_{\boldsymbol{\theta}_*}\left[d(\boldsymbol{\theta}_*, \hat{\boldsymbol{\theta}}_{\hat{m}_k})\right] + o_n(1) \tag{9}$$

where

$$\alpha(k, q_k) = \mathrm{E}_{\chi^2_k}\left[\max\{z_1, \ldots, z_{q_k}\}\right] \tag{10}$$

and $z_1, \ldots, z_{q_k}$ are independently and identically distributed χ^2_k variates with k degrees of freedom.

Proof: Following the procedure in [5] the cross-entropy risk can be written

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}_*}\left[d(\boldsymbol{\theta}_*, \hat{\boldsymbol{\theta}}_m)\right] = \mathrm{E}_{\boldsymbol{\theta}_*}\left[d(\boldsymbol{\theta}_*, \hat{\boldsymbol{\theta}}_m)\right] - d(\boldsymbol{\theta}_*, \boldsymbol{\theta}_*) \\ + d(\boldsymbol{\theta}_*, \boldsymbol{\theta}_*) - 2\mathrm{E}_{\boldsymbol{\theta}_*}\left[\log 1/p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m)\right] \\ + 2\mathrm{E}_{\boldsymbol{\theta}_*}\left[\log 1/p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m)\right]\end{aligned} \tag{11}$$

From regularity conditions the following approximations hold

$$2\log 1/p(\mathbf{y}|\boldsymbol{\theta}_*) + 2\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m) = (\boldsymbol{\theta}^*_m - \hat{\boldsymbol{\theta}}_m)'\mathbf{H}(\hat{\boldsymbol{\theta}}_m, \mathbf{y})(\boldsymbol{\theta}^*_m - \hat{\boldsymbol{\theta}}_m) + o(k) \tag{12}$$

$$d(\boldsymbol{\theta}_*, \hat{\boldsymbol{\theta}}_m) - d(\boldsymbol{\theta}_*, \boldsymbol{\theta}_*) = (\boldsymbol{\theta}^*_m - \hat{\boldsymbol{\theta}}_m)'\mathbf{J}(\boldsymbol{\theta}^*_m)(\boldsymbol{\theta}^*_m - \hat{\boldsymbol{\theta}}_m) + o(k) \tag{13}$$

where

$$\mathbf{H}(\hat{\boldsymbol{\theta}}_m, \mathbf{y}) = \left[\left.\frac{\partial^2 \log 1/p(\mathbf{y}|\boldsymbol{\theta}_m)}{\partial\boldsymbol{\theta}_m \partial\boldsymbol{\theta}'_m}\right|_{\boldsymbol{\theta}_m = \hat{\boldsymbol{\theta}}_m}\right], \quad \mathbf{J}(\boldsymbol{\theta}^*) = \left[\left.\frac{\partial^2 \Delta(\boldsymbol{\theta}^*, \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^*}\right]$$

are the observed and expected Fisher information matrices respectively. Denote the right hand side of (12) and (13) by a_m and b_m respectively. The first term, a_m, is twice the decrease in the negative log-likelihood due to fitting a model $\hat{\boldsymbol{\theta}}_m$, and the second term, b_m, is twice the K–L divergence between the generating model $\boldsymbol{\theta}^*$ and the fitted model $\hat{\boldsymbol{\theta}}_m$. Since there are q_k models with k parameters, there are q_k random variables a_m and b_m.

Selecting the model with k parameters that minimises the negative log-likelihood is equivalent to solving

$$\hat{m}_k = \arg\max_{m \in \Gamma_k}\{a_m\}$$

Then we have

$$2\mathrm{E}_{\boldsymbol{\theta}^*}\left[\log 1/p(\mathbf{y}|\boldsymbol{\theta}^*) - \log 1/p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\hat{m}_k})\right] = \mathrm{E}_{\boldsymbol{\theta}^*}\left[a_{\hat{m}_k}\right] + o_n(1) \tag{14}$$

$$\mathrm{E}_{\boldsymbol{\theta}^*}\left[d(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_{\hat{m}_k}) - d(\boldsymbol{\theta}^*, \boldsymbol{\theta}^*)\right] = \mathrm{E}_{\boldsymbol{\theta}^*}\left[b_{\hat{m}_k}\right] + o_n(1) \tag{15}$$

For large n, the random variables satisfy $a_m = b_m + o_n(1)$ and therefore coincide. From the properties of the maximum likelihood estimator $\mathbf{H}(\hat{\boldsymbol{\theta}}_m, \mathbf{y}) \to \mathbf{J}(\boldsymbol{\theta}^*_m)$ as $n \to \infty$, rendering the quadratic forms in (12) and (13) identical. Furthermore, a_m converge to centrally distributed χ^2_k variates with k degrees of freedom. Thus,

$$\mathrm{E}_{\boldsymbol{\theta}^*}\left[a_{\hat{m}_k}\right] = \mathrm{E}\left[\max\{z_1, \ldots, z_{q_k}\}\right] \tag{16}$$

where $z_1, \ldots, z_{q_k}$ are independently and identically distributed χ^2_k variates with k degrees of freedom, with an identical expression for $\mathrm{E}_{\boldsymbol{\theta}^*}\left[b_{\hat{m}_k}\right]$. Substituting these expectations into the expression for $\mathrm{E}_{\boldsymbol{\theta}^*}[d(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_{\hat{m}_k})]$ given by (11) completes the proof. □

4 Discussion and Impact

We now discuss the impact of Theorem 1. In words, the result states that if we consider more than one candidate model with the same number of parameters, say k, then the usual AIC complexity penalty of $2k$ (or alternatively, the bias correction) for these models will be insufficient. A further negative result is that under the above conditions, the required bias correction depends on the number of models with k parameters, q_k, and k, but not on the sample size n, and will not disappear even as $n \to \infty$. The primary effect an underestimation of bias will have in practice is to lead to an increased probability of overfitting.

As an example, consider the situation in which the "true" model, γ^*, has no free parameters, and we are considering as alternatives, based on regular AIC scores, a set of $q_1 \geq 1$ "independent" models with one free parameter. In the usual case of a nested sequence of models $q_1 = 1$, and noting that twice the difference in log-likelihoods between the fit of γ^* and the alternative one parameter model is approximately χ^2_1 distributed, we determine that AIC has approximately a 16% probability of eroneously preferring the one parameter model (overfitting). This probability will increase with increasing q_k: using the results of Theorem 1, we see that if $q_k > 1$ then twice the difference in negative log-likelihoods between the initial model we fit, γ_*, and the best of the one parameter models, $\gamma_{\hat{m}_1}$, is distributed as per the maximum of q_1 χ^2_1 variates with one degree of freedom. Using standard results on distributions of order statistics [6], we can compute the probability of overfitting in this scenario for various values of q_1; these are summarised in Table 1. It is clear that even if we consider only four models with $k = 1$ parameters, the probability of overfitting is almost one half, and that it rapidly rises towards one as q_1 increases. This demonstrates just how poorly regular AIC may perform when applied to a non-nested sequence of models.

Table 1. Probability of AIC overfitting by one parameter for various values of q_1

q_1	1	2	3	4	5	8	10	15	25	50	100
P(overfit)	0·157	0·290	0·402	0·496	0·575	0·746	0·819	0·923	0·986	0·999	1·000

4.1 Theorem 1 as an Upper Bound

The most restrictive assumption used by Theorem 1 is the requirement that the models be "independent" (Assumption 4 in Section 3). For many models, this will not be the case; a simple example is "all subsets" regression models, where many of the subsets of two or more features will have several features in common. If one feature is strongly associated with the target, then all subsets containing this feature will reduce the negative log-likelihood by a similarly large amount, i.e., the a_m variates from Theorem 1 will be correlated . However, even in the case that Assumption 4 is violated, the result of Theorem 1 offers a novel *upper bound*: noting that if $\{w_1, \ldots, w_q\}$ are q correlated variates and $\{z_1, \ldots, z_q\}$ are uncorrelated variates, with both sets of variates possessing the same marginal distribution, then

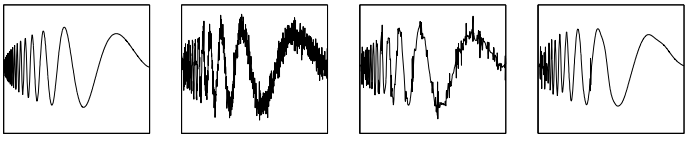
$$\mathrm{E}\left[\max\left\{w_1, \ldots, w_q\right\}\right] < \mathrm{E}\left[\max\left\{z_1, \ldots, z_q\right\}\right]$$

so that the asymptotic bias correction term in this case will be less than $2\alpha(k, q)$. Thus, the result in Theorem 1 acts as an upper bound on the asymptotic bias correction.

5 Forward Selection of Regression Features

A common application of model selection procedures in machine learning and data mining is *feature selection*. Here, one is presented with many features (explanatory variables, covariates) and a single target variable $\mathbf{y}$ we wish to explain with the aid of some of these features. The AIC criterion is often used to determine if a feature is useful in explaining the target; this is a type of "all subsets" regression, in which any combination of features is considered plausible *a priori*, the data itself being used to determine whether the features are significant or statistically useful. Unfortunately, as the number of features may often be very large, the results of Section 3 suggest that the usual AIC is inappropriate, and choosing features by minimising an AIC score will generally lead to large numbers of "spurious" features being included in the final model. We propose a forward-selection AIC-like procedure, called AIC_m, based on the results of Theorem 1. Forward selection of features acts by iteratively enlarging the current model to include the feature that most improves the fit, and produces a type of nested sequence of models; unfortunately, the sequence is determined by the available data rather than *a priori* and so violates the usual AIC conditions.

The main idea behind our procedure is to note that, with high probability, the important non-spurious features will yield the best improvements in fit and be included before the spurious features. Thus, if there are k^* non-spurious features, the first k^* subsets created by the forward selection procedure will, with high probability, be the same irrespective of the random noise corrupting



(a) "Doppler" (b) Noisy (c) SURE (d) AIC_m

Fig. 1. Denoising of the "Doppler" Test Signal by *SureShrink* and AIC_m

our data, and thus form a usual nested sequence of models. However, once all k^* non-spurious features have been included, *the remaining* $(q - k^*)$ *subsets depend entirely on the random noise and form a non-nested sequence of models*; the results of Theorem 1 may be used to avoid selecting these spurious features.

The AIC_m procedure may be summarised as follows. Let $\gamma[k]$ denote the set of the q features included at step k, so that $\gamma[0] = \emptyset$, i.e., we start with an empty model and let $\bar{\gamma}[k] = \{1, \ldots, q\} - \gamma[k]$ denote the set of features not in $\gamma[k]$. Then, for $k = 0$

1. Find the unused feature that most decreases the negative log-likelihood

$$\gamma[k+1] = \underset{j \in \bar{\gamma}[k]}{\arg\min} \left\{ \log 1/p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\gamma[k] \cup j}) \right\}$$

2. If $\left(\log 1/p\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\gamma[k]}\right) - \log 1/p\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\gamma[k+1]}\right) \right) < (\alpha(1, q-k)+1)/2$ the feature is rejected and algorithm terminates
3. $k \leftarrow k + 1$; if $k = q$, algorithm terminates, otherwise go to Step 1.

The threshold for rejection is based on two observations; the first is that even if all $(q - k)$ remaining features at step k were spurious we would still expect to see, on average, an improvement in negative log-likelihood of $\alpha(1, q - k)/2$ from the best one amongst them (from the expectation of the a_m variates in Theorem 1). This accounts for the first term in the threshold. The second term arises by noting that if the improvement exceeds the first threshold, we are deciding the feature is non-spurious; at this point, we can use the regular AIC penalty of $1/2$ a unit to account for the variance introduced by estimating the extra parameter.

5.1 Application: Signal Denoising by Wavelet Thresholding

An interesting example of regression in which the number of features is very large is denoising or smoothing of a signal using orthonormal basis functions called "wavelets". An excellent discussion of wavelets, and their properties for smoothing, can be found in [7]), and it is from this paper we take our four test signals. These signals, called "Bumps", "Blocks", "HeaviSine" and "Doppler" are benchmarks in the wavelet literature and are designed to caricature various types of signals found in real applications.

We tested our AIC_m procedure on the wavelet smoothing problem by first applying the discrete wavelet transform to the noise corrupted versions of the test

Table 2. Squared prediction errors for denoising of the Donoho-Johnston test signals

	"Bumps"		"Blocks"		"HeaviSine"		"Doppler"	
SNR	Sure	AIC_m	Sure	AIC_m	Sure	AIC_m	Sure	AIC_m
1	0·4129	0·3739	0·3140	0·2302	0·2164	0·0315	0·2686	0·0963
10	0·0578	0·0478	0·0530	0·0498	0·0239	0·0079	0·0355	0·0157
100	0·0070	0·0055	0·0066	0·0055	0·0033	0·0015	0·0047	0·0020

signals, and then using our criterion to determine which wavelets (our "features") to include, the maximum number of wavelets possible being restricted to $n/2$ to ensure that the asymptotic conditions are not violated. The closeness of the resulting smoothed signal to the true signal was assessed using average mean squared error, and our AIC_m procedure was compared against the well known *SureShrink* algorithm [7]. Three levels of signal-to-noise ratio (SNR) (the ratio of signal variance to noise variance) were used, and for each combination of test signal and SNR level, the two criterion were tested one thousand times. The mean squared errors presented in Table 2 clearly demonstrate the effectiveness of the AIC_m procedure; in contrast, applying regular AIC resulted in the maximum number of $n/2$ wavelets being included in every case, with correspondingly poor performance. Figure 1 demonstrates the difference in performance between AIC_m and *SureShrink* for the "Doppler" signals at an SNR of ten; the AIC_m smoothing is visually superior to that obtained by *SureShrink*.

6 Conclusion

This paper examined the failings of AIC as a model selection criterion when the set of candidate models forms a non-nested sequence. The main contribution was a novel theorem quantifying the bias in the regular AIC estimate of the Kullback–Leibler risk, which demonstrated that this bias may not be overcome even as the sample size $n \to \infty$. This result was used to derive an AIC-like procedure for forward selection in regression models, and simulations suggested the procedure was competitive when applied to wavelet denoising.

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6), 716–723 (1974)
2. Hurvich, C.M., Tsai, C.L.: A crossvalidatory AIC for hard wavelet thresholding in spatially adaptive function estimation. Biometrika 85, 701–710 (1998)
3. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics 22(1), 79–86 (1951)
4. Linhart, H., Zucchini, W.: Model Selection. Wiley, New York (1986)
5. Cavanaugh, J.E.: A large-sample model selection criterion based on Kullback's symmetric divergence. Statistics & Probability Letters 42(4), 333–343 (1999)
6. Cramér, H.: Mathematical methods of statistics. Princeton University Press, Princeton (1957)
7. Donoho, D.L., Johnstone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. Journal of the Amer. Stat. Ass. 90(432), 1200–1224 (1995)