OPTIMAL SELECTION OF MODEL ORDER FOR A CLASS OF NONLINEAR SYSTEMS USING THE BOOTSTRAP

Abdelhak M. Zoubir *Jonathon C. Ralston*[1] *D. Robert Iskander*

Signal Processing Research Centre, QUT
GPO box 2434, Brisbane, Q. 4001, Australia
a.zoubir@qut.edu.au

ABSTRACT

Nonlinear system identification involves selecting the order of the given model based on the input-output data. A bootstrap model selection procedure which selects the model by minimising bootstrap estimates of the prediction error is developed. Bootstrap based model selection procedures are attractive because the bootstrap observations generated for the model selection can also be used in subsequent inference procedures. The proposed method is simple and computationally efficient.

1. INTRODUCTION

Nonlinear system identification has been receiving increased interest recently due to the need to more accurately characterise real-life phenomena. A basic problem in nonlinear system identification lies in the judicious selection of the order of the given model so as to avoid the effects of under or over parametrisation [5]. Several model order selection procedures exists for the case where the relationship between the input and the output process is linear. Popular techniques are Akaike's information criterion [1], Rissanen's minimum description length criterion [7], and Hannan and Quinn's criterion [4]. These criteria are suggested in the context of estimating the parameters p and q of an autoregressive moving average process of order (p, q). Experimental as well as theoretical results indicated that the model criteria do not yield definitive results. For example, it is known that Akaike's criterion is not consistent. On the other hand, Rissanen's and Hannan and Quinn's criteria are consistent [3]. In the absence of any prior information regarding the physical process that resulted in the data, one is often left with trying different model orders and different criteria and, ultimately, interpreting the different results.

In this paper we describe a procedure for selecting the order of a class of nonlinear models using the bootstrap [2, 9, 10]. Bootstrap procedures for model selection have recently attracted the attention of statisticians [8], but have not seen much application among signal processing practitioners.

Besides the theoretical and empirical properties of bootstrap selection procedures such as the ones discussed in [8], there are good reasons to use a bootstrap model selection procedure. Bootstrap methods are simple and computationally efficient. If one uses a bootstrap approach for the model selection and for the subsequent inference, then the bootstrap observations generated for model selection can also be used in the inference procedure. Thus, the model order selection procedure can be done at no extra computational cost.

Our approach, presented in Section 3, is demonstrated on the Hammerstein series but can be easily extended to Volterra series or other nonlinear models. An outline of the paper follows.

In Section 2, we briefly discuss the Hammerstein series used in this application. Section 3 introduces our approach to selecting the model order of the Hammerstein series using the bootstrap. In Section 4, we demonstrate our approach on simulated data and give some results on the empirical probabilities of selecting various models, before we conclude.

2. THE HAMMERSTEIN SERIES

The Hammerstein series was introduced recently [6] as a model for identification of nonlinear systems driven by non-Gaussian stationary input signals. The Hammerstein series is defined by the input-output relationship

$$\begin{aligned} Y(t) &= \sum_{\tau=-\infty}^{\infty} g_1(\tau)X(t-\tau) + \sum_{\tau=-\infty}^{\infty} g_2(\tau)X(t-\tau)^2 \\ &+ \sum_{\tau=-\infty}^{\infty} g_3(\tau)X(t-\tau)^3 + \cdots, \end{aligned} \tag{1}$$

where $X(t)$, $Y(t)$, $t = 0, \pm1, \pm2, \ldots$, are the input and output of the nonlinear system, respectively, assumed to be stationary, and the functions $\{g_n(\tau)\}$, $n = 1, 2, \ldots$, $\tau \in \mathbb{Z}$, characterise the linear, quadratic, and higher order responses of the system, and are called the *Hammerstein kernels*. The Hammerstein series in (1) is depicted in Figure 1.

We shall consider a Hammerstein series of order p and finite memory m, i.e.,

$$Y(t) = \sum_{k=1}^{p}\sum_{\tau=0}^{m} g_k(\tau)X(t-\tau)^k + \varepsilon(t), \tag{2}$$

where we have allowed for a stationary noise process $\varepsilon(t)$, $t = 0, \pm1, \pm2, \ldots$, which we assume to be a sequence of independently and identically distributed (i.i.d.) variates. We further assume that $X(t)$ and $\varepsilon(t)$ are independent for all t.

[1]Now with CSIRO, Queensland Centre for Advanced Technologies, PO Box 883, Kenmore, Q. 4096, Australia.

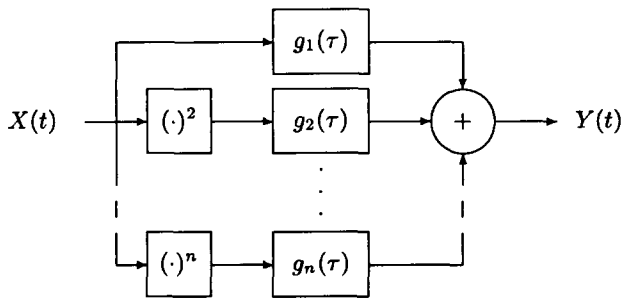

Figure 1. A schematic of an nth order Hammerstein series.

2.1. Estimation of the Hammerstein Kernels

Suppose we are given observations $Y(t)$, $X(t)$, for $t = 0,\ldots,T-1$. Given the parameters p and m, a solution for the kernels is obtained as follows.

Let

$$\boldsymbol{X}^{\odot k}=\begin{pmatrix} X(0)^k & X(-1)^k & \cdots & X(-m)^k \\ X(1)^k & X(0)^k & \cdots & X(1-m)^k \\ \vdots & \vdots & \ddots & \vdots \\ X(T-1)^k & X(T-2)^k & \cdots & X(T-1-m)^k \end{pmatrix}, \tag{3}$$

for $k = 1,\ldots,p$, be a matrix of size $T \times (m+1)$. The notation $\boldsymbol{X}^{\odot k}$ mean that each element in the matrix is raised to the kth power. Let $\boldsymbol{Y} = (Y(0),\ldots,Y(T-1))'$ and $\boldsymbol{\varepsilon} = (\varepsilon(0),\ldots,\varepsilon(T-1))'$. Define for each $k = 1,\ldots,p$ the parameter vector

$$\boldsymbol{g}_k = (g_k(0),\ldots,g_k(m))'.$$

Given $X(t)$ and $Y(t)$, for $t = 0,\ldots,T-1$, one can write (2) in the following vector form

$$\boldsymbol{Y} = \left(\boldsymbol{X}^{\odot 1}\,\boldsymbol{X}^{\odot 2}\cdots\boldsymbol{X}^{\odot p}\right)\begin{pmatrix} \boldsymbol{g}_1 \\ \boldsymbol{g}_2 \\ \vdots \\ \boldsymbol{g}_p \end{pmatrix} + \boldsymbol{\varepsilon}, \tag{4}$$

which is equivalent to

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{g} + \boldsymbol{\varepsilon}, \tag{5}$$

where $\boldsymbol{X}$ is a $T\times(m+1)p$ block matrix with elements $\boldsymbol{X}^{\odot k}$ and $\boldsymbol{g}$ is the $(m+1)p$-vector valued parameter with elements $g_k(\tau)$, $\tau = 0,\ldots,m$, and $k = 1,\ldots,p$.

The least-squares estimate (LSE) of $\boldsymbol{g}$ is then obtained from

$$\hat{\boldsymbol{g}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{6}$$

provided the inverse exists. It is worth noting that the nonlinear system model can be expressed now as a linear one in the unknown parameter $\boldsymbol{g}$.

2.2. Model order selection

In practical situations neither the model order p nor the memory m is known. We wish to select a subset of the parameters $\{g_k(\tau)\}$, $\tau = 0,\ldots,m$, $k = 1,\ldots,p$, to fit the Hammerstein series to $X(t)$ and $Y(t)$.

The problem is the following: given $Y(0),\ldots,Y(T-1)$ and $X(0),\ldots,X(T-1)$ estimate the parameters p and m. This can be formulated as a model selection problem in which we select α from $\{1,\ldots,p\}$ and β from $\{0,\ldots,m\}$ and each α and β corresponds to the model in (2) of order α and memory β, i.e.,

$$Y(t) = \sum_{k=1}^{\alpha}\sum_{\tau=0}^{\beta} g_k(\tau)X(t-\tau)^k + \varepsilon(t).$$

Under α and β, we have

$$\boldsymbol{g}_{\alpha\beta} = \begin{pmatrix} (g_1(0),\ldots,g_1(\beta))' \\ (g_2(0),\ldots,g_2(\beta))' \\ \vdots \\ (g_\alpha(0),\ldots,g_\alpha(\beta))' \end{pmatrix} = \begin{pmatrix} \boldsymbol{g}_1 \\ \boldsymbol{g}_2 \\ \vdots \\ \boldsymbol{g}_\alpha \end{pmatrix},$$

where $\boldsymbol{g}_k = (g_k(0),\ldots,g_k(\beta))'$, $k = 0,\ldots,\alpha$. Assuming the inverse exists, the parameter $\boldsymbol{g}_{\alpha\beta}$ is estimated by the LSE

$$\hat{\boldsymbol{g}}_{\alpha\beta} = \left(\boldsymbol{X}'_{\alpha\beta}\boldsymbol{X}_{\alpha\beta}\right)^{-1}\boldsymbol{X}'_{\alpha\beta}\boldsymbol{Y}, \tag{7}$$

where $\boldsymbol{X}_{\alpha\beta} = (\boldsymbol{X}_\beta^{\odot 1}\,\boldsymbol{X}_\beta^{\odot 2}\cdots,\boldsymbol{X}_\beta^{\odot\alpha})$, and $\boldsymbol{X}_\beta^{\odot k}$ is as (3) replacing m by β, for $k = 1,\ldots,\alpha$. In the following section, we present a method based on the bootstrap that will select α and β by minimising bootstrap estimates of the prediction error.

3. A BOOTSTRAP APPROACH FOR MODEL SELECTION

We assume that $\alpha = p$ and $\beta = m$. The *optimal* model is

$$(\alpha_0,\beta_0) = \max\{(k,\tau) : 1 \le k \le p, 0 \le \tau \le m,\ g_k(\tau) \ne 0\}.$$

Let $\varepsilon^*(t)$, $t = 0 \pm 1, \pm 2,\ldots$ be i.i.d. from the distribution putting mass T^{-1} to

$$\sqrt{\frac{T}{L_T}}\left(\hat{r}(t) - \frac{1}{T}\sum_{t=0}^{T-1}\hat{r}(t)\right),\quad t = 0\ldots,T-1,$$

where

$$\hat{r}(t) = Y(t) - \boldsymbol{X}_{pm}(t)\hat{\boldsymbol{g}}_{pm},$$

with $\boldsymbol{X}_{pm}(t) = \big(X(t)\cdots X(t-m)\,X(t)^2\cdots X(t-m)^2$ $\cdots X(t)^p\cdots X(t-m)^p\big)$ is the tth residual under the largest model $\alpha = p$ and $\beta = m$.

By multiplying the residuals by the factor $\sqrt{T/L_T}$ one increases the variablity among the bootstrap observations and achieves consistency, i.e.,

$$\lim_{T\to\infty}\Pr\{(\hat{\alpha}_{T,L_T},\hat{\beta}_{T,L_T}) = (\alpha_0,\beta_0)\} = 1,$$

provided that L_T is such that $\lim_{T\to\infty}\frac{L_T}{T} = 0$ and $\lim_{T\to\infty} L_T = \infty$.

The bootstrap analog $\hat{\boldsymbol{g}}^*_{\alpha\beta}$ of $\hat{\boldsymbol{g}}_{\alpha\beta}$ is defined in (7) with $Y(t)$ replaced by

$$Y^*(t) = \sum_{k=1}^{\alpha}\sum_{\tau=0}^{\beta}\hat{g}_k(\tau)X(t-\tau)^k + \varepsilon^*(t), \tag{8}$$

$t = 0, \ldots, T-1$. The parameters selected by the bootstrap, denoted by $\hat{\alpha}_{T,L_T}$ and $\hat{\beta}_{T,L_T}$, are then the minimiser of

$$\hat{\Gamma}_{T,L_T}(\alpha,\beta) = \mathsf{E}_* \sum_{t=0}^{T-1} \frac{\left(Y(t) - \sum_{k=1}^{\alpha} \sum_{\tau=0}^{\beta} \hat{g}_k^*(\tau) X(t-\tau)^k \right)^2}{T} \tag{9}$$

over $\alpha = 1, \ldots, p$ and $\beta = 0, \ldots, m$, where E_* is the asymptotic expectation conditioned on the input-output data [2]. A detailed procedure for parameter selection is given in Table 1.

Table 1. Bootstrap-based selection procedure for the parameters of a Hammerstein series model.

1. Select $\alpha = \alpha_{\max}$ $\beta = \beta_{\max}$, and find the estimate $\hat{\boldsymbol{g}}_{\alpha\beta}$ of $\boldsymbol{g}_{\alpha\beta}$ and compute

$$\hat{Y}(t) = \sum_{k=1}^{\alpha_{\max}} \sum_{\tau=0}^{\beta_{\max}} \hat{g}_k(\tau) X(t-\tau)^k$$

2. Compute the residuals $\hat{r}(t)$ as

$$\hat{r}(t) = Y(t) - \hat{Y}(t), \quad t = 0, \ldots, T-1.$$

3. Rescale the empirical residuals

$$\tilde{r}(t) = \sqrt{\frac{T}{L_T}} \left(\hat{r}(t) - \frac{1}{T} \sum_{t=1}^{T} \hat{r}(t) \right),$$

where L_T is such that $\lim_{T\to\infty} \frac{L_T}{T} = 0$ and $\lim_{T\to\infty} L_T = \infty$, e.g. $L_T = T^\gamma$, $0 < \gamma < 1$.

4. For all $1 \le \alpha \le \alpha_{\max}$ and $0 \le \beta \le \beta_{\max}$
   (a) calculate $\hat{\boldsymbol{g}}_{\alpha\beta}$ and $\hat{Y}(t)$ as in step 1.
   (b) Using a pseudo-random number generator, draw independent bootstrap residuals $\tilde{r}^*(t)$ with replacement, from the empirical distribution of $\tilde{r}(t)$.
   (c) Define the bootstrap output

$$Y^*(t) = \hat{Y}(t) + \tilde{r}^*(t).$$

   (d) Using $Y^*(t)$ as the *new* system output, compute the least-squares estimate of $\boldsymbol{g}_{\alpha\beta}$, $\hat{\boldsymbol{g}}^*_{\alpha\beta}$, and calculate

$$\hat{Y}^*(t) = \sum_{k=1}^{\alpha} \sum_{\tau=0}^{\beta} \hat{g}_k^*(\tau) X(t-\tau)^k$$

   and

$$\mathsf{SSE}^*_{T,L_T}(\alpha,\beta) = \frac{1}{T} \sum_{t=0}^{T-1} \left(Y(t) - \hat{Y}^*(t) \right)^2.$$

   (e) Repeat steps (b)–(d) a large number of times (e.g. 100) to obtain a total of B bootstrap statistics

$$\mathsf{SSE}^*_{T,L_T}(\alpha,\beta)_1, \ldots, \mathsf{SSE}^*_{T,L_T}(\alpha,\beta)_B,$$

   and estimate the bootstrap mean-square error

$$\hat{\Gamma}_{T,L_T}(\alpha,\beta) = \frac{1}{B} \sum_{i=1}^{B} \mathsf{SSE}^*_{T,L_T}(\alpha,\beta)_i.$$

5. Choose α and β for which $\hat{\Gamma}_{T,L_T}(\alpha,\beta)$ is a minimum.

4. SIMULATION RESULTS

We now demonstrate the method using two examples of the Hammerstein series. Consider first Kim and Powers' example [5], where the dynamic nonlinear system was given by

$$Y(t) = -0.64\,X(t) + X(t-2) + 0.9\,X(t)^2 + X(t-1)^2 + \varepsilon(t). \tag{10}$$

Here $\varepsilon(t)$ is an additive, i.i.d. non-Gaussian (double exponential) noise process, and $\varepsilon(t)$ and $X(t)$ are independent. The signal to noise ratio at the output of the system was approximately 3 dB. Comparing the models in (10) and (2), it is clear that $p = 2$ and $m = 2$ represents the optimal parameters. We generated a white Gaussian noise input sequence of length T, and evaluated the output using the Hammerstein series as in (10). Using the bootstrap based procedure of Table 1, we found

$$\hat{\Gamma}_{T,L_T}(\alpha,\beta) = \begin{pmatrix} 8.68 & 8.80 & 8.34 & 8.54 & 8.77 \\ 5.29 & 2.01 & \mathbf{1.70} & 2.05 & 2.32 \\ 5.01 & 2.10 & 2.34 & 2.37 & 3.14 \\ 4.62 & 2.35 & 2.81 & 3.87 & 7.45 \\ 4.75 & 3.00 & 3.07 & 4.51 & 12.94 \end{pmatrix} \tag{11}$$

and

$$\hat{\Gamma}_{T,L_T}(\alpha,\beta) = \begin{pmatrix} 12.59 & 12.37 & 12.01 & 11.87 & 11.69 \\ 5.26 & 1.80 & \mathbf{1.02} & 1.08 & 1.17 \\ 5.28 & 1.88 & 1.09 & 1.22 & 1.33 \\ 5.06 & 1.92 & 1.24 & 1.36 & 1.43 \\ 5.11 & 1.91 & 1.31 & 1.40 & 1.53 \end{pmatrix} \tag{12}$$

for $T = 32$ and $T = 64$, respectively, where $\alpha_{\max} = 5$ and $\beta_{\max} = 4$.

The bootstrap based method clearly shows the optimal solution, where the minimum of $\hat{\Gamma}_{T,L_T}(\alpha,\beta)$ corresponds to $p = 2$ and $m = 2$ (in bold). Note particularly how the estimated mean-square error *increases* for model orders above $p = 2$ and $m = 2$. In this case the parameter γ was set to 0.51.

We have evaluated also the empirical probability of selecting a particular model for the example given in (10). For 100 independent runs, the empirical probability was (in

percentages)

$$\hat{\mathsf{Pr}} = \begin{pmatrix} 6 & 0 & 1 & 0 & 0 \\ 11 & 7 & \mathbf{43} & 8 & 3 \\ 0 & 0 & 8 & 1 & 3 \\ 0 & 0 & 2 & 1 & 1 \\ 0 & 0 & 3 & 1 & 1 \end{pmatrix} \tag{13}$$

and

$$\hat{\mathsf{Pr}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & \mathbf{94} & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{14}$$

for $T = 32$, and $T = 64$, respectively, where

$$\hat{\mathsf{Pr}} = \mathsf{Pr}\{(\hat{\alpha}_{T,L_T}, \hat{\beta}_{T,L_T}) = (\alpha_0, \beta_0)\}.$$

Consider another example of the Hammerstein series, where the nonlinear dynamic is given by

$$\begin{aligned} Y(t) &= 0.4X(t) + 0.3X(t-1) + 0.2X(t-2) \\ &+ 0.1X(t-3) + X(t)^2 + X(t-1)^2 \\ &+ 0.5X(t-2)^2 + 4X(t-3)^2 + 0.5X(t)^3 \\ &+ X(t-1)^3 + 0.5X(t-2)^3 + X(t)^4 + X(t-2)^4 \\ &+ \varepsilon(t). \end{aligned}$$

In this case $p = 4$ and $m = 3$. We have repeated the bootstrap procedure for $\alpha_{\max} = 7$ and $\beta_{\max} = 6$ with all other settings as in the previous example. The empirical probabilities of selecting a particular model (evaluated over 100 runs) for $T = 64$ were

$$\hat{\mathsf{Pr}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{83} & 4 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}. \tag{15}$$

Discussion. In the test we ran, it was clear that the method proposed performs well. Over 100 replications, the empirical probability that the method selects the true order was close to 1, even when the data length was only $T = 64$. Increasing the data length improved the performance as long as the choice of γ was less than 1. If γ was close or equal to one, we observed that the method becomes inconsistent. This result is in accordance with what was discussed by Shao in [8]. One may argue that the choice of L_T may be a problem in practical situations. Some guidelines as to the choice of L_T in linear regression are given in [8]. In our simulation we did not observe any problem with the choice of γ, except where γ was close or equal to one and T is large. We also ran the simulations with different models and different noise levels (lower SNR). Again, the performance was promising.

The procedure presented in this paper was developed for the Hammerstein series. This is by no means a limitation. If one chooses to use a Volterra series one would still be able to formulate a linear regression like (5) (see for example [5]).

The bootstrap model selection procedure depicted in Table 1 can be altered so as to draw much less than T bootstrap observations. The bootstrap sampling procedure is consistent if we draw L_T bootstrap observations with $\lim_{T\to\infty} L_T \to \infty$ and $\lim_{T\to\infty} L_T/T \to 0$ [8]. Tests with this technique were performed and promising results were obtained. An extensive analysis has not been performed as yet and results will be presented elsewhere.

5. CONCLUSIONS

We have proposed a procedure for determining the order and the memory of a Hammerstein series using the bootstrap. The method is based on minimising bootstrap estimates of the prediction error. We have also presented some simulation results based on a nonlinear model used in [5] which can be represented by a Hammerstein series. At very low SNR and with only a small size of data points we have also been able to achieve a high probability of selecting the true order, irrespective of the statistical distributions of the input and the noise time series. The method presented is not restricted to Hammerstein series but can be easily applied to other nonlinear systems.

REFERENCES

[1] H. Akaike, "A New Look at the Statistical Model Identification", *IEEE Transaction on Automatic Control*, vol. 19, pp. 716–723, 1974.

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

[3] E. J. Hannan, "The estimation of the order of an ARMA process," *Annals of Statistics*, vol. 8, pp. 1071–1081, 1980.

[4] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression", *J. R. Statist. Soc. B*, vol. 41, pp. 190–195, 1979.

[5] K. I. Kim and E. J. Powers, "A digital method of modelling quadratically nonlinear systems with a general random input," *IEEE Transactions on ASSP*, vol. 36, no. 11, pp. 1758–1769, 1988.

[6] J. C. Ralston, A. M. Zoubir, and B. Boashash, "Identification of a class of nonlinear systems under stationary non-Gaussian excitation," *IEEE Transactions on Signal Processing*, 1996. To appear.

[7] J. Rissanen, "A universal prior integers and estimating by minimum description length," *Annals of Statistics*, vol. 11, pp. 416–431, 1983.

[8] J. Shao, "Bootstrap Model Selection," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 655–665, 1996.

[9] A. M. Zoubir, "Bootstrap: Theory and Applications," in *Advanced Signal Processing Algorithms, Architectures and Implementations* (T. Luk, ed.), vol. 2027, (San Diego), pp. 216–235, SPIE, 1993.

[10] A. M. Zoubir and B. Boashash, "The Bootstrap: signal processing applications", *IEEE Signal Processing Magazine*, 1997 To appear.