

Parameter-based hypothesis tests for model selection

J.A. Stark*, W.J. Fitzgerald

Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK

Received 9 June 1994; revised 9 January 1995 and 29 May 1995

Abstract

This paper explores parameter-based hypothesis tests for selecting between candidate models that predict an unknown variable from observations. This is the form of many time series models, classifiers, and data-fitting models. The basis for this paper is that if a model contains redundant terms the associated parameters can be set to zero without penalty. Hypothesis tests are proposed for assessing the statistical evidence for parameters taking non-zero values. These compare closely with standard criteria such as Akaike's and the Bayesian information criterion. A numerical simulation is presented to illustrate the criteria. The link between selection criteria based on parameter distributions and those based on data distributions is relevant to techniques such as changepoint methods. Resampling and other similar techniques may be applied using this framework.

Zusammenfassung

Diese Arbeit untersucht auf Parametern basierende Hypothesentests zur Auswahl zwischen Modellen, die eine unbekannte Variable auf der Grundlage von Beobachtungen vorhersagen. Dies entspricht vielen Zeitreihenmodellen, Klassifizierern und Datenanpassungsmodellen. Die Grundlage dieser Arbeit ist die folgende Tatsache: wenn ein Modell redundante Terme enthält, können die entsprechenden Parameter ohne Verluste gleich null gesetzt werden. Es werden Hypothesentests zur Beurteilung der statistischen Aussagekraft hinsichtlich Parametern vorgeschlagen, die Werte ungleich null annehmen. Diese Hypothesentests sind Standardkriterien wie dem Akaike-Kriterium und dem Bayesschen Informationskriterium sehr ähnlich. Die Kriterien werden durch eine numerische Simulation illustriert. Die Verbindung zwischen auf Parameterverteilungen beruhenden Selektionskriterien und auf Datenverteilungen beruhenden Selektionskriterien ist relevant für Verfahren wie z.B. Changepoint-Methoden. Resampling und ähnliche Methoden können in diesem Rahmen angewandt werden.

Résumé

Cet article explore les différents tests d'hypothèse basés sur les paramètres pour la sélection parmi des modèles candidats qui prédisent une variable inconnue à partir d'observations. C'est la forme de nombreux modèles de séries temporelles, de classificateurs et de modèles d'ajustement aux données. L'idée de cet article est que si un modèle contient des termes redondants, les paramètres associés peuvent être mis à zéro sans perte. Des tests d'hypothèse sont proposés pour estimer l'importance statistique des paramètres ayant des valeurs autres que zéro. Ces tests peuvent aisément être comparés à des critères standards, tels le critère d'Akaike ou le critère d'Information Bayésien. Une simulation numérique

* Corresponding author. Tel.: (+44) 1223 33 27 67. Fax: (+44) 1223 33 2662. E-mail: jas25@cam.ac.uk.

est présentée afin d'illustrer les critères. Le lien entre les critères basés sur la distribution des paramètres et ceux basés sur la distribution des données est pertinent pour des techniques comme les méthodes du point de changement. Le rééchantillonnage et autres techniques similaires peuvent être appliquées en utilisant ce cadre.

Keywords: Model selection; Akaike's information criterion; Bayesian information criterion; Signal modelling; Data fitting; Change-point methods; Neural networks; Polynomial fitting; System identification

1. Introduction

This paper explores one approach to the selection of models, a problem with widespread applications. The modelling problem being considered is the prediction of a data value from a set of realisations of measured data. Such model selection is applicable to data fitting, time series model selection, feature selection in classification, and complexity reduction in neural networks. In classification and learning problems the training time and the failure rate are increased by the presence of redundant complexity. The dangers of overfitting data are well-known. In this paper the analysis is restricted to candidate models that calculate a prediction using a *linear* combination of realisations of data which are either the measured data values directly or a non-linear (Volterra) expansion of the data. For example, the second-order expansion of the variables x_1, x_2, x_3 is $\{x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_2x_3, x_1x_3\}$. These are 'linear-in-the-parameters' models.

Consider a model that has redundant complexity. The coefficients or parameters associated with the unnecessary terms should be zero. Suppose we compare the correct model for a data-set to another with additional terms (that is *nested* within a larger model). These terms will be redundant. This paper explores statistical tests on the parameters of a model, and by testing the hypothesis that the values of the additional parameters are zero we can evaluate the evidence for believing that the additional complexity provides a useful contribution to the model. Only the case of prediction errors with a Gaussian distribution is considered, although the concept of testing the distributions of parameters is one that can readily be extended using numerical methods and resampling.

The first section discusses the type of model and the requirements of an efficient and consistent selection criterion. The following sections present two fundamentally different approaches to testing the zero-parameter hypothesis. The first uses a classical confidence interval test for both marginalised and unmarginalised distributions of the additional parameters, and this leads to a statistic in a similar form to AIC (Akaike's information criterion) [1]. The second uses a Bayesian hypothesis test and the expressions for the values of the posterior probabilities of the two hypotheses lead to a modified form of the BIC (Bayesian information criterion) [10]. Throughout this discussion the concepts and expressions are compared with standard techniques. The results of a numerical experiment are then presented to illustrate the performance of the criteria.

2. The model selection problem

2.1. Assessment of model selection criteria

Model selection criteria are *relative*. No absolute measure of model fit exists, and if we do not include the 'correct' model in the set that we consider, then we will certainly make the wrong choice. It is generally implicitly assumed that any criterion that will select the correct model over all others given a large amount of data will select the 'most appropriate' model from a set of wrong but approximately correct models. This is also assumed when too few data points have been measured to select between candidates that are hard to distinguish. Exactly what the most appropriate model is depends on the situation, but in many cases we are interested in minimising the prediction errors when new data are presented.

It is important to decide on a definition of the correct model. Consider a model that provides a prediction based on the observed data. Suppose that the prediction error is zero mean and is completely uncorrelated with the observed data. Since the error has a mean of zero we cannot obtain a reduction in error variance by changing the parameters of the model. If we are unable to find any further observations that provide information about the prediction error, that is are correlated with it in any way, then it is not possible to reduce this error and we have the ‘correct’ model. It may be possible to find equivalent models if new observations can be found that are exact functions (that is without additional random variations) of the required set. An important result of this definition is that the correct model has the *minimum error variance*.

The performance of a model selection criterion can be assessed in a number of ways. The minimum error-variance property is useful in designing tests that exhibit one form of consistency. This is that if model ‘A’ is preferred to ‘B’, and model ‘B’ to ‘C’, then model ‘A’ will be preferred to model ‘C’. There is also an *asymptotic consistency* condition that requires that a minimum data-set size can be found for each false model, such that the correct model will be selected in preference with a probability of error less than a given value. This condition is important for defining an effective selection criterion. There are two effects of increasing the data-set size. First, the criterion must be able to distinguish between increasingly similar models with decreasing probability of error. Second, the criterion must include a penalty against complexity, or ‘Ockham factor’, so that models with redundant complexity are less likely to be chosen.

The discussion is limited to the case where the errors in predictions take a Gaussian distribution. This is generally considered to be a reasonable assumption for linear models, but in the absence of better information it is perhaps the weakest assumption for non-linear models as well. If a different error model is proposed, this could easily be incorporated in the same manner. If it is then impossible to find the posterior distribution of the parameters analytically, numerical methods could be used.

2.2. Linear-in-the-parameters models

Consider an observed data series $\{d_i\}$ of size N with corresponding vectors \mathbf{x}_i of realisations of measured data. An example of a set of realisations of underlying variables u_j is the second-order Volterra expansion $\{u_1, u_1^2, u_2, u_2^2, u_1 u_2, \dots\}$. Linear-in-the-parameters models form a prediction of the data series from a linear combination of these realisations. Putting the parameters into a vector $\boldsymbol{\theta}$ we can write the prediction as $\mathbf{x}_i^T \boldsymbol{\theta}$, and the mean-square prediction error estimate function as

$$\begin{aligned} V(\boldsymbol{\theta}) &= \frac{1}{N} \sum_{i=1}^N |\mathbf{x}_i^T \boldsymbol{\theta} - d_i|^2 \\ &= \frac{1}{N} \sum_{i=1}^N d_i^2 - 2\boldsymbol{\theta}^T \frac{1}{N} \sum_{i=1}^N d_i \mathbf{x}_i \\ &\quad + \boldsymbol{\theta}^T \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^T \boldsymbol{\theta} \\ &= r_{dd} - 2\boldsymbol{\theta}^T \mathbf{r}_d + \boldsymbol{\theta}^T \mathbf{R} \boldsymbol{\theta}, \end{aligned} \tag{1}$$

where \mathbf{R} , \mathbf{r}_d and r_{dd} are the correlations of the realisations of the measured data and the data sequence being predicted.

2.3. Probability distributions for models

In this paper we will only consider models for which the distribution of prediction errors is Gaussian (see Section 2.1). Denoting the model structure as \mathcal{M} , the probability of a particular set of data being measured under that model and parameters $\boldsymbol{\theta}$ is

$$p(\mathbf{d} | \boldsymbol{\theta}, \mathcal{M}) = \frac{1}{(2\pi\sigma_e^2)^{N/2}} \exp \left[\frac{-NV(\boldsymbol{\theta})}{2\sigma_e^2} \right]. \tag{2}$$

For our present purposes, suppose that we do not have prior knowledge to say that the parameters should be, say, positive or negative. It seems reasonable then to assign a zero-mean prior distribution to the parameters. At this point we will assume that something is known about their magnitude. Using the principle of maximum entropy to

minimise the effect of the constraints on the uncertainty, this would suggest a Gaussian distribution. Let the variance of this be γ^2 . This is the approximate magnitude that we expect the parameters to take. This variance can in some cases be allowed to grow so that the prior becomes uniform in the limit.

Before finding the marginalised distribution $p(\mathbf{d}|\mathcal{M})$, let P be the number of parameters and define

$$\mathbf{R}_\gamma = \mathbf{R} + \frac{\sigma_e^2}{N\gamma^2} \mathbf{I}_P \quad (3)$$

and $V_\gamma(\boldsymbol{\theta})$ as $V(\boldsymbol{\theta})$ but with \mathbf{R}_γ substituted in.

To find the marginalised distribution we have to integrate out the parameters:

$$\begin{aligned} p(\mathbf{d}|\mathcal{M}) &= \int p(\mathbf{d}|\boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta} \\ &= \int \frac{1}{(2\pi\sigma_e^2)^{N/2}} \exp\left[-\frac{NV(\boldsymbol{\theta})}{2\sigma_e^2}\right] \\ &\quad \times \frac{1}{(2\pi\gamma^2)^{P/2}} \exp\left[-\frac{\boldsymbol{\theta}^T \boldsymbol{\theta}}{2\gamma^2}\right] d\boldsymbol{\theta} \\ &= \int \frac{1}{(2\pi\sigma_e^2)^{N/2}} \frac{1}{(2\pi\gamma^2)^{P/2}} \exp\left[-\frac{NV_\gamma(\boldsymbol{\theta})}{2\sigma_e^2}\right] d\boldsymbol{\theta} \\ &= \int \frac{1}{(2\pi\sigma_e^2)^{N/2}} \frac{1}{(2\pi\gamma^2)^{P/2}} \\ &\quad \times \exp\left[\frac{-N}{2\sigma_e^2} ((\boldsymbol{\theta} - \mathbf{R}_\gamma^{-1} \mathbf{r}_d)^T \mathbf{R}_\gamma (\boldsymbol{\theta} - \mathbf{R}_\gamma^{-1} \mathbf{r}_d) \right. \\ &\quad \left. + r_{dd} - \mathbf{r}_d^T \mathbf{R}_\gamma^{-1} \mathbf{r}_d)\right] d\boldsymbol{\theta} \\ &= \frac{1}{(2\pi\sigma_e^2)^{N/2}} \left(\frac{\sigma_e^2}{N\gamma^2}\right)^{P/2} \frac{1}{\det(\mathbf{R}_\gamma)^{1/2}} \\ &\quad \times \exp\left[-\frac{NV_\gamma(\hat{\boldsymbol{\theta}})}{2\sigma_e^2}\right], \quad (4) \end{aligned}$$

where $\hat{\boldsymbol{\theta}}$ is chosen to minimise the prediction-error variance:

$$V_\gamma(\hat{\boldsymbol{\theta}}) = r_{dd} - \mathbf{r}_d^T \mathbf{R}_\gamma^{-1} \mathbf{r}_d. \quad (5)$$

2.4. Distributions over some parameters

In the next section we will consider methods of comparing two models, one being nested within the other. To do this we need to split the data and parameter vectors into elements that are common to both models and additional to the larger model ($\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_A$). Quantities referring to the model with only the common terms are subscripted with C.

The common parameters can be integrated out of $p(\mathbf{d}|\boldsymbol{\theta}, \mathcal{M})$ as follows:

$$\begin{aligned} p(\mathbf{d}|\boldsymbol{\theta}_A, \mathcal{M}) &= \int p(\mathbf{d}|\boldsymbol{\theta}_A, \boldsymbol{\theta}_C, \mathcal{M}) p(\boldsymbol{\theta}_C|\mathcal{M}) d\boldsymbol{\theta}_C \\ &= \int \frac{1}{(2\pi\sigma_e^2)^{N/2}} \exp\left[-\frac{NV(\boldsymbol{\theta})}{2\sigma_e^2}\right] \\ &\quad \times \frac{1}{(2\pi\gamma^2)^{P_C/2}} \exp\left[-\frac{\boldsymbol{\theta}_C^T \boldsymbol{\theta}_C}{2\gamma^2}\right] d\boldsymbol{\theta}_C \\ &= \int \frac{1}{(2\pi\sigma_e^2)^{N/2}} \frac{1}{(2\pi\gamma^2)^{P_C/2}} \\ &\quad \times \exp\left[-\frac{NV_\gamma(\boldsymbol{\theta})}{2\sigma_e^2}\right] \exp\left[-\frac{\boldsymbol{\theta}_A^T \boldsymbol{\theta}_A}{2\gamma^2}\right] d\boldsymbol{\theta}_C. \quad (6) \end{aligned}$$

In the case when $\boldsymbol{\theta}_A = \mathbf{0}$ the integration proceeds as in the previous section, substituting $V_{\gamma C}(\boldsymbol{\theta}_C)$ for $V_\gamma(\boldsymbol{\theta})$:

$$\begin{aligned} p(\mathbf{d}|\boldsymbol{\theta}_A = \mathbf{0}, \mathcal{M}) &= \frac{1}{(2\pi\sigma_e^2)^{N/2}} \left(\frac{\sigma_e^2}{N\gamma^2}\right)^{P_C/2} \frac{1}{\det(\mathbf{R}_{\gamma C})^{1/2}} \\ &\quad \times \exp\left[-\frac{N}{2\sigma_e^2} (V_{\gamma C}(\hat{\boldsymbol{\theta}}_C))\right]. \quad (7) \end{aligned}$$

In the general case it is easier to note that $\boldsymbol{\theta}_A$ will appear in a quadratic form and that Eq. (7) is a special case:

$$\begin{aligned} p(\mathbf{d}|\boldsymbol{\theta}_A, \mathcal{M}) &= \frac{1}{(2\pi\sigma_e^2)^{N/2}} \left(\frac{\sigma_e^2}{N\gamma^2}\right)^{P_C/2} \frac{1}{\det(\mathbf{R}_{\gamma C})^{1/2}} \\ &\quad \times \exp\left[-\frac{N}{2\sigma_e^2} (\boldsymbol{\theta}_A^T \mathbf{G} \boldsymbol{\theta}_A - 2\mathbf{f}^T \boldsymbol{\theta}_A + V_{\gamma C}(\hat{\boldsymbol{\theta}}_C))\right], \quad (8) \end{aligned}$$

where \mathbf{G} and \mathbf{f} are expressions in terms of components of \mathbf{R}_γ .

3. Confidence interval analysis

3.1. Derivation of the test statistic

Suppose that we wish to compare two models which are nested so that the larger has some additional terms over the other. The model selection question is, then, ‘are these additional terms redundant?’. If the additional terms do not contribute usefully to the model, it should be possible to assign zero values to their corresponding parameters without penalty. By calculating the posterior probability distribution of the additional parameters by integrating out the common parameters, we can assess the evidence for this. Using Bayes’ rule,

$$p(\theta_A | \mathbf{d}, \mathcal{M}) = \frac{p(\mathbf{d} | \theta_A, \mathcal{M}) p(\theta_A | \mathcal{M})}{p(\mathbf{d} | \mathcal{M})}. \quad (9)$$

We can take a classical statistics approach using the following hypothesis test.

\mathcal{H}_0 : the additional parameters all have the value 0.

\mathcal{H}_1 : at least one of the additional parameters is non-zero.

If we calculate the magnitude of the mean after normalisation by the covariance matrix, we can see whether there is statistical evidence in the classical sense to reject the null hypothesis. Consider Eq. (8). If $p(\theta_A | \mathcal{M})$ is incorporated, then \mathbf{G} is modified in a similar way to \mathbf{R} and the numerator term of (9) is in the form

$$p(\mathbf{d} | \theta_A, \mathcal{M}) p(\theta_A | \mathcal{M}) \propto \exp \left[\frac{-N}{2\sigma_e^2} \left([\theta_A^T - \mathbf{f}^T \mathbf{G}_y^{-1}] \mathbf{G}_y [\theta_A - \mathbf{G}_y^{-1} \mathbf{f}] + V_{\gamma C}(\hat{\theta}_C) - \mathbf{f}^T \mathbf{G}_y^{-1} \mathbf{f} \right) \right]. \quad (10)$$

If θ_A were then integrated out, the result would be $p(\mathbf{d} | \mathcal{M})$ – Eq. (4). Thus,

$$V_{\gamma C}(\hat{\theta}_C) - \mathbf{f}^T \mathbf{G}_y^{-1} \mathbf{f} = V_{\gamma}(\hat{\theta}).$$

Now, θ_A has mean $\mathbf{G}_y^{-1} \mathbf{f}$ and covariance $(\sigma_e^2/N) \mathbf{G}_y^{-1}$. As the number of data points increases, this covariance estimate becomes more accurate and the following test statistic takes a χ^2

distribution with the number of degrees of freedom equal to the number of additional parameters:

$$\begin{aligned} z^2 &= \boldsymbol{\mu}^T \mathbf{S}^{-1} \boldsymbol{\mu} \\ &= \frac{N}{\sigma_e^2} \mathbf{f}^T \mathbf{G}_y^{-1} \mathbf{f} \\ &= \frac{N}{\sigma_e^2} [V_{\gamma C}(\hat{\theta}) - V_{\gamma}(\hat{\theta})]. \end{aligned} \quad (11)$$

3.2. Comparison with AIC

If the statistic given by Eq. (11) exceeds a certain value, we must reject the null hypothesis and choose the alternative one. Thus we say that, for a predetermined significance level, we have evidence that the additional parameters have non-zero values and so contribute usefully to the model. As the data-set size increases over, say, 30 points the χ^2 threshold becomes approximately proportional to the number of degrees of freedom or the number of additional parameters, $P - P_C$. Variations from proportionality imply that we use a slightly different significance level depending on the number of additional parameters. We can then rewrite the test as, for some constant k ,

$$\frac{N}{\sigma_e^2} [V_{\gamma C}(\hat{\theta}) - V_{\gamma}(\hat{\theta})] > (P - P_C)k, \quad (12)$$

$$kP_C + \frac{N}{\sigma_e^2} V_{\gamma C}(\hat{\theta}) > kP + \frac{N}{\sigma_e^2} V_{\gamma}(\hat{\theta}). \quad (13)$$

Thus we should select the model with the minimum value of

$$\frac{kP}{2} + \frac{N}{2\sigma_e^2} V_{\gamma}(\hat{\theta}) \quad (14)$$

or

$$\ln(p_{\text{MAP}}(\hat{\theta})) + \frac{kP}{2}, \quad (15)$$

where $p_{\text{MAP}}(\hat{\theta})$ is the maximum a posteriori probability for a known value of the innovations variance. This is similar to Akaike’s information criterion (AIC), since for a uniform prior, or as γ^2 increases, $p_{\text{MAP}}(\hat{\theta}) \rightarrow L(\hat{\theta})$.

The main drawback is that the probability of making the wrong selection remains constant as the data-set size is increased. This is an inconsistency and is known, for statistical hypothesis tests, as Lindley's or Jeffrey's paradox [6]. Suppose that the simpler model is correct: even with a large number of data points, the probability with which the classical statistical hypothesis test chooses the more complex model does not fall!

4. Bayesian hypothesis test

4.1. Bayesian sharp null-hypothesis test

In Section 3.1 we introduced the idea of testing two hypotheses:

\mathcal{H}_0 : $\theta_A \in \Theta_0$, the additional parameters all have the value 0.

\mathcal{H}_1 : $\theta_A \in \Theta_1$, at least one of the additional parameters is non-zero.

The Bayesian approach to model selection involves assigning probabilities to models to represent the degree of belief we have in them. We need to assign prior probabilities to the two hypotheses and these are generally denoted as π_0 and π_1 . On the whole we do not have any reason to favour either model so we make each equal to a half. Denote the posterior probabilities of the two hypotheses based on the evidence provided by the measured data as p_0 and p_1 . We reject the more complex model if $p_0 > p_1$. The *Bayes factor* B is generally defined as

$$B = \frac{p_0/\pi_0}{p_1/\pi_1} = \frac{p_0 \pi_1}{p_1 \pi_0}. \quad (16)$$

We can write

$$p_0 = p(\mathcal{H}_0 | \mathbf{d}), \quad (17)$$

$$\pi_0 = p(\mathcal{H}_0), \quad (18)$$

with corresponding expressions for \mathcal{H}_1 . Using Bayes' rule (9),

$$p_0 = \frac{\pi_0 p(\mathbf{d} | \mathcal{H}_0)}{p(\mathbf{d})}, \quad (19)$$

where $p(\mathbf{d})$ takes a value to satisfy $p_0 + p_1 = 1$. But

$$p(\theta_A | \mathcal{H}_0) = \begin{cases} 1, & \theta_A = \mathbf{0}, \\ 0 & \text{otherwise.} \end{cases} \quad (20)$$

So if $\pi_0 = \pi_1$,

$$B = \frac{p(\mathbf{d} | \theta_A = \mathbf{0})}{p_1(\mathbf{d})}, \quad (21)$$

where $p_1(\mathbf{d})$ is the probability of the data occurring under \mathcal{H}_1 . Using this, one can compare a sequence of models constructed by progressively adding terms. The simpler of any two models is to be preferred if $B > 1$. Thus one criterion can be used to make all the comparisons. This criterion is the probability of the data occurring under the model, integrating out all parameters in the model and zeroing all those excluded. From Eq. (4),

$$\left(\frac{\sigma_e^2}{N\gamma^2} \right)^{P/2} \frac{1}{\det(\mathbf{R}_\gamma)^{1/2}} \exp \left[\frac{-NV_\gamma(\hat{\theta})}{2\sigma_e^2} \right]. \quad (22)$$

If we multiply the logarithm of the result by -2 we obtain

$$\frac{NV_\gamma(\hat{\theta})}{\sigma_e^2} + \ln \det \left(\frac{\mathbf{R}_\gamma}{\sigma_e^2} \right) + P [\ln N + \ln(\gamma^2)], \quad (23)$$

which should be minimised.

4.2. Comparison with Bayesian evidence

The principles of Bayesian evidence date back to 1939 [5] although the form shown here is more recent [9]. 'Bayes factors' in the sense of odds ratios of model probabilities [4, 11] perform essentially the same function. The Bayesian evidence criterion is similar to the hypothesis test – a degree of belief is assigned to each model by finding posterior probabilities. We make use of Bayes' rule again:

$$p(\mathcal{M} | \mathbf{d}) = \frac{p(\mathbf{d} | \mathcal{M}) p(\mathcal{M})}{p(\mathbf{d})}. \quad (24)$$

We can use this to decide between two model hypotheses \mathcal{M}_0 and \mathcal{M}_1 in the same fashion as deciding between two statistical hypotheses \mathcal{H}_0 and \mathcal{H}_1 . We do not normally have any reason

to favour either model, and so assign equal values to $p(\mathcal{M})$. The probability of the data realisation occurring in either model, $p(\mathbf{d})$, is common to the expressions for the two models, so the model with the higher value of the *Bayesian evidence* $p(\mathbf{d}|\mathcal{M})$ should be chosen. This can be extended to all model comparisons, so the one with the highest evidence should be selected. In the light of the discussion in this paper, a reasonable formulation of the models might be as follows. Make all models have the same superset of parameters. Assign a Gaussian prior to those parameters that are ‘in’ each model, and a sharp zero prior to those that are ‘out’. These parameters are then integrated out,

$$p(\mathbf{d}|\mathcal{M}_i) = \int p(\mathbf{d}|\boldsymbol{\theta}, \mathcal{M}_i) p(\boldsymbol{\theta}|\mathcal{M}_i) d\boldsymbol{\theta}. \quad (25)$$

But we are considering a particular class of model with a common parameter set, hence we can write

$$p(\mathbf{d}|\mathcal{M}_i) = \int p(\mathbf{d}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{M}_i) d\boldsymbol{\theta}, \quad (26)$$

and so the evidence criterion is the same as the Bayesian hypothesis test.

4.3. Comparison with BIC

The form of the Bayesian information criterion for the models in question is

$$\frac{NV_{\gamma P}(\hat{\boldsymbol{\theta}}_P)}{\sigma_e^2} + P \ln N, \quad (27)$$

where the subscript P denotes the quantities for a P th-order model fit. This clearly comprises some of the terms of Eq. (23). One of the additional terms is $(\ln \det(\mathbf{R}_P/\sigma_e^2))$, which remains approximately constant as redundant model terms are added, and is relatively small in magnitude even for modest N . The $P \ln N$ term is a penalty (‘Ockham’) factor against complexity. By adding $\ln(\gamma^2)$ into this the penalty is changed. Note that, because of overfitting, the $NV_{\gamma}(\hat{\boldsymbol{\theta}})$ term always decreases as the model is enlarged. The Ockham factor should compensate for this decrease. If it is believed that the parameters are likely to take small values, then more weight

should be given to the interpretation that the error-variance decrease is due to better modelling, and therefore we wish to reduce the complexity penalty.

5. Numerical simulation

A Gaussian white noise sequence was filtered using a fourth-order autoregressive model with poles at 0.8, -0.75 , $0.48 + 0.64i$ and $0.48 - 0.64i$. The innovations sequence had a variance of 1. The output sequence was split into 100 000 segments, each of length 100. AIC ($k = 2$), BIC and Bayesian evidence (BEV) were calculated for models up to order 10. In order to perform a simulation of this size, efficient decompositions of the matrices were required. The autocorrelation method was used instead of the covariance method [13] so that the matrices \mathbf{R}_γ were Toeplitz and more easily analysed. This has a slight effect on the results such as introducing a bias into the parameter estimates. A linear autoregressive model was chosen so that a very large number of sequences could be simulated; the example illustrates the relationship between parameter estimates and the preference of one model over another. The behaviour for non-linear models is the subject of ongoing research.

There is no intention to explore the issue of priors for parameter distributions in this paper. It should in principle be possible to incorporate knowledge such as the need for the model to be asymptotically stable, although this would be difficult in practice. Some estimate of the parameter variance is necessary in order to calculate the criteria. The parameters, and especially the last parameters, of an AR model typically have values less than 1. Therefore a standard deviation $\gamma = 0.5$ was chosen for this experiment.

The results of the simulation are shown in Table 1. The fact that, in general, AIC overestimates and BIC underestimates the model order is clearly demonstrated. Since the Ockham factor is proportional to P and the determinant term is small, the comparison of a model with the next smallest one amounts to a confidence-interval analysis for each criterion. This is illustrated in Table 2; there is an erroneous acceptance rate of about 16% for AIC

Table 1
The frequency (%) of preference of each order of AR model

	1	2	3	4	5	6	7	8	9	10
AIC	0.1	1.5	1.0	68.7	11.2	6.1	3.8	2.7	2.3	2.6
BIC	2.3	10.0	2.8	80.4	3.0	0.7	0.2	0.1	0.2	0.3
BEV	0.7	4.5	1.9	81.2	6.5	2.3	1.0	0.6	0.5	0.8

Table 2
The frequency (%) with which each order of model was selected in preference to the next smallest

	2	3	4	5	6	7	8	9	10
AIC	92.6	66.6	98.0	16.5	16.8	16.7	16.5	16.5	16.7
BIC	84.1	45.6	92.0	3.7	3.9	3.8	3.7	3.9	3.9
BEV	88.2	55.0	95.9	8.2	8.5	8.4	8.4	8.6	8.8

that would remain the same if more data were available. The results for BEV are interesting – the rate with which the correct model is selected is about the same, whereas there is more balance between over- and underestimation. We have assumed knowledge of the variance of the excitation sequence; errors in the estimation of this will also lead, effectively, to a slightly different Ockham penalty and hence a different balance between the selection of higher- and lower-order models.

The comparison between models of adjacent order is illustrated further by the distribution of the estimate of the additional parameter, that is the last parameter in the larger model (Fig. 1). The distributions for models of order greater than 4 are nearly the same. The segments for which orders higher than 4 were erroneously selected correspond to estimates of parameters that lie in the tails of these distributions. The distributions for orders 2 and 4 are more easily identifiable as non-zero than that for order 3, and this is reflected in Table 2. This illustrates an interesting phenomenon: some parameters are better determined from the data than others. In this case the last parameter of the third-order AR model is less well-determined. Some models are more easily identified than others; the magnitude of a parameter is not as important as its magnitude with respect to its estimation variance.

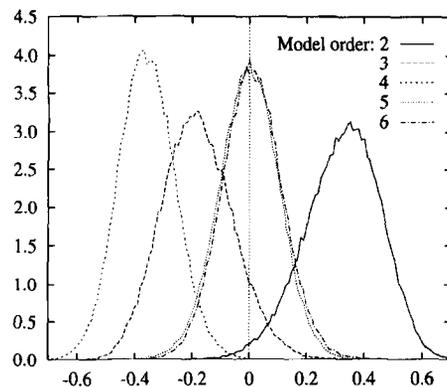


Fig. 1. Distributions of the maximum a posteriori estimates of the last parameters in candidate AR models. The plots were generated from the numerical simulation. The distributions for larger models are similar to those of orders 5 and 6.

It is important to point out that the model, data-set size and parameter priors can be chosen to 'prove' almost anything. The example presented here was chosen to illustrate the main features of the selection criteria.

6. Discussion

A number of very different approaches have been used to assess model suitability. One common basis

has been that the received data must be a reasonable realisation of the model [12, 8]. Some criteria involve more direct statistical or information-based comparisons of the residuals; the F-test arises out of such an approach. These have often yielded expressions similar to AIC, BIC and Bayesian evidence [7]. Bayes factors have already been mentioned [11]. In this paper these expressions have been derived as hypothesis tests on the *parameters*. This shows that the ability of a method to correctly identify the correct terms of a model depends on how well determined the parameters are. That is to say that if the variability of the estimate of a parameter is large with respect to its correct value then it is hard to identify the need to include that term.

The issue of priors on the parameters is a difficult one and often overshadows the benefits of Bayesian techniques. This is true for model comparisons using Bayesian evidence. We have shown that the standard Bayesian methods for model selection can be viewed as hypothesis tests on parameter values. It might therefore be argued that, rather than being an unfortunate necessity in assessing the posterior probabilities of models, these priors should be seen as a useful mechanism of incorporating belief in parameter magnitudes. Nevertheless, the fact that standard non-informative priors are usually unsuitable presents a serious difficulty. There has been no intention to discuss priors in detail in this paper. However, techniques have been developed recently to use training sample and other resampling paradigms [2]. These allow the use of a non-informative prior by inferring a posterior distribution of the parameters from a subset of data samples and then using this as the basis for the model selection test using some or all of the remaining samples.

Parameter-based model selection might make use of numerical techniques such as bootstrap [3] in a variety of ways. They could be an alternative to the training sample methods for dealing with non-informative priors. Estimates of the parameters of additional terms (when comparing nested models) could be obtained from resampled data. The hypothesis that the parameters were zero could be tested using the estimated distributions. Numerical techniques might also be utilised when the statist-

ical assumptions or the model structure restrictions made in this paper were relaxed. For example, if the prediction-error model distribution were non-Gaussian or if the model were non-linear in its parameters, then finding the maximum a posteriori parameters would be more difficult. In such situations the probability of the observed data being a realisation of the model could be assessed for randomly sampled parameter combinations, again generating an estimate distribution of additional parameters.

Finally, the link between the standard criteria and hypothesis tests on parameters may be of wider interest. For example, the detection of changepoints in signals often involves comparison of models in different segments. The suitability of candidate models might be assessed for each segment and compared, and we have seen that some of the criteria for this assessment are equivalent to parameter tests. One might instead look for changes in model parameters. Parameter-based hypothesis testing provides a common framework in which to consider these two approaches.

References

- [1] H. Akaike, "A Bayesian analysis of the minimum AIC procedure", *Ann. Inst. Statist. Math.*, Part A, Vol. 30, 1978, pp. 9–14.
- [2] J.O. Berger and L.R. Pericchi, The intrinsic Bayes factor for model selection and prediction, Tech. Rep. #93-43C, Department of Statistics, Purdue University, 1993.
- [3] B. Efron, "Computers and the theory of statistics: Thinking the unthinkable", *SIAM Rev.*, Vol. 21, No. 4, 1979, pp. 460–480.
- [4] I.J. Good, *Probability and the Weighting of Evidence*, Griffin, London, 1950.
- [5] H. Jeffreys, *Theory of Probability*, Oxford University Press, Oxford, 3rd Edition, 1961.
- [6] D.V. Lindley, "A statistical paradox", *Biometrika*, Vol. 44, 1957, pp. 187–192.
- [7] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
- [8] H. Lütkepohl, "Comparison of criteria for estimating the order of a vector autoregression process", *J. Time Series Anal.*, Vol. 6, No. 1, 1985, pp. 35–52.
- [9] D.J.C. MacKay, Bayesian methods for adaptive models, Chapter 2, Ph.D. Thesis, California Institute of Technology, Pasadena, CA, 1991.

- [10] G. Schwarz, "Estimating the dimension of a model", *Ann. Statist.*, Vol. 6, No. 2, 1978, pp. 461–464.
- [11] A.F.M. Smith and D.J. Spiegelhalter, "Bayes factors and choice criteria for linear models", *J. Roy. Statist. Soc. B.* Vol. 42, No. 2, 1980, pp. 213–220.
- [12] T. Söderström, "On model structure testing in system identification", *Internat. J. Control*, Vol. 26, No. 1, 1977, pp. 1–18.
- [13] C.W. Therrien, *Discrete Random Signals and Statistical Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1992.