USING THE BOOTSTRAP TO SELECT MODELS

Petar M. Djurić

Department of Electrical Engineering
State University of New York at Stony Brook
Stony Brook, NY 11794-2350
internet: djuric@sbee.sunysb.edu

ABSTRACT

In this paper the problem of model selection is addressed by the Bayesian methodology and the bootstrap technique. As a rule for choosing the best model from a set of proposed models, the maximum a posteriori principle is used. The evaluation of the maximum a posteriori probability (MAP) of each model amounts to computation of integrals whose integrands may be very peaked functions. We carry out the integration by importance sampling, where the importance function is a multivariate Gaussian whose samples are obtained by the bootstrap technique. The performance of the MAP rule is examined by computer simulations, and comparisons with the widely used AIC (Akaike information criterion) and MDL (minimum description length) rules are made.

1. INTRODUCTION

Model selection is an important problem in signal processing. Very often when we observe data whose generating mechanism is not completely known, we first propose a set of models for the generating mechanism and then we try to choose the best of them using some predefined criterion. Such applications are common in sonar, radar, image processing, communications, and biomedical signal processing.

The problem of model selection is basically a multiple hypotheses testing problem. In many practical situations a reasonable principle for choosing the best model is the maximum a posteriori probability. To implement this principle, a standard approach is to apply the Bayesian methodology. The strict implementation of the MAP principle requires evaluation of intractable integrals. Their computation is a very difficult task even for low dimensional parameter spaces. To alleviate this difficulty one can use Laplace's asymptotic approximation, which is based on the assumption that the likelihood function is highly peaked near the maximum likelihood parameter estimate. This approximation exploits the Taylor expansion, which in the literature has been used in several variations [1], [2], [7]. In this paper we propose a different approach based on Monte Carlo computation of the required integrals. More specifically, we carry out the integration by importance sampling, where the importance function is the posterior density of the model parameters. This probability density function is assumed to be a multivariate Gaussian which in many cases is a good approximation, even when the number of available samples is relatively small. The sampling from the posterior density is implemented by the bootstrap method [4], [5]. The usage of the bootstrap method to applications in model selection has already been proposed [3], [8], [9], but there the overall selection procedure is either based on a completely different concept or the simulation from the posterior is done in a dissimilar way.

In the paper we first state the problem. Then we derive the criterion and continue with the outline of its implementation by the bootstrap method. The paper ends with three experiments that show the performance of the selection rule and compare it with the popular AIC and MDL.

2. PROBLEM STATEMENT

We formulate the problem using standard assumptions. A set of data $\mathbf{x}$ is observed and a family of parametric models $\mathcal{M}_1$, $\mathcal{M}_2$, ..., $\mathcal{M}_q$, which describe the data generation, is proposed. The models can be of any type, nested or nonnested. The parameters associated with the k-th model are denoted by $\boldsymbol{\theta}_k$, and they are assumed unknown. Each model $\mathcal{M}_k$ is described by a parametric probability distribution function whose form is known and given by $f(\mathbf{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)$. In addition, the a priori probability that the model $\mathcal{M}_k$ is the one that generated the data is given by $p(\mathcal{M}_k)$.

This work was supported by the National Science Foundation under Award No. MIP-9506743

The objective is to choose the best model from the candidates $\mathcal{M}_1$, $\mathcal{M}_2$, ..., $\mathcal{M}_q$, using the MAP principle, i.e., we want to select the model according to

$$\mathcal{M}_{\hat{k}} = \arg \max_{k=1,2,\ldots,q} p(\mathcal{M}_k|\mathbf{x}) \quad (1)$$

where $p(\mathcal{M}_k|\mathbf{x})$ is the a posteriori probability that the model $\mathcal{M}_k$ has generated the data $\mathbf{x}$.

3. CRITERION FOR MODEL SELECTION

The quantity of interest is the a posteriori probability $p(\mathcal{M}_k|\mathbf{x})$, which is obtained from Bayes' theorem

$$p(\mathcal{M}_k|\mathbf{x}) = \frac{f(\mathbf{x}|\mathcal{M}_k)p(\mathcal{M}_k)}{f(\mathbf{x})} \quad (2)$$

where $f(\mathbf{x}|\mathcal{M}_k)$ is the marginal density of the data given they are generated by $\mathcal{M}_k$. The marginal density is found from

$$f(\mathbf{x}|\mathcal{M}_k) = \int_{\Theta_k} f(\mathbf{x}|\boldsymbol{\theta}_k, \mathcal{M}_k) f(\boldsymbol{\theta}_k|\mathcal{M}_k) d\boldsymbol{\theta}_k \quad (3)$$

where $f(\boldsymbol{\theta}_k|\mathcal{M}_k)$ is the prior density of the model parameters, and Θ_k is the parameter space of $\mathcal{M}_k$. If the a priori probability $p(\mathcal{M}_k)$ is uniform, the selection rule (1) simplifies to

$$\mathcal{M}_{\hat{k}} = \arg \max_{k=1,2,\ldots,q} f(\mathbf{x}|\mathcal{M}_k) \quad (4)$$

because the marginal density of the data $f(\mathbf{x})$ in (2) does not depend on the model $\mathcal{M}_k$. This rule can be rewritten as

$$\mathcal{M}_{\hat{k}} = \arg \max_{k=1,2,\ldots,q} \int_{\Theta_k} f(\mathbf{x}|\boldsymbol{\theta}_k, \mathcal{M}_k) f(\boldsymbol{\theta}_k|\mathcal{M}_k) d\boldsymbol{\theta}_k. \quad (5)$$

In many practical problems, the main difficulty in using (5) is the evaluation of the integrals, since almost always they are quite intractable. Their numerical computation is usually so inefficient that in general (5) is of very little use [7]. An approximation of (5) can be obtained by Taylor expanding $\ln f(\mathbf{x}|\boldsymbol{\theta}_k, \mathcal{M}_k) f(\boldsymbol{\theta}_k|\mathcal{M}_k)$ around the maximum likelihood estimates of the model parameters $\hat{\boldsymbol{\theta}}_k$, exponentiating the approximation, and carrying out the integration analytically. Here, we propose a different approach.

The evaluation of the integral

$$I_k = \int_{\Theta_k} f(\mathbf{x}|\boldsymbol{\theta}_k, \mathcal{M}_k) f(\boldsymbol{\theta}_k|\mathcal{M}_k) d\boldsymbol{\theta}_k \quad (6)$$

can be carried out by the Monte Carlo method, which approximates it by

$$\tilde{I}_k = \frac{1}{B} \sum_{b=1}^{B} f(\mathbf{x}|\boldsymbol{\theta}_k^b, \mathcal{M}_k) f(\boldsymbol{\theta}_k^b|\mathcal{M}_k) \quad (7)$$

where the $\boldsymbol{\theta}_k^b$'s are samples from the density $f(\boldsymbol{\theta}_k|\mathcal{M}_k)$. The estimate in (7) may converge very slowly to the true value of I_k, especially if $f(\boldsymbol{\theta}_k|\mathcal{M}_k)$ is a function that quantifies vague prior knowledge about $\boldsymbol{\theta}_k$. A better approach is to rewrite (6) as

$$I_k = \int_{\Theta_k} \frac{f(\mathbf{x}|\boldsymbol{\theta}_k, \mathcal{M}_k) f(\boldsymbol{\theta}_k|\mathcal{M}_k)}{h(\boldsymbol{\theta}_k)} h(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k. \quad (8)$$

where $h(\boldsymbol{\theta}_k)$ is a probability density function whose values are much larger in the region where the magnitude of $f(\mathbf{x}|\boldsymbol{\theta}_k, \mathcal{M}_k) f(\boldsymbol{\theta}_k|\mathcal{M}_k)$ is big than in any other part of the parameter space Θ_k. Then, analogously to the estimate $\tilde{I}_k$ in (7), we can write

$$\hat{I}_k = \frac{1}{B} \sum_{b=1}^{B} \frac{f(\mathbf{x}|\boldsymbol{\theta}_k^b, \mathcal{M}_k) f(\boldsymbol{\theta}_k^b|\mathcal{M}_k)}{h(\boldsymbol{\theta}_k^b)} \quad (9)$$

where the samples $\boldsymbol{\theta}_k^b$ are generated from $h(\boldsymbol{\theta}_k)$. The estimate (9) is known as the importance sampling estimate of (6) because in evaluating the integral the samples $\boldsymbol{\theta}_k^b$ come most often from the regions where the integrand $f(\mathbf{x}|\boldsymbol{\theta}_k^b, \mathcal{M}_k) f(\boldsymbol{\theta}_k^b|\mathcal{M}_k)$ is large. It can be shown that this estimate is unbiased and, provided the importance function is properly chosen, $\hat{I}_k$ has a smaller variance than $\tilde{I}_k$.

Theoretically, if for any $\boldsymbol{\theta}_k^b$, $f(\mathbf{x}|\boldsymbol{\theta}_k^b, \mathcal{M}_k) f(\boldsymbol{\theta}_k^b|\mathcal{M}_k) > 0$ and $h(\boldsymbol{\theta}_k^b) > 0$, then any function $h(\boldsymbol{\theta}_k)$ can be used in (9). However, the art of the importance sampling technique is to choose a function which, (a) is easy to sample from, and (b) resembles the integrand as closely as possible. It should be noted that the effectiveness of the importance sampling is mainly determined by how closely $h(\boldsymbol{\theta}_k)$ approximates $f(\mathbf{x}|\boldsymbol{\theta}_k, \mathcal{M}_k) f(\boldsymbol{\theta}_k|\mathcal{M}_k)$.

Now, since we have assumed that the importance sampling function is a multivariate Gaussian, we write

$$h(\boldsymbol{\theta}_k) = \frac{1}{(2\pi)^{\frac{d_k}{2}} |\mathbf{R}_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{\theta}_k - \boldsymbol{\mu}_k)^T \mathbf{R}_k^{-1} (\boldsymbol{\theta}_k - \boldsymbol{\mu}_k)} \quad (10)$$

where $\boldsymbol{\mu}_k$ and $\mathbf{R}_k$ are the mean vector and the covariance matrix of $\boldsymbol{\theta}_k$, respectively, and d_k is the length of the vector $\boldsymbol{\theta}_k$. This implies

$$\hat{I}_k = \frac{(2\pi)^{\frac{d_k}{2}} |\mathbf{R}_k|^{\frac{1}{2}}}{B} \sum_{b=1}^{B} \frac{f(\mathbf{x}|\boldsymbol{\theta}_k^b, \mathcal{M}_k) f(\boldsymbol{\theta}_k^b|\mathcal{M}_k)}{e^{-\frac{1}{2}(\boldsymbol{\theta}_k^b - \boldsymbol{\mu}_k)^T \mathbf{R}_k^{-1} (\boldsymbol{\theta}_k^b - \boldsymbol{\mu}_k)}} \quad (11)$$

and the selection rule becomes

$$\mathcal{M}_{\hat{k}} = \arg \min_{k=1,2,\ldots,q} \left\{ -\ln \left(\sum_{b=1}^{B} f(\mathbf{x}|\boldsymbol{\theta}_k^b, \mathcal{M}_k) \times \frac{f(\boldsymbol{\theta}_k^b|\mathcal{M}_k)}{e^{-\frac{1}{2}(\boldsymbol{\theta}_k^b - \boldsymbol{\mu}_k)^T \mathbf{R}_k^{-1} (\boldsymbol{\theta}_k^b - \boldsymbol{\mu}_k)}} \right) - \frac{d_k}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{R}_k| \right\}. \quad (12)$$

To implement the selection according to (12), a critical step is the sampling of $\boldsymbol{\theta}_k^b$. As was mentioned before, it is important that the Gaussian distribution of $\boldsymbol{\theta}_k^b$ is as close as possible to $f(\mathbf{x}|\boldsymbol{\theta}_k^b, \mathcal{M}_k)\ f(\boldsymbol{\theta}_k^b|\mathcal{M}_k)$. We propose that the sampling of $\boldsymbol{\theta}_k^b$ is done by the bootstrap method.

4. IMPLEMENTATION BY THE BOOTSTRAP METHOD

The bootstrap method is a computer based technique that is generally applied to evaluate the performance of various estimators. When it was originally proposed, the underlying assumptions were that the observed data are independent and identically distributed [4]. Since then, however, the method has been generalized to treat problems that include stochastic processes where the data are highly dependent [4], [5]. Its implementation does not require theoretical calculations, and it can readily be applied to any data model irrespectively of the model's complexity.

In our problem, we want to sample from a multivariate Gaussian, and then apply the so obtained samples in (12). The bootstrap, however, does not guarantee that the samples come from a Gaussian distribution. Instead, we approximate the distribution of the bootstrap samples by a Gaussian whose mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\mathbf{R}_k$ are determined from $\boldsymbol{\theta}_k^b$. They are found from

$$\boldsymbol{\mu}_k = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{\theta}_k^b \tag{13}$$

and

$$\mathbf{R}_k = \frac{1}{B}\sum_{b=1}^{B} (\boldsymbol{\theta}_k^b - \boldsymbol{\mu}_k)(\boldsymbol{\theta}_k^b - \boldsymbol{\mu}_k)^T \tag{14}$$

where B is the number of bootstrap samples. Clearly, in cases where the Gaussian approximation is inappropriate, this approach is invalid.

The bootstrap is applied as follows. First we find the estimates of the model parameters, $\hat{\boldsymbol{\theta}}_k$. Then using these estimates and the residuals from the original data, we generate B sets of bootstrap data $\mathbf{x}^{*b}$, which are subsequently used for parameter estimation as if the $\mathbf{x}^{*b}$'s were the measured data. From each data set then, we obtain an estimate, $\hat{\boldsymbol{\theta}}_k^b$, which is used in the evaluation of the mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\mathbf{R}_k$, and later in the computation of (12).

5. SIMULATION RESULTS

We tested the MAP rule by performing three experiments. In the testing we included the AIC and MDL rules, which are typically used in everyday practice, and compared them with the MAP. For the AIC rule we used

$$\mathcal{M}_{\hat{k}} = \arg \min_{k=1,2,\ldots,q} \left\{ -2 \ln f(\mathbf{x}|\hat{\boldsymbol{\theta}}_k, \mathcal{M}_k) + 2d_k \right\} \tag{15}$$

and for the MDL

$$\mathcal{M}_{\hat{k}} = \arg \min_{k=1,2,\ldots,q} \left\{ -\ln f(\mathbf{x}|\hat{\boldsymbol{\theta}}_k, \mathcal{M}_k) + \frac{d_k}{2} \ln N \right\} \tag{16}$$

where $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimate of the model parameters and N is the number of observed data samples. It can be argued that this MDL rule is not necessarily the correct one since its penalty for overparametrization is fixed to $(d_k/2)\ln N$ regardless of the model's structure. Here we use (16) anyway because it is the rule that is typically applied by practitioners. We refer to it as a 'naive' MDL rule and denote it as 'MDL.'

In the experiments the data were generated by the following model:

$$x_n = 1 + 100\, n + a_2\, n^2 + w_n \qquad n = 0, 1, 2, ..., 49 \tag{17}$$

where the x_n's are the observed data, and the w_n's are the noise samples. The noise samples were independent and identically distributed according to the Gaussian distribution with mean zero and variance one. The polynomial coefficients and the noise variance were assumed unknown. In the three experiments the coefficient a_2 was varied. In the first experiment $a_2 = 0.0043$, in the second, $a_2 = 0.0076$, and in the third, $a_2 = 0.0135$. There were five candidate models, and they were polynomials of degrees zero, one, two, three, and four, respectively. The number of bootstrap samples B was equal to 200.

From the assumptions, we had $\boldsymbol{\theta}_k^b = [a_0^b\ a_1^b\ \cdots a_{k-1}^b\ \sigma^{2^b}]^T$, where $\mathbf{a}_k^b = [a_0^b, a_1^b, ..., a_{k-1}^b]^T$ are the coefficients of the polynomial of k-th degree, and σ^{2^b} is the noise variance. We could also write

$$f(\mathbf{x}|\boldsymbol{\theta}_k^b, \mathcal{M}_k) = \frac{1}{(2\pi\sigma^{2^b})^{\frac{N}{2}}} e^{-\frac{1}{2\sigma^{2^b}}(\mathbf{x}-\mathbf{s}(\mathbf{a}_k^b))^T(\mathbf{x}-\mathbf{s}(\mathbf{a}_k^b))} \tag{18}$$

where $\mathbf{s}(\mathbf{a}_k)$ is the signal vector whose n-th element is given by $s_n = a_0 + a_1 n + ... + a_{k-1} n^{k-1}$.

Next, we had to specify the priors $f(\boldsymbol{\theta}_k)$, which in general quantified our prior knowledge about $\boldsymbol{\theta}_k$. In our experiments we wanted to introduce as less information about the model parameters as possible. The Jeffreys' noninformative priors were one possibility, but they are improper and proportional to unknown constants. Instead, we assumed that the $f(\boldsymbol{\theta}_k)$'s are equal to a constant over the parameter space $\boldsymbol{\Theta}_k$, that is

$$f(\boldsymbol{\theta}_k^b) = C, \quad \boldsymbol{\theta}_k^b \in \boldsymbol{\Theta}_k. \tag{19}$$

	$k=1$	$k=2$	$k=3$	$k=4$	k=5
MAP	0	7	93	0	0
AIC	0	0	69	21	10
'MDL'	0	0	94	5	1

Table 1: Performance of the MAP, MDL, and AIC rules in 100 trials ($a_2 = 0.0043$). The correct model is $\mathcal{M}_3$.

	$k=1$	$k=2$	$k=3$	$k=4$	k=5
MAP	0	0	100	0	0
AIC	0	0	70	14	16
'MDL'	0	0	92	6	2

Table 2: Performance of the MAP, MDL, and AIC rules in 100 trials ($a_2 = 0.0076$). The correct model is $\mathcal{M}_3$.

Note that this constant is identical for every prior ($k = 1, 2, \cdots, q$), which implies that the parameters' priors did not affect the model selection.

In the experiments, we performed 100 independent trials and the results are shown in Tables 1, 2 and 3. The entries represent the number of times the MAP, AIC, and 'MDL' rules chose the models $\mathcal{M}_1$ – $\mathcal{M}_5$, respectively, out of 100 trials. From the results we see that the AIC did not perform well, even in the case when the coefficient a_2 was relatively large (Table 3). The 'MDL' showed very good performance when $a_2 = 0.0043$ (for the smallest value of a_2 in the three experiments), and did not improve when a_2 was increased, which does not seem appropriate. When the coefficient a_2 is larger, it should be easier to choose the correct degree of the polynomial. The results then suggest that something could be wrong with the penalty of the 'MDL'. The MAP performed perfectly in the second and third experiment and was comparable to the 'MDL' in the first experiment.

6. CONCLUSIONS

A model selection approach based on Bayesian theory and the bootstrap method has been proposed. The principle for model selection is the maximum a posteriori probability. The approach does not require theoretical evaluation of penalties for overparametrization, and it is straightforwardly implemented by exploiting (12), where the samples of the model parameters are generated by the bootstrap method.

	$k=1$	$k=2$	$k=3$	$k=4$	k=5
MAP	0	0	100	0	0
AIC	0	0	76	10	14
'MDL'	0	0	90	7	3

Table 3: Performance of the MAP, MDL, and AIC rules in 100 trials ($a_2 = 0.0135$). The correct model is $\mathcal{M}_3$.

7. REFERENCES

[1] P.M. Djurić, "Model Selection Based on Asymptotic Bayes Theory," 7-th IEEE Signal Processing Workshop on Statistical Signal and Array Processing, pp. 7-10, Quebec City, Canada, 1994.

[2] P.M. Djurić, "A Model Selection Rule for Sinusoids in White Gaussian Noise," IEEE Transactions on Signal Processing, vol. 44, pp. 1744-1751, 1996.

[3] P.M. Djurić, "A Novel Approach to Rank Determination of Multichannel Covariance Matrices," 8-th IEEE Signal Processing Workshop on Statistical Signal and Array Processing, pp. 40-43, Corfu, Greece, 1996.

[4] B. Efron and R. J. Tibshurani, *An Introduction to the Bootstrap*, New York: Chapman and Hall, 1993.

[5] J. S. Hjorth, *Computer Intensive Statistical Methods,* Chapman & Hall, NY, 1994.

[6] M.H. Kalos and P. H. Whitlock, *Monte Carlo Methods*. New York: Wiley, 1986.

[7] R. Kass and A.E. Raftery, "Bayes Factors," Journal of the American Statistical Association, vol. 90, pp. 773-795, 1995.

[8] M. A. Newton and A.E. Raftery, "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap," Journal of the Royal Statistical Society, B, pp. 3-48, 1994.

[9] J. Shao, "Bootstrap Model Selection," Journal of the American Statistical Association, vol. 91, pp. 655-665, 1996.