MODEL SELECTION: A BOOTSTRAP APPROACH

A. M. Zoubir

Communications and Information Processing Group
School of Electrical & Electronic Systems Engineering
Queensland University of Technology,
GPO Box 2434, Brisbane, QLD 4001, Australia

ABSTRACT

The problem of model selection is addressed. Bootstrap methods based on residuals are used to select the best model according to a prediction criterion. Both the linear and the nonlinear models are treated. It is shown that bootstrap methods are consistent and in simulations that in most cases they outperform classical techniques such as Akaike's information criterion and Rissanen's minimum description length. We also show how the methods apply to dependent data models such as autoregressive models.

1. INTRODUCTION

Model selection is a fundamental problem in many areas of signal processing, including system identification [5], radar [8] and sonar [12]. Among signal processing practitioners, two approaches for model selection have gained popularity and are widely used [7]. These are Akaike's Information Criterion (AIC) [1] and Rissanen's Minimum Description Length [9]. Although there exist many model selection procedures, the development of new techniques that outperform the popular ones is still growing and continues to grow (see for example the recently developed methods based on the generalised Kullback-Leibler information [11]).

The objective of this paper is to introduce methods for model selection based on the bootstrap in a signal processing framework. Besides the good statistical properties of bootstrap selection procedures there are other reasons for the use of the bootstrap for model selection.

The bootstrap is a powerful tool in that it requires very little in the way of modelling, assumptions, or analysis, and it can be applied in an automatic way when only a small set of data is available and standard methods that invoke the central limit theorem are inapplicable.

Usually, model selection is associated with parameter estimation and inference such as variance or mean squared error estimation of parameter estimators and hypothesis testing (e.g., signal detection). Inference based on the bootstrap has proved to be asymptotically more accurate than methods based on the Gaussian assumption. Therefore, it is preferable to use the bootstrap for both, model selection and subsequent inference applied to the selected model. This does not involve extra cost because the observations generated by the bootstrap for model selection can be used for inference.

Bootstrap model selection is not limited to linear models but can be extended to more complicated models.

Some methods for model selection in signal processing based on the bootstrap have been reported in [2, 14]. Here we present the general theory of model selection with the bootstrap based on residuals and explain why the methods are attractive. We give several examples and compare the results with those based on classical techniques.

2. MODEL SELECTION

Let $\boldsymbol{y} = (y_0, \ldots, y_{n-1})'$ be observations of a set of random variables $\boldsymbol{Y} = (Y_0, \ldots, Y_{n-1})'$. Based on the observations $\boldsymbol{y}$ we have a choice among q parameter dependent models $\mathcal{M}_1, \ldots, \mathcal{M}_q$. The objective of model selection is to choose the model which best explains the data $\boldsymbol{y}$.

Assume that a model $\mathcal{M}$ is specified by a probability density function $f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})$ of $\boldsymbol{Y}$ with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ being a parameter to be estimated based on $\boldsymbol{y}$. Provided the probability density function of the data is known, one may use the method of maximum likelihood. An intuitive solution to the model selection problem may be as follows: Given $\mathcal{M}_1, \ldots, \mathcal{M}_q$, find for each $\mathcal{M}_k$ the corresponding maximum value of the log-likelihood for $k = 1, \ldots, q$. A naive approach would then be to choose the model corresponding to the largest value with respect to k. However, it is known that this approach fails because it tends to pick the model with the largest number of parameters [9]. This is a problematic statistical solution as it contradicts the principle of parsimony. A modification of the log-likelihood function such that "parsimonious" models are favoured while "generous" models are penalised is a compromise. For example, Akaike's information criterion penalises a model with p parameters by subtracting the number of parameters p from the maximising log-likelihood function. Many other criteria are based on a similar principle. They include Rissanen's MDL [9] and Hannan and Quinn's criterion [4]. The statistical properties of these criteria have been well studied.

In our study, we focus on bootstrap methods. With little assumptions, these are shown to be consistent. In an extensive simulation study, we also show that in most cases they outperform most popular techniques. We first consider the simplest linear model and then extend the study to nonlinear models. We also demonstrate their performance in autoregressions.

3. MODEL SELECTION IN LINEAR MODELS

Consider the linear model

$$Y_t = \boldsymbol{x}_t'\boldsymbol{b} + Z_t, \qquad t = 0, \ldots, n-1, \tag{1}$$

where Z_t is a noise sequence of identically and independently distributed (iid) random variables of unknown distribution with mean

zero and variance σ_Z^2. The iid case is considered here for the sake of simplicity, but the methods presented can be extended to the case where Z_t is a correlated process. A discussion on this will be provided in Section 4. In (1), $\boldsymbol{b}$ is an unknown p-vector parameter and $\boldsymbol{x}_t$ is the t-th value of the p vector of explanatory variables. The output Y_t is sometimes called the response at $\boldsymbol{x}_t$.The vector $\boldsymbol{x}_t$ can be assumed to be random. This will affect the resampling schemes discussed below. For simplicity, we omit a random $\boldsymbol{x}_t$ and will only consider the case where $\boldsymbol{x}_t$ is fixed. A comprehensive treatment of model selection procedures when $\boldsymbol{x}_t$ is random can be found, for example, in [10].

Model (1) can be re-written as

$$\mu_t = \mathsf{E}[Y_t|\boldsymbol{x}_t] = \boldsymbol{x}_t'\boldsymbol{b}, \qquad \mathsf{var}[Y_t|\boldsymbol{x}_t] = \sigma_Z^2,$$

for $t = 0, \ldots, n-1$, and in vector form $\boldsymbol{Y} = \boldsymbol{x}\boldsymbol{b} + \boldsymbol{Z}$, where $\boldsymbol{Y} = (Y_0, \ldots, Y_{n-1})'$, the matrix $\boldsymbol{x} = (\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n-1})'$ is full rank, $\boldsymbol{b} = (b_1, \ldots, b_p)'$ and $\boldsymbol{Z} = (Z_0, \ldots, Z_{n-1})'$.

Let β be a subset of $\{1, \ldots, p\}$, $\boldsymbol{b}_\beta$ be a sub-vector of $\boldsymbol{b}$ containing the components of $\boldsymbol{b}$ indexed by integers in β, and let $\boldsymbol{x}_\beta$ be a matrix containing the columns of $\boldsymbol{x}$ indexed by integers in β. Then, a model corresponding to β is

$$\boldsymbol{Y} = \boldsymbol{x}_\beta\boldsymbol{b}_\beta + \boldsymbol{Z}. \tag{2}$$

Let β represent a model from now on. Define the optimal model as the model β_o such that $\boldsymbol{b}_{\beta_o}$ contains all non-zero components of $\boldsymbol{b}$ only. The problem of model selection is to estimate β_o based on the data $y_0, \ldots, y_{n-1}$. Our treatment will be based on an estimator of the mean-squared prediction error $\mathsf{E}(Y_{f,t} - \boldsymbol{x}_{f,t}'\hat{\boldsymbol{b}})^2$, where $\boldsymbol{x}_{f,t}'\hat{\boldsymbol{b}}$ is the prediction of the future response $Y_{f,t}$ at a given $\boldsymbol{x}_{f,t}$. For model β this estimator is given by

$$\Gamma_n(\beta) = \frac{1}{n}\sum_{t=0}^{n-1}\left(Y_t - \boldsymbol{x}_{\beta t}'\hat{\boldsymbol{b}}_\beta\right)^2 = \frac{\|\boldsymbol{Y} - \boldsymbol{x}_\beta\hat{\boldsymbol{b}}_\beta\|^2}{n} \tag{3}$$

where $\boldsymbol{x}_{\beta t}'$ is the t-th row of $\boldsymbol{x}_\beta$ and $\|\boldsymbol{a}\| = \sqrt{\boldsymbol{a}'\boldsymbol{a}}$ for any vector $\boldsymbol{a}$. One can show that the expected value of (3), taken with respect to Y_t, is equivalent to

$$\mathsf{E}[\Gamma_n(\beta)] = \sigma_Z^2 - \frac{\sigma_Z^2 p_\beta}{n} + \Delta_n(\beta),$$

where p_β is the size of $\boldsymbol{b}_\beta$, $\Delta_n(\beta) = n^{-1}\boldsymbol{\mu}'(\boldsymbol{I} - \boldsymbol{h}_\beta)\boldsymbol{\mu}$, with $\boldsymbol{\mu} = \mathsf{E}[\boldsymbol{Y}] = (\mu_0, \ldots, \mu_{n-1})'$ and $\boldsymbol{I}$ and $\boldsymbol{h}_\beta = \boldsymbol{x}_\beta(\boldsymbol{x}_\beta'\boldsymbol{x}_\beta)^{-1}\boldsymbol{x}_\beta'$ being the $p \times p$ identity and projection matrix, respectively. If β is a correct model in that $\boldsymbol{b}_\beta$ contains all non-zero components of $\boldsymbol{b}$ such that for any $\boldsymbol{x}$, $\boldsymbol{x}\boldsymbol{b} = \boldsymbol{x}_\beta\boldsymbol{b}_\beta$, then $\Delta_n(\beta)$ is identical zero.

An estimate of $\mathsf{E}[\Gamma_n(\beta)]$ minimised over β will lead to an optimal model. This principle is also used in AIC, for example. With the bootstrap, we would consider the estimate

$$\tilde{\Gamma}_n(\beta) = \frac{1}{n}\sum_{t=0}^{n-1}\mathsf{E}_*\left(y_t - \boldsymbol{x}_{t\beta}'\hat{\boldsymbol{b}}_\beta^*\right)^2 = \mathsf{E}_*\frac{\|\boldsymbol{y} - \boldsymbol{x}_\beta\hat{\boldsymbol{b}}_\beta^*\|^2}{n}, \tag{4}$$

where E_* denotes expectation operation with respect to bootstrap sampling [3], $\hat{\boldsymbol{b}}_\beta^*$ is the bootstrap analog of the least-squares estimate $\hat{\boldsymbol{b}}_\beta$, calculated in the same manner as $\hat{\boldsymbol{b}}_\beta$, but with $(y_t^*, \boldsymbol{x}_{\beta t})$ replacing $(y_t, \boldsymbol{x}_{\beta t})$. To obtain observations $y_t^*, t = 0, \ldots, n-1$, we use the following bootstrap method based on residuals.

Let $\hat{\boldsymbol{b}}$ be the least-squares estimate of $\boldsymbol{b}$ and define the t-th residual by $\hat{z}_t = y_t - \boldsymbol{x}_{\alpha t}'\hat{\boldsymbol{b}}_\alpha$, $t = 0, \ldots, n-1$, where $\alpha = \{1, \ldots, p\}$. Bootstrap resamples $\hat{z}_t^*$ can be generated by resampling with replacement from $(\hat{z}_t - \hat{z}_\bullet)/\sqrt{1 - p/n}$ (the inclusion of the divider $\sqrt{1-p/n}$ is for the purpose of bias correction), and computing $y_t^* = \boldsymbol{x}_{\beta t}'\hat{\boldsymbol{b}}_\beta + \hat{z}_t^*$, $t = 0, \ldots, n-1$, where $\hat{z}_\bullet = n^{-1}\sum_{t=0}^{n-1}\hat{z}_t$.

A refined bootstrap approach for estimating $\mathsf{E}[\Gamma_n(\beta)]$ first estimates the bias in $\Gamma_n(\beta)$ as an estimator of the true prediction error and then corrects $\Gamma_n(\beta)$ by subtracting its estimated bias [3]. The average difference between the true prediction error and its estimate over data sets $\boldsymbol{x}$, called the average optimism [3], can be estimated by the bootstrap, yielding

$$\hat{e}_n(\beta) = \mathsf{E}_*\left[\frac{\|\boldsymbol{y} - \boldsymbol{x}_\beta\hat{\boldsymbol{b}}_\beta^*\|^2}{n} - \frac{\|\boldsymbol{y}^* - \boldsymbol{x}_\beta\hat{\boldsymbol{b}}_\beta^*\|^2}{n}\right] = \frac{2\hat{\sigma}_Z^2 p_\beta}{n}$$

The final estimate of $\mathsf{E}[\Gamma_n(\beta)]$ is then given by

$$\hat{\Gamma}_n(\beta) = \frac{\|\boldsymbol{y} - \boldsymbol{x}_\beta\hat{\boldsymbol{b}}_\beta\|^2}{n} + \frac{2\hat{\sigma}_Z^2 p_\beta}{n}$$

Evaluation of the previous expression leads to

$$\begin{aligned}\hat{\Gamma}_n(\beta) &= \frac{\|\boldsymbol{z}\|^2}{n} + \frac{\|(\boldsymbol{I} - \boldsymbol{h}_\beta)\boldsymbol{\mu}\|^2}{n} \\ &- \frac{\|\boldsymbol{h}_\beta\boldsymbol{z}\|^2}{n} + \frac{2\boldsymbol{z}'(\boldsymbol{I} - \boldsymbol{h}_\beta)\boldsymbol{\mu}}{n} + \frac{2\hat{\sigma}_Z^2 p_\beta}{n}.\end{aligned}$$

Under some mild regularity conditions (see [10] for details),

$$\hat{\Gamma}_n(\beta) = \mathsf{E}[\Gamma_n(\beta)] + o_p(1) \tag{5}$$

and

$$\hat{\Gamma}_n(\beta) = \frac{\|\boldsymbol{z}\|^2}{n} + \frac{2\sigma_Z^2 p_\beta}{n} - \frac{\|\boldsymbol{h}_\beta\boldsymbol{z}\|^2}{n} + o_p(n^{-1})$$

for an incorrect and a correct model, respectively (a model is incorrect if $\boldsymbol{\mu} \neq \boldsymbol{x}\boldsymbol{b}$). This result indicates that the model selection procedure based on minimising $\hat{\Gamma}_n(\beta)$ over β is inconsistent in that $\lim_{n\to\infty}\mathsf{P}\{\hat{\beta} = \beta_0\} < 1$, unless $\beta = \{1, \ldots, p\}$ is the only correct model. A consistent model selection procedure is obtained if we replace $\hat{e}_n(\beta)$ by $\hat{e}_m(\beta)$ where m is chosen such that, with $h_{\beta t} = \boldsymbol{x}_{\beta t}'(\boldsymbol{x}_\beta'\boldsymbol{x}_\beta)^{-1}\boldsymbol{x}_{\beta t}$,

$$\frac{m}{n} \to 0 \quad \text{and} \quad \frac{n}{m}\max_{t\le n} h_{\beta t} \to 0$$

for all β in the class of models to be selected. Then,

$$\hat{\Gamma}_{n,m}(\beta) = \frac{\|\boldsymbol{z}\|^2}{n} + \frac{\sigma_Z^2 p_\beta}{m} + o_p(m^{-1})$$

when β is a correct model, otherwise $\hat{\Gamma}_{n,m}(\beta)$ is as in Eq. (5). These results suggest that we estimate $\mathsf{E}[\Gamma_{n,m}(\beta)]$ through

$$\hat{\Gamma}_{n,m}^*(\beta) = \mathsf{E}_*\left[\frac{\|\boldsymbol{y} - \boldsymbol{x}\hat{\boldsymbol{b}}_{\beta,m}^*\|^2}{n}\right], \tag{6}$$

where $\hat{\boldsymbol{b}}_{\beta,m}^*$ is the bootstrap analog of $\hat{\boldsymbol{b}}$ obtained from $y_t^* = \boldsymbol{x}_{\beta t}'\hat{\boldsymbol{b}}_\beta + \hat{z}_t^*$, $t = 0, \ldots, n-1$, where $\hat{z}_t^*$ denotes the bootstrap resample from $\sqrt{n/m}(\hat{z}_t - \hat{z}_\bullet)/\sqrt{1-p/n}$. To evaluate the ideal expression in (6), we use Monte Carlo approximations, in which we repeat the resampling stage B times to obtain $\hat{\boldsymbol{b}}_{\beta,m}^{*(i)}$ and $\hat{\Gamma}_{n,m}^{*(i)}(\beta)$, and average $\hat{\Gamma}_{n,m}^{*(i)}(\beta)$ over $i = 1, \ldots, B$.

	$\mathcal{N}(0,1)$			t_3		
Model β	$\hat{\Gamma}^*$	AIC	MDL	$\hat{\Gamma}^*$	AIC	MDL
$(0,0,b_2,b_3)$	**100**	**91**	**98**	**99**	**89**	**98**
$(0,b_1,b_2,b_3)$	0	5	1	1	5	1
$(b_0,0,b_2,b_3)$	0	3	1	0	3	1
(b_0,b_1,b_2,b_3)	0	2	0	0	3	0

Table 1: Estimates of the empirical probabilities (in percent) on selecting models for a trend with $\boldsymbol{b} = (0, 0, 0.035, -0.0005)'$, embedded in Gaussian and t_3 distributed noise, $n = 64$, $m = 2$.

3.1. Example: Trend Estimation

We give a simple example where we estimate the model for a trend in a stationary iid process of unknown distribution. Let $Y_t = \boldsymbol{x}_t'\boldsymbol{b} + \boldsymbol{Z}_t$, $t = 0, \ldots, n-1$ where $\boldsymbol{x}_t = (1, t, \ldots, t^p)$, $t = 0, \ldots, n-1$, $\boldsymbol{b}$ is the vector of polynomial coefficients chosen to be $\boldsymbol{b} = (0, 0, 0.035, -0.0005)'$ and $n = 64$. We simulate Y_t by adding Gaussian and t_3-distributed noise of variance of 1 and 3, respectively.

The bootstrap procedure was run using $B = 100$ and $m = 2$. The minimiser of $\hat{\Gamma}^*_{n,m}(\beta)$ was selected as the optimal model. Table 1 shows the empirical probabilities (based on 1,000 simulations) on selecting some models (models not shown were not selected by any of the methods). Clearly, in this example the bootstrap outperforms the AIC and the MDL criterion.

4. MODEL SELECTION IN NONLINEAR MODELS

The principles discussed in the previous section are easily extendible to nonlinear models. We define a nonlinear model by

$$Y_t = g(\boldsymbol{x}_t, \boldsymbol{b}) + Z_t, \qquad t = 0, \ldots, n-1, \tag{7}$$

where Z_t is a noise sequence of iid random variables of unknown distribution with mean zero and variance σ_Z^2. The model in (7) can also be written as

$$\mu_t = \mathsf{E}[Y_t|\boldsymbol{x}_t] = g(\boldsymbol{x}_t, \boldsymbol{b}), \qquad \mathsf{var}[Y_t|\boldsymbol{x}_t] = \sigma_Z^2,$$

for $t = 0, \ldots, n-1$. Herein, g is a known function. Let $\mathcal{B}$ be a collection of subsets of $\{1, \ldots, p\}$, and let $g_{\beta t}(\boldsymbol{b}_\beta) = g_\beta(\boldsymbol{x}_{\beta t}, \boldsymbol{b}_\beta)$, where $\beta \in \mathcal{B}$ and g_β is the restriction of the function g to the admissible set of $(\boldsymbol{x}_{\beta t}, \boldsymbol{b}_\beta)$. Let $\tilde{\mathcal{B}}$ be the admissible set for $\boldsymbol{b}$.

A consistent bootstrap procedure for selecting β is given in Table 2, where $\dot{\boldsymbol{g}}(\boldsymbol{\gamma}) = \frac{\partial g(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$ and $\boldsymbol{m}_\beta(\boldsymbol{\gamma}) = \sum_{t=0}^{n-1} \dot{\boldsymbol{g}}_{\beta t}(\boldsymbol{\gamma})\dot{\boldsymbol{g}}_{\beta t}(\boldsymbol{\gamma})'$.

The proof for consistency of this procedure requires more regularity conditions than the one in section 3. Specifically, conditions for the asymptotic normality of $\hat{\boldsymbol{b}}_\beta$ and its bootstrap analog are needed [10]. The performance of this method is highlighted in an example.

4.1. Example: Oscillations in noise

Consider the model $\cos\omega_1 t(1 + \cos\omega_2 t) + Z_t, t \in \mathbb{Z}$, In this case $\mathcal{B} = \{\beta_k, k = 1, 2, 3\}$. Then, for example, $g_{\beta_1 t}(\boldsymbol{b}_{\beta_1}) = 2\cos\omega_1 t$ $(\omega_2 = 0)$, $g_{\beta_2 t}(\boldsymbol{b}_{\beta_2}) = 1 + \cos\omega_2 t$ $(\omega_1 = 0)$, and $g_{\beta_3 t}(\boldsymbol{b}_{\beta_3}) = \cos\omega_1 t(1 + \cos\omega_2 t)$ $(\omega_1, \omega_2 \neq 0)$. We run simulations at -1.2 dB signal-to-noise power ratio with $n = 40$ and $m = 35$. The frequencies were selected to be $\omega_1 = 0.2\pi$ and $\omega_2 = 0.1\pi$. The empirical probabilities (based on 100 simulations) are given in Table 3. It appears that in this example all three methods perform equally well. Similar results were obtained under different conditions and parameter settings.

Step 1. With $y_t, t = 0, \ldots, n-1$, find $\hat{\boldsymbol{b}}_\alpha$, the solution of

$$\sum_{t=0}^{n-1} (y_t - g_{\alpha t}(\boldsymbol{\gamma}))\, \dot{\boldsymbol{g}}_{\alpha t}(\boldsymbol{\gamma}) = 0,$$

for all $\boldsymbol{\gamma} \in \tilde{\mathcal{B}}$ with $\alpha = \{1, \ldots, p\}$.

Step 2. Compute the residuals $\hat{z}_t = y_t - g_{\alpha t}(\hat{\boldsymbol{b}}_\alpha)$ for $t = 0, \ldots, n-1$.

Step 3. Get $\hat{z}_t^*$, $t = 0, \ldots, n-1$, iid samples from the empirical distribution putting mass n^{-1} on each

$$\sqrt{n/m}(\hat{z}_t - \hat{z}_\bullet)/\sqrt{1 - p/n}, \quad t = 0, \ldots, n-1.$$

Step 4. Compute $\hat{\boldsymbol{b}}^*_{\beta,m} = \hat{\boldsymbol{b}}_\beta + \boldsymbol{m}_\beta(\hat{\boldsymbol{b}}_\beta)^{-1} \sum_{t=0}^{n-1} \hat{z}_t^* \dot{\boldsymbol{g}}_{\beta t}(\hat{\boldsymbol{b}}_\beta)$.

Step 5. Repeat steps 3-4 a large number of times to obtain $\hat{\boldsymbol{b}}^{*(i)}_{\beta,m}$, $i = 1, \ldots, B$.

Step 6. Compute

$$\hat{\Gamma}^*_{n,m}(\beta) = B^{-1} \sum_{i=1}^{B} \sum_{t=0}^{n-1} \frac{\left(y_t - g_{\beta t}(\hat{\boldsymbol{b}}^{*(i)}_{\beta,m})\right)^2}{n}.$$

Step 7. Minimise $\hat{\Gamma}^*_{n,m}(\beta)$ over β to find $\hat{\beta}$.

Table 2: Bootstrap procedure for model selection in nonlinear regression.

Method	β_1	β_2	β_3
Boot	3	0	**97**
AIC	0	3	**97**
MDL	0	5	**95**

Table 3: Probabilities (in percent) of selecting the true model.

In Sections 3 and 4, we considered the case where the errors are iid. However, the methods can be extended to the correlated case. For this, we could use an alternative resampling scheme such as the method of *sub-sampling* described in [6], which works well for a coloured noise sequence. An alternative would be to model the coloured noise sequence as an autoregressive process, for example. Then, the residuals of the autoregressive process could be used for resampling. Resampling autoregressive processes for model selection is discussed below.

Choice of m. The methods described in this Sections 3 and 4 require the choice of the parameter m with the consistency properties indicated in the sections. An optimal m may depend on model parameters and thus may be difficult or even impossible to determine. One guideline for choosing m is such that p/m should be reasonably small.

5. ORDER SELECTION IN AUTOREGRESSIVE MODELS

The methods discussed above can be generalised to linear processes. Here, we consider model selection of an autoregressive process. Consider

$$Y_t = b_1 Y_{t-1} + b_2 Y_{t-2} + \cdots b_p Y_{t-p} + Z_t, \quad t \in \mathbb{Z},$$

where p is the order, b_k, $k = 1, \ldots, p$, are unknown parameters and Z_t are iid random variables with mean zero and variance σ_Z^2. Let $(y_{-p}, \ldots, y_{-1}, y_0, \ldots, y_{n-1})$ be observations and collect the parameters into a vector $\boldsymbol{b}$ whose least squares estimator is $\hat{\boldsymbol{b}}$.

A resampling procedure for estimating the variance of the estimator of the parameter of an AR(1) process has been described in [13]. The principle can be used here in a similar fashion to estimate the order of an AR process. We thus select a model β from $\mathcal{B} = \{1, \ldots, p\}$ and *each β corresponds to the autoregressive model of order β*, i.e., $Y_t = b_1 Y_{t-1} + b_2 Y_{t-2} + \cdots b_\beta Y_{t-\beta} + Z_t$. The optimal order is $\beta_0 = \max\{k : 1 \le k \le p, b_k \neq 0\}$, where p is the largest order. The bootstrap approach is described in Table 4.

Step 1. Resample the residuals $(\hat{z}_t - \hat{z}_\bullet)$ to obtain $\hat{z}_t^*$.

Step 2. Find $\hat{\boldsymbol{b}}_{\beta,m}^*$ the least-squares estimate of $\boldsymbol{b}_\beta$ under β from $y_t^* = \sum_{k=1}^{\beta} \hat{b}_k y_{t-k}^* + \hat{z}_t^*$ for $t = -p, \ldots, m-1$, with m replacing n and where the initial bootstrap observations $\{y_{-2p}^*, \ldots, y_{-p-1}^*\}$ are chosen to be equal to $\{y_{-p}^*, \ldots, y_0^*\}$.

Step 3. Repeat steps 1-2 to obtain $\hat{\boldsymbol{b}}_{\beta,m}^{*(1)}, \ldots, \hat{\boldsymbol{b}}_{\beta,m}^{*(B)}$ and

$$\hat{\Gamma}_{n,m}^*(\beta) = B^{-1} \sum_{i=1}^{B} \sum_{t=0}^{n-1} \frac{\left(y_t - \sum_{k=1}^{\beta} y_{t-k+1} \hat{b}_{k,m}^{*(i)}\right)^2}{n}$$

Step 5. Minimise $\hat{\Gamma}_{n,m}^*(\beta)$ over β to find $\hat{\beta}$.

Table 4: Procedure for order selection in an AR model.

The procedure described in Table 4 is consistent in that $\mathsf{P}\{\hat{\beta} = \beta_0\} \to 1$ as $n \to \infty$, provided m satisfies $m \to \infty$ and $m/n \to 0$ as $n \to \infty$. The proof requires stability of the recursive filter and Cramér's condition. Details can be found in [10]. Note that in Section 3 and 4 m was a scalor of the residuals while here it determines the size of the data used for the bootstrap estimates.

5.1. Example: Order Selection in an AR Model

In this example, we consider the problem of determining the order of the process described by

$$Y_t = -0.4 Y_{t-1} + 0.2 Y_{t-2} + Z_t, \quad t \in \mathbb{Z},$$

where Z_t is a standard Gaussian variable. A number of $n = 128$ observations was considered. Results of the procedure described in Table 4 as well as a comparison with AIC and the MDL criterion are given in Table 5.

Similar results were obtained with different constellations and noise types. In the simulations we run the choice of m does not appear to have an effect on the results, as long as it satisfies the condition given above.

Method	$\beta = 1$	$\beta = 2$	$\beta = 3$	$\beta = 4$
Boot	28.0	**65.0**	5.0	2.0
AIC	17.8	**62.4**	12.6	7.2
MDL	43.2	**54.6**	2.1	0.1

Table 5: Empirical Probabilities (in percent) of selecting the true AR model, $p = 2$. $n = 128$ and $m = 40$.

6. CONCLUSIONS

We have discussed model selection techniques based on bootstrap residuals. We have considered linear, nonlinear models as well as autoregressions. The methods are based on a predictive measure which is estimated by the bootstrap. The methods are shown to be consistent when the residuals are scaled appropriately. The examples described show that the techniques outperform Akaike's information criterion and Rissanen's minimum description length.

7. REFERENCES

[1] H. Akaike. A New Look at the Statistical Model Identification. *IEEE Trans. on Automat. Contr.*, 19:716–723, 1974.

[2] P. M. Djurić. Using the Bootstrap to Select Models. In *Proc. of the IEEE ICASSP–97*, vol. V, pp. 3729–3732, Munich, Germany, 1997.

[3] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[4] E. J. Hannan and B. G. Quinn. The Determination of the Order of an Autoregression. *J. R. Statist. Soc., B*, 41:190–195, 1979.

[5] Ljung L. *System Identification. Theory for the User*. Prentice-Hall, 1987.

[6] D. N. Politis. A Primer on Bootstrap Methods in Statistics. *IEEE Signal Processing Magazine*, 15(1):39–55, 1998.

[7] B. Porat. *Digital Processing of Random Signals; Theory and Methods*. Prentice-Hall, 1994.

[8] A. W. Rihaczek. *Principles of High-Resolution Radar*. Peninsula Publishing, 1985.

[9] J. Rissanen. A universal prior integers and estimating by minimum description length. *Ann. Statist.*, 11:416–431, 1983.

[10] J. Shao. Bootstrap Model Selection. *J. Am. Statist. Assoc.*, 91:655–665, 1996.

[11] P. Shi and C. L. Tsai. A note on the unification of the Akaike information criterion. *J. R. Statist. Soc. B*, 60:551–558, 1998.

[12] M Wax and T. Kailath. Detection of Signals by Information Theoretic Criteria. *IEEE Trans. Acoust., Speech, and Signal Processing*, 33:387–392, 1985.

[13] A. M. Zoubir and B. Boashash. The Bootstrap and Its Application in Signal Processing. *IEEE Signal Processing Magazine*, 15(1):56–76, 1998.

[14] A. M. Zoubir and D. R. Iskander. Bootstrap Model Selection for Polynomial Phase Signals. In *Proc. IEEE ICASSP-98*, vol. 4, pp. 2229–2232, Seattle, USA, 1998.