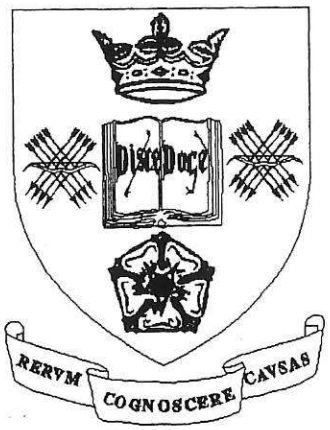
# Improved Parameter Estimates for Nonlinear Dynamical Models Using a Bootstrap Method

H.L. Wei and S.A. Billings
Department of Automatic Control and Systems Engineering, University of Sheffield
Mappin Street, Sheffield, S1 3JD, UK




Research Report No. 888

Department of Automatic Control and Systems Engineering
The University of Sheffield
Mappin Street, Sheffield,
S1 3JD, UK

April 2005

# Improved Parameter Estimates for Nonlinear Dynamical Models Using a Bootstrap Method

H.L. Wei and S.A. Billings
Department of Automatic Control and Systems Engineering, University of Sheffield
Mappin Street, Sheffield, S1 3JD, UK
S.Billings@Sheffield.ac.uk , W.Hualiang@Sheffield.ac.uk

April 6, 2005

It is known that the least-squares class of algorithms produce unbiased estimates providing certain assumptions are met. There are many practical problems, however, where the required assumptions are violated. Typical examples include nonlinear dynamical system identification problems, where the input and output observations are affected by measurement uncertainty and possibly correlated noise. This will result in biased least-squares estimates and the identified model will exhibit poor generalization properties. Model estimation for this type of error-in-variables (EIV) problem is investigated in this study, and a new identification scheme based on a bootstrap algorithm is proposed to improve the model estimates for nonlinear dynamical system identification.

**Keywords**: Bootstrap, error-in-variables, measurement uncertainty, NARX and NARMAX models, nonlinear systems identification, orthogonal least-squares, parameter estimation

## 1. Introduction

A wide range of nonlinear dynamical systems can be described by the NARX (N*onlinear* A*uto*R*egressive* with e*X*o*genous* inputs) model

$$y(t) = f(y(t-1),\cdots,y(t-n_y),u(t),\cdots,u(t-n_u)) + e(t) \tag{1}$$

where $f$ is an unknown nonlinear mapping, $u(t)$, $y(t)$ and $e(t)$ are the sampled input, output and noise sequences, respectively; $n_u$ and $n_y$ are the maximum input and output lags. The NARX model (1) can be constructed using a variety of local or global basis functions including polynomials, kernel functions, splines, ridge and radial bases, neural networks and wavelets. The NARX model can often be expressed as a linear-in-the-parameters form

$$y(t) = \sum_{m=1}^{M_0} \theta_m \Pi_m(t) + e(t) \tag{2}$$

where $\Pi_m(t) = \Pi_m(\varphi(t))$ are model terms generated from the regression vector $\varphi(t) = [y(t-1),\cdots, y(t-n_y), u(t),\cdots, u(t-n_u)]^T$, $\theta_m$ are unknown parameters, and $M_0$ is the number of total potential model terms involved. The linear-in-the-parameters model (2) can be solved using the orthogonal least-squares and error reduction ratio (OLS-ERR) algorithm (Korenberg *et al.* 1988, Billings *et al.* 1989, Chen *et al.* 1989) and revised versions of these algorithms (Zhu and Billings 1996). In practice, only a relatively small number of significant model terms may be necessary to describe a given nonlinear system with a given accuracy. In other words, there exists an integer $M$ (generally $M << M_0$), such that the model

$$y(t)=\sum_{m=1}^{M}\beta_m\Pi_{i_m}(t)+e(t)=\sum_{m=1}^{M}\beta_m X_m(t)+e(t) \tag{3}$$

provides a satisfactory representation over the range considered for the measured input-output data, where $X_m(t)=\Pi_{i_m}(t)$ and $i_m\in\{1,2,\cdots,M_0\}$ for $m$=1,2, ..., $M$.

It is known that the least-squares class of algorithms and the standard statistical analysis for these algotihms require certain assumptions, these include: 1) the system input $u(t)$ and the output $y(t)$ are measured without errors, this is to guarantee that the regressor vector $\varphi(t)=[y(t-1),\cdots, y(t-n_y), u(t),\cdots, u(t-n_u)]^T$ and thus the regression matrix $X=[X_1,\cdots,X_M]$, with $X_j=[X_j(1),\cdots,X_j(N)]^T$ for $j$=1,2, ..., $M$, is deterministic, where $N$ is the length of the observational data set; 2) $e(t)$ does not depend on the regression matrix $X=[X_1,\cdots,X_M]$; 3) the error $e(t)$ is identically distributed as $e(t)\sim N(0,\sigma_e^2)$ (the assumption of a normal distribution is required for some standard statistical analysis methods but not for parameter estimation). For most practical nonlinear identification problems, however, these assumptions may not be satisfied and this can lead to biased estimates. It was noted by Vinod and Ullan (1981) that "many researchers tend to forget that the term '*unbiased*' means that in a large number of repetitions the average of the deviations of the estimator from the true parameter is zero. In practice, we do not generally have a large number of replications, and mere averaging to zero of large positive and negative deviations may not be satisfactory." In fact, in most cases only one or at best a few sets of input-output observational data are available for identification problems. In summary, whatever basis functions are chosen and whatever modelling schemes are adopted, it may be difficult to obtain unbiased models in some practical nonlinear system identification problems, where the system input and output variables are measured with errors or the system is corrupted by correlated noise.

Various studies have been made both to investigate and to reduce the bias in parameter estimation. The total-least-squares (TLS) method (Golub and Van Loan 1980, 1989) is a well known regression approach which was introduced to solve the error-in-variables problem by taking into account the effects of additive noise on both sides of the regression equation (3). However, the TLS method has a major limitation when applied to nonlinear dynamical system identification problems: it is assumed that the SNR for each column of the regression vector $X=[X_1,\cdots,X_M]$ is the same. For most nonlinear systems described by a NARX model, this assumption cannot be satisfied. Although some modified versions, such as the constrained total-least-squares (CTLS) (Abatzoglou et al 1991) and the structured total-least-squares (STLS) algorithms (De More 1994), provide better results compared with the original TLS method when handling general linear estimation problems, no advantages are gained when these methods are applied to identify nonlinear models. For example, the CTLS method requires a priori information of the possible range of the unknown parameters, and the STLS method requires that the regression matrix is of a particular structure including the Hankel or Toeplitz matrix form. These assumptions are not realistic for NARX modelling.

One approach which has been used to reduce bias and minimize the variance of the estimates is the extended least-squares (ELS) class of algorithms (Soderstrom and Stoica 1989, Billings and Voon 1984), where the lagged noise variables $e(t-k)$, as well as the lagged input and output variables $u(t-j)$ and $y(t-i)$ are introduced to form an extended regression vector $\widetilde{\varphi}(t)=[y(t-1),\cdots, y(t-n_y), u(t),\cdots, u(t-n_u), e(t-1),\cdots,$

$e(t-n_e)]^T$. The well known NARMAX methodology (Leontaritis and Billings 1985, Billings and Chen 1998), where the extended regression vector $\tilde{\varphi}(t)$ is employed to accommodate correlated or coloured noise and possible nonlinear noise effects including power and cross product noise terms, can produce unbiased estimates for most nonlinear dynamical systems. By introducing the lagged noise variables $e(t-k)$, the bias associated with a NARX model can often be eliminated. For a more comprehensive discussion on the NARMAX methodology, see Billings and Chen (1998) and the references therein.

In this paper, an alternative approach is proposed to effectively reduce and minimize the variance of the parameter estimates associated with a NARX model by introducing a bootstrap method. The bootstrap method, which was initially introduced by Efron (Efron 1979a,1979b) for point estimation and statistical analysis for non-parametric models, has attracted much attention in recent years (Duchesne and MacGregor 2001, Kukreja et al 2004). The method, which in fact can be viewed as a special class of Monte Carlo methods, provides a simulation procedure for parameter estimation and statistical analysis with the assumption that the associated error is an independent, identically distributed random sequence with zero mean. These properties make the bootstrap method more flexible for solving estimation problems where conventional approaches might be inappropriate. The main objective in the current study is to develop a new improved algorithm to produce unbiased parameter estimates for any linear-in-the-parameters model estimation whatever basis functions are used and whatever modelling scheme is employed.

The paper is organized as follows. The effects of measurement errors on parameter estimates are discussed in Section 2 and a result on asymptotic bias is given. The basic idea of the bootstrap approach is briefly summarized in Section 3. The new algorithm, where the bootstrap method is developed for error-in-variables dynamical system identification, is proposed in Section 4. Several examples are given in Section 5 to illustrate the application of the proposed algorithm to model estimation of error-in-variables dynamical systems. Finally, the study is summarized in Section 6.

## 2. The Effect of Measurement Uncertainty on Parameter Estimates

Consider a nonlinear dynamical system described by the NARX representation (1), where either the input $u(t)$ or the output $y(t)$, or both are measured with errors. The NARX model can often be converted into a linear-in-the-parameters form (2) by introducing some basis functions. Assume that the initial linear-in-the-parameters model involves a total of $M_0$ model terms. It has been demonstrated (Wei and Billings 2004) that the forward OLS-ERR algorithm (Korenberg *et al.* 1988, Billings *et al.* 1989, Chen *et al.* 1989) can be used to detect the model structure and select the significant model terms for this error-in-variables identification problem. The final identified model is of the form (3) with $M$ ($M << M_0$) significant model terms. The detail of the model structure and term detection problem will not be repeated here rather the focus will be on the effects of measurement uncertainty on the parameter estimation.

Consider initially the noise free case, where it is assumed that both the input and the output of the system are 'clean' (measured without any errors), and where the system can be represented using a linear-in-the-parameters model

$$\bar{\mathbf{y}} = \bar{X}\beta + \mathbf{e} \tag{4}$$

where $\breve{\mathbf{y}} = [\breve{y}(1), \cdots, \breve{y}(N)]^T$ is the vector of true output values, $\breve{X} = [\breve{X}_1, \cdots, \breve{X}_M]$ ,with $\breve{X}_j = [\breve{X}_j(1), \cdots, \breve{X}_j(N)]^T$ for $j$=1,2, ..., $M$, is the corresponding regression matrix, $\mathrm{e} = [e(1), \cdots, e(N)]^T$ is the model error vector with $\mathrm{E(e)} = 0$ and $\mathrm{E(ee}^T) = \sigma_e^2 I$, $\beta = [\beta_1, \cdots, \beta_M]^T$ is the unknown parameter vector. Let $\mathbf{y}$ and $X$ be the actually observed vector (matrix) for $\breve{\mathbf{y}}$ and $\breve{X}$, respectively. Since the true values $\{\breve{y}(t)\}_{t=1}^N$ and $\{\breve{u}(t)\}_{t=1}^N$ for the system output and input can never be perfectly measured, there will always exist some measurement errors $\delta$ and $\Lambda$ between $\mathbf{y}$ and $\breve{\mathbf{y}}$, and between $X$ and $\breve{X}$, respectively, in the sense that

$$\delta = \breve{\mathbf{y}} - \mathrm{y} \tag{5}$$

$$\Lambda = \breve{X} - X \tag{6}$$

where $\delta = [\delta(1), \cdots, \delta(N)]^T$ , $\Lambda = [\Lambda_1, \cdots, \Lambda_M]$ ,with $\Lambda_j = [\Lambda_j(1), \cdots, \Lambda_j(N)]^T$ for $j$=1,2, ..., $M$ . Three different approaches can be used to analyse the effects of the measurement errors on the parameter estimates, namely, the asymptotic, perturbation and simulation approaches (Hodges and Moore 1972, Davies and Hutton 1975, Beaton et al 1976, Chatterjee and Hadi 1988). Following the asymptotic approach (Chatterjee and Hadi 1988), the discussion below will reveal that unbiased and consistent estimates cannot be obtained for any NARX model, using the ordinary least squares algorithm, where the regression vector is chosen as $\varphi(t) = [y(t-1), \cdots, y(t-n_y), \ u(t), \cdots, \ u(t-n_u)]^T$ , under the condition that the elements of the regression matrix $X$ are measured with errors, even if only part of $X$ are measured with errors. This means that measurement errors in either the input $u(t)$ or the output $y(t)$ will result in inconsistent, asymptotically biased estimates.

The following assumptions are considered to be satisfied:

(*A1*) The measurement error $\delta$ and the model error e are uncorrelated with the true regression matrix $\breve{X}$, that is, $\mathrm{E}(\breve{X}^T\delta) = 0$ and $\mathrm{E}(\breve{X}^T\mathrm{e}) = 0$.

(*A2*) The measurement error $\Lambda$ is a random matrix with $\mathrm{E}(\Lambda) = 0$, and is uncorrelated with the observation vector y, the model error e and the output measurement error $\delta$, that is, $\mathrm{E}(\Lambda^T\delta) = 0$ and $\mathrm{E}(\Lambda^T\mathrm{e}) = 0$.

(*A3*) The matrices $(X^TX)/N$ and $(\Lambda^T\Lambda)/N$ converge to limit matrices $C_{xx}$ (non-singular) and $D$, respectively, when the data length $N$ approaches infinity, that is,

$$\operatorname*{plim}_{N\to\infty}\left(\frac{X^TX}{N}\right) = C_{xx} \tag{7}$$

$$\operatorname*{plim}_{N\to\infty}\left(\frac{\Lambda^T\Lambda}{N}\right) = D \tag{8}$$

Note that the symbol 'plim $g(N)= g_0$ ' refers to the 'probability limit' of a given function $g(N)$, implying that the probability distribution of the associated function $g(N)$ approaches $g_0$ as $N \to \infty$. In practice, the true values

for $\breve{y}$ and $\breve{X}$ are nearly always unknown. The unknown parameter vector $\beta$ is usually estimated from the actually observed values and the ordinary least-squares estimator is given by

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y} \tag{9}$$

From (4), (5), (6) and (9)

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T [X\beta + \Lambda\beta + (\mathrm{e} - \delta)] \\ &= \beta + (X^T X)^{-1} (\breve{X}^T - \Lambda^T)[\Lambda\beta + (\mathrm{e} - \delta)] \\ &= \beta + (X^T X)^{-1} [\breve{X}^T \Lambda\beta - \Lambda^T \Lambda\beta + \breve{X}^T (\mathrm{e} - \delta) - \Lambda^T (\mathrm{e} - \delta)]\end{aligned} \tag{10}$$

Thus,

$$\begin{aligned}\operatorname*{plim}_{N\to\infty} \hat{\beta} &= \beta + \operatorname*{plim}_{N\to\infty} \left( \frac{X^T X}{N} \right)^{-1} \\ &\quad \times \left[ \operatorname*{plim}_{N\to\infty} \left( \frac{\breve{X}^T \Lambda}{N} \right) \beta - \operatorname*{plim}_{N\to\infty} \left( \frac{\Lambda^T \Lambda}{N} \right) \beta + \operatorname*{plim}_{N\to\infty} \left( \frac{\breve{X}^T (\mathrm{e} - \delta)}{N} \right) - \operatorname*{plim}_{N\to\infty} \left( \frac{\Lambda^T (\mathrm{e} - \delta)}{N} \right) \right] \\ &= [I - C_{xx}^{-1} D]\beta\end{aligned} \tag{11}$$

Therefore,

$$\operatorname*{plim}_{N\to\infty} \hat{\beta} - \beta = -C_{xx}^{-1} D\beta \tag{12}$$

From (12) it can be seen that $\hat{\beta}$ is not a consistent estimate of $\beta$, the magnitude of the asymptotic bias for the estimate is characterized by $C_{xx}^{-1} D\beta$, which shows that parameter estimates for all model terms are affected, even those that are measured without errors. Keep in mind that $X_j$ for $j$=1,2,...,$M$ are generated by the regression vector $\varphi(t) = [y(t-1), \cdots, y(t-n_y), u(t), \cdots, u(t-n_u)]^T$, measurement errors in either the input $u(t)$ or the output $y(t)$ will result in an inconsistent, asymptotically biased estimate. This is why the lagged noise variables $e(t-k)$ are included in the extended regression vector $\tilde{\varphi}(t)$ in the extended least squares algorithm, to accommodate correlated or coloured noise and possible nonlinear noise effects to produce unbiased estimates. But this solution is only appropriate for noise on the output and is not effective if there is noise on the input. In the present study, a new alternative approach will be introduced to improve parameter estimation for both the case of noise on the output and on the input measurements. In the new algorithm, the lagged noise variables $e(t-k)$ do not directly appear in the regression vector in an explicit way.

Further results can be obtained from (12) by introducing a distance measure, say by taking the standard $L^2$-norm (the squared norm). Taking the $L^2$-norm of both sides of (12), yields

$$\left\| \operatorname{plim} \hat{\beta} - \beta \right\| \le \left\| C_{xx}^{-1} D \right\| \left\| \beta \right\| \tag{13}$$

Thus,

$$\frac{\left\|\text{plim}\hat{\beta}-\beta\right\|}{\|\beta\|}\le\left\|C_{xx}^{-1}D\right\| \tag{14}$$

Assume the measurement errors $\Lambda_1,\Lambda_2\cdots,\Lambda_M$ are uncorrelated with each other, $D$ will then reduce to a diagonal matrix with the diagonal entries $\sigma_1^2,\sigma_2^2,\cdots,\sigma_M^2$. Let $\lambda_1^2,\lambda_2^2,\cdots,\lambda_M^2$ be the eigenvalues of the matrix $C_{xx}$. From (14) it follows by using the properties of the $L^2$-norm that

$$\frac{\left\|\text{plim}\hat{\beta}-\beta\right\|}{\|\beta\|}\le\frac{\sigma_{\max}^2}{\lambda_{\min}^2} \tag{15}$$

where $\sigma_{\max}^2=\max\{\sigma_m^2,1\le m\le M\}$ and $\lambda_{\min}^2=\min\{\lambda_m^2,1\le m\le M\}$. Eq. (15) shows that the upper bound for the relative asymptotic bias is large if the matrix $X^TX$ is close to singularity with small eigenvalue $\lambda_{\min}^2$, or the variance of the measurement errors are large ( $\sigma_{\max}^2$ is large).

## 3. The Bootstrap Method

The bootstrap method, which is similar to the jackknife method, is a Monte Carlo simulation based statistical technique for estimating standard errors and bias, and requires a minimum number of mild assumptions and does not require a large number of samples (Efron and Tibshirani 1993). The bootstrap estimate and the associated standard errors for an unknown parameter whose probability distribution is unavailable are summarized below.

Assume that a random sample $\xi=\{\xi_1,\xi_2,\cdots,\xi_n\}$ from an unknown probability distribution $F$ has been observed, and an estimate for a parameter $\theta=\pi(F)$ is of interest on the basis of the observation $\xi$. A bootstrap method is associated with the notion of a bootstrap sample. Let $\hat{F}$ be an empirical distribution function of $F$, which is defined to be

$$\xi\prec F\overset{e.d.}{\rightarrow}\xi^*\prec\hat{F} \tag{16}$$

where $\xi=\{\xi_1,\xi_2,\cdots,\xi_n\}$ and $\xi^*=\{\xi_1^*,\xi_2^*,\cdots,\xi_n^*\}$, the symbol '$\xi\prec F$' implies that the random sample $\xi$ comes from the distribution $F$, the symbol '$\overset{e.d.}{\rightarrow}$' implies that $\hat{F}$ assigns to a set $\Omega=\Omega(F,\xi)$ in the sample space of $\xi$ with an empirical probability $p=1/n$ in the sense that for any $i,j\in\{1,2,\cdots,n\}$

$$\Pr\{\xi_i^*=\xi_j\}=\frac{1}{n} \tag{17}$$

The star notation indicates that $\xi^*$, which is referred to as a *bootstrap sample* of size *n*, is not the actual data set $\xi$ but rather a randomized, or *resampled*, version of $\xi$. For example, a possible bootstrap sample for a given random sample of size 10, $\xi=\{\xi_1,\xi_2,\cdots,\xi_{10}\}$, is $\xi=\{\xi_6,\xi_{10},\xi_2,\xi_7,\xi_1,\xi_4,\xi_5,\xi_4,\xi_3,\xi_7\}$ with $\xi_1^*=\xi_6$, $\xi_2^*=\xi_{10}$, $\cdots$, $\xi_{10}^*=\xi_7$.

From the *plug-in principle*, the *plug-in estimate* of $\theta=\pi(F)$ can be calculated by

$$\hat{\theta} = \pi(\hat{F}) \tag{18}$$

This means that the function $\theta = \pi(F)$ of the unknown distribution $F$ can be estimated using the same function of the empirical distribution $\hat{F}$ with the form of (18). In practice the function $\pi$ is unknown, and the unknown parameter $\theta$ is often approximated using a statistical estimator $\hat{\theta} = s(\xi)$ based on the observed data $\xi = \{\xi_1, \xi_2, \cdots, \xi_n\}$. The *bootstrap replication* of $\hat{\theta}$ is therefore given by $\hat{\theta}^* = s(\xi^*)$. The function $s(\cdot)$ here can be any statistic of interest, for example, the mean, the standard derivation, or the second moment of the sampled data.

To obtain an efficient *bootstrap estimate*, it is often required to generate a number of independent bootstrap samples, say resampling the original observed data $\xi = \{\xi_1, \xi_2, \cdots, \xi_n\}$ $B$ times, where the number $B$ is ordinarily chosen in the range 25-200 and $B$=50 is often enough to give a good estimate (Efron and Tibshirani 1993). Assume that the bootstrap replication at step $b$ is $\hat{\theta}^*(b) = s(\xi^*(b))$ for $b$=1,2,...,$B$. The bootstrap estimate $\hat{\theta}^{*B}$ can then be obtained by synthesizing all the individual bootstrap replications $\hat{\theta}^*(b) = s(\xi^*(b))$ via a given function $\phi$ with the form $\hat{\theta}^{*B} = \phi(\theta^*(1), \cdots, \theta^*(B))$, where $g$ is some known function. For example, the bootstrap estimate of the standard error of $\hat{\theta} = s(\xi)$ from the observed data $\xi$ is given by

$$\hat{s}^B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left[ \hat{\theta}^*(b) - \bar{\hat{\theta}} \right]^2 \right\}^{1/2} \tag{19}$$

where $\bar{\hat{\theta}} = (1/B) \sum_{b=1}^{B} \hat{\theta}^*(b)$. Readers are referred to Efron and Tibshirani (1993) for more details of the bootstrap methods.

## 4. Applying the Bootstrap Method in Parameter Estimation

It is known that a variety of local and global basis functions, including polynomials, kernel functions, splines, radial bases, neural networks and wavelets, can be adopted to describe a nonlinear dynamical system in a linear-in-the-parameters model form (2). Such a functional expansion based modelling approach often involves several steps: 1) choose appropriate basis functions; 2) select the significant variables; 3) detect the model structure and determine the significant model terms; 4) estimate the unknown parameters; and 5) assess the model performance. Procedures 1), 2), 3) and 5) have been extensively studied in the literature. This study focuses on the fourth step with an intention to reduce any bias in the parameter estimates and therefore to improve the capability of the model to predict the future behaviour of the system.

### 4.1 The bootstrap estimator

Assume that the significant variables and model terms have been correctly selected, and that the system output can be expressed using a linear-in-the-parameters model form given by (3). The estimates for the unknown parameter $\beta$ can then be calculated by (9). The model residuals (one-step-ahead prediction errors) can be estimated as

$$\varepsilon = \mathrm{y} - \hat{\mathrm{y}} = \mathrm{y} - X\hat{\beta} \tag{20}$$

where $\varepsilon = [\varepsilon(1), \cdots, \varepsilon(N-l)]^T$ with $E(\varepsilon) = 0$, where $N$ is the length of the observed data set and $l$ is the model order defined to be the maximum lag of the system input or output, whichever is the larger. Let $\Xi = \{\varepsilon(1), \cdots, \varepsilon(N-l)\}$. By performing the bootstrap resampling theory described in Section 3 over the set $\Xi$, a bootstrap sample $\Xi^* = \{\varepsilon^*(l+1), \cdots, \varepsilon^*(N)\}$ will be obtained. The *bootstrap response* (*output*) of the system is defined to be $y^*(t) = y(t)$ for $1 \le t \le l$ and

$$y^*(t) = \hat{y}(t) + \mu\varepsilon^*(t) \tag{21}$$

for $l+1 \le t \le N$, where $\mu$ is some constant satisfying $0 < \mu \le 1$. Corresponding to the above bootstrap observations, a bootstrap estimate of $\beta$ can be obtained by substituting the bootstrap output vector $\mathrm{y}^*$ and the corresponding bootstrap regression matrix $X^*$ into (19). Denote this bootstrap estimate by $\hat{\beta}^* = [\hat{\beta}_1^*, \cdots, \hat{\beta}_M^*]^T$. This procedure will be repeated $B$ times. Let $\hat{\beta}^*(b)$ be the bootstrap estimate at step $b$ for $b$=1,2, …, $B$. Define

$$\overline{\hat{\beta}}^* = [\overline{\hat{\beta}}_1^*, \cdots, \overline{\hat{\beta}}_M^*]^T \tag{22}$$

where $\overline{\hat{\beta}}_m^* = (1/B)\sum_{b=1}^{B} \hat{\beta}_m^*(b)$. The bootstrap estimate (22) will be used to replace the original estimate $\hat{\beta}$.

The bootstrap response defined by Eq. (21) is based on one-step-ahead predictions. This response can also be defined in terms of the model predicted outputs. Assume that the model structure is correctly identified and can be expressed as

$$\hat{y}(t) = g(y(t-1), \cdots, y(t-p), u(t), \cdots, u(t-q)) \tag{23}$$

Note that (23) can be expressed using a linear-in-the-parameters form similar to (3). The least-squares estimate of the unknown parameter $\beta$ can then be calculated using (19). The *model predicted bootstrap response* (*output*) of the system for $t$=1,2, …, $l$, with $l$=max($p$,$q$), is defined to be $y^*(t) = y(t)$ and for $t > l$,

$$y^*(t) = g(y^*(t-1), \cdots, y^*(t-p), u(t), \cdots, u(t-q)) + \mu\varepsilon^*(t) \tag{24}$$

Note that the bootstrap response defined by (24) requires that the model produce ideal long term predictions since unlike Eq. (23), Eq. (24) involves measured values $y^*$ on the right hand side. In many cases, however, an identified model for a dynamical system might fail to produce ideal long term predictions. Therefore the bootstrap observations may become of no use and the bootstrap would lose its function. In this case, the one-step-ahead bootstrap response defined by (21) would be considered and the corresponding bootstrap estimate $\hat{\beta}^*$ defined by (22) will be used to replace the original estimate $\hat{\beta}$ defined by (19) to obtain an improved estimate. Furthermore, experience on simulation studies shows that for time series modelling, where the observational data are often generated from autonomous systems (or models), improved model estimation is often achieved when the one-step-ahead prediction $\hat{y}(t)$ is replaced by the observational value $y(t)$. The advantages of the bootstrap estimate $\hat{\beta}^*$ introduced above will be demonstrated in Section 5 via several examples.

### 4.2 A procedure for bootstrap estimation

Given an observed data set in the presence of measurement errors, the first step is to choose proper basis functions to represent the observed data. Once the basis functions and associated modelling approach have been adopted, a procedure for parameter estimation using the bootstrap method can be summarized below.

*Step*1. *Select the key variables.* It would be advantageous to describe a system using a model that involves a minimum number of significant variables (lagged output and/or input of the system).

*Step*2. *Detect the significant model terms.* To obtain a parsimonious model, all redundant model terms or most of the insignificant model terms should be removed from the candidate set $\Gamma_0 = \{\Pi_1(t), \cdots, \Pi_{M_0}(t)\}$ so that only the significant model terms are kept in the final model. The aim of model term selection is to find a subset $\Gamma_M = \{\Pi_{i_1}(t), \cdots, \Pi_{i_M}(t)\} \subseteq \Gamma_0$, where $M << M_0$ and $1 \le i_k \le M_0$, so that $\mathbf{y}$ can be correctly approximated using a linear combination of the elements in $\Gamma_M$.

*Step*3. *Estimate the unknown parameters using a bootstrap procedure on the basis of Step 2.* The bootstrap resampling procedure will be repeated $B$ times.

*Step*4. *Assess the performance of the identified model.* A commonly used approach to check the validity of the identified model is to use higher order statistical correlation analysis (Billings and Voon 1986, Billings and Zhu 1994). An alternative is to check both the short and the long term predictive ability of the model.

## 5. Numerical examples

This section provides four numerical examples to illustrate the effectiveness of the new parameter estimation scheme using the bootstrap method. The systems given in these examples were assumed to be corrupted by noise and the observational data were assumed to be contaminated with noise (or measured with errors). These data sets were used for model identification and parameter estimation but the true model structure for each example was assumed completely unknown once the models had been simulated and the data were obtained. The parameter $\mu$ in Eq. (20) was set to $\mu = 1/b$ at the $b$th step bootstrap, where $b$=1,2, …, $B$.

The performance of the identified model in Examples 3 and 4 will be measured by introducing the normalized root-mean-squared-errors (NRMSE) defined as below:

$$\text{NRMSE} = \sqrt{\frac{\sum_{t=1}^{N_{test}} [y(t) - \hat{y}(t)]^2}{\sum_{t=1}^{N_{test}} [y(t) - \Delta]^2}} \tag{25}$$

where $N_{test}$ is the length of a test data set, $y(t)$ and $\hat{y}(t)$ are the measurements and associated predictions, respectively, over the test data set, $\Delta$ can be set to zero or $\bar{y} = (1/N_{test}) \sum_{t=1}^{N_{test}} y(t)$.

### 5.1 Example 1—an AR model

Consider a fourth order AR model

$$y(t) + a_1 y(t-1) + a_2 y(t-2) + a_3 y(t-3) + a_4 y(t-4) = e(t) \tag{26}$$

where $a_1$ =-1.8, $a_2$ =1.99, $a_3$ =-1.422 and $a_4$ =0.493, $e(t)$ was Gaussian white noise with mean zero and variance $\sigma_e^2$ =1. The poles of the AR process are at $p_{1,2} = 0.2 \pm j0.9$ and $p_{1,2} = 0.7 \pm j0.3$. Starting from an initial value of zero and driven by a Gaussian white noise sequence $e(t)$, the AR process (26) was simulated and 500 data points were recorded. The SNR for this data set was approximately 8.84dB.

The initial starting model order for the process was deliberately set to $n$=10. The model term selection procedure started from a candidate polynomial model with a nonlinear degree $\ell$ =2, which contained 66 candidate model terms with the form $z_1^{i_1}(t)z_2^{i_2}(t)$, where $z_j^{i_j}(t) \in \{y(t-1), \cdots, y(t-10)\}$, $j$=1,2, $0 \le i_j \le 2$ and $0 \le i_1 + i_2 \le 2$. This over-specification of the candidate model terms was deliberate to produce an identification scenario which is more realistic to the situation that would arise when identifying from a real data set. An OLS-ERR algorithm (Korenberg *et al.* 1988, Billings *et al.* 1989) was used for model structure detection. The ordinary least-squares algorithm and the proposed bootstrap method (with $B$=100) were then applied to estimate the unknown parameters based on the structure detected model. The resulting model terms and the corresponding parameters are listed in Table 1, which clearly shows that the algorithm has correctly removed all the insignificant model terms including all the nonlinear terms, and the ordinary least-squares (LS) estimator has been improved by introducing the bootstrap method.

Table 1 The parameter estimates for the system (26) in Example 1, using least-squares algorithm and the bootstrap method.

| Model Terms | | y(t-1) | y(t-2) | y(t-3) | y(t-4) |
|---|---|---|---|---|---|
| Parameters | True Value | -1.8 | 1.99 | -1.42 | 0.493 |
| | LS | -1.8378 | 2.0626 | -1.5193 | 0.5339 |
| | Bootstrap | -1.8076 | 2.0133 | -1.4723 | 0.5074 |

### 5.2 Example 2—a nonlinear model corrupted by white noise

A nonlinear system is described by the following model

$$x(t) = -0.605x(t-1) - 0.163x^2(t-2) + 0.588u(t-1) - 0.240u(t-2) + \xi(t) \tag{27a}$$

$$y(t) = x(t) + \eta(t) \tag{27b}$$

where $\xi(t)$ and $\eta(t)$ were independently distributed Gaussian white noise sequences with mean and standard deviation $\sigma_\xi^2 = \sigma_\eta^2$=0.1. By setting the input $u(t)$ as a random sequence that is uniformly distributed in [-1,1], the model (26) was simulated and 800 input-output data points were collected. An additive noise, $\zeta(t)$, a Gaussian white noise sequence with mean and standard deviation $\sigma_\zeta^2$=0.1, was also added to the sampled input data. The objective was then to identify both the model structure and the unknown model parameters based on the recorded data set, where the system input, output and state variables were corrupted by noise, and the SNR to the input, output and state variables were all approximately 15dB.

The noise corrupted data set was used to identify a NARX model with an assumption that the true model structure was unknown. Using a variable selection algorithm proposed in Wei *et al.* (2004), 4 significant variables were chosen as $y(t-1)$, $y(t-2)$, $u(t-1)$ and $u(t-2)$, which were used to form an initial full model

containing 35 candidate terms with the form $z_1^{i_1}(t)z_2^{i_2}(t)z_3^{i_3}(t)$ , where $z_j^{i_j}(t) \in \{y(t-1), y(t-2), u(t-1), u(t-2)\}$ for $j$=1,2,3, $0 \le i_j \le 3$ and $0 \le i_1 + i_2 + i_3 \le 3$. An OLS-ERR algorithm (Korenberg *et al.* 1988, Billings *et al.* 1989) was used for model structure detection. The ordinary least-squares algorithm and the proposed bootstrap method (with $B$=200) were then applied to estimate the unknown parameters based on the structure detected model. The resulting model terms and the corresponding parameters are listed in Table 2, which clearly shows that the bootstrap method provided an improved result compared to the ordinary least-squares (LS) estimator.

Table 2 The parameter estimates for the system (27) in Example 1, using least-squares algorithm and the bootstrap method.

| Model Terms | | y(t-1) | $y^2$(t-2) | u(t-1) | u(t-2) |
|---|---|---|---|---|---|
| Parameters | True Value | -0.6050 | -0.1630 | 0.5880 | -0.2400 |
| | LS | -0.5957 | -0.1559 | 0.5766 | -0.2523 |
| | Bootstrap | -0.6013 | -0.1609 | 0.5819 | -0.2468 |

### 5.3 Example 3—a nonlinear system characterized by a rational model

In Examples 1 and 2, the deterministic part of the models representing the systems was of a polynomial form, for which it is possible to compare the parameter estimates with the corresponding true values to assess the performance of the estimation algorithm. For a system that is accurately characterized by a non-polynomial model and that can be approximated by a polynomial model, the comparison between the parameter estimates and the true values may not be available. In this case, the performance of the identified model can be assessed by considering the predictive capability of the model.

Consider a nonlinear system described by the following rational model

$$x(t) = \frac{0.8x(t-1) + 0.4x(t-2) + 0.5u^2(t-1) + 0.5x(t-1)x(t-2)u(t-1) + 0.2x(t-1)x(t-2)u(t-2)}{1 + x^2(t-1) + x^2(t-2)} + \xi(t) \quad (28a)$$

$$y(t) = x(t) + \eta(t) \quad (28b)$$

where

$$\xi(t) = v(t) + 0.3v(t-1) + 0.6v(t-2) \quad (29)$$

$$\eta(t) = v(t) + 0.4v(t-1) + 0.8v(t-2) \quad (30)$$

and $v(t)$ was a Gaussian white noise with zero mean and standard deviation $\sigma = 0.02$ . This model was simulated by setting the input $u(t)$ to be a random sequence that is uniformly distributed in [-1,1], and 1000 input-output data points were collected. An additive noise, $\zeta(t)$, which was produced by the following filter model

$$\zeta(t) = 0.5w(t) + 0.25w(t-1) \quad (31)$$

was added to the sampling input data, where $w(t)$ is a random sequence uniformly distributed in [-0.2,0.2]. For a noise corrupted data set used for identification of this system, the SNR for the system input, output and state variables was approximately 14dB, 14dB and 19dB, respectively.

The significant variables were chosen as $y(t-1)$, $y(t-2)$, $u(t-1)$ and $u(t-2)$ using a variable selection algorithm (Wei *et al.* 2004), and these were used to form an initial model of degree 3, which contained 35 candidate terms of the form $z_1^{i_1}(t)z_2^{i_2}(t)z_3^{i_3}(t)$, where $z_j^{i_j}(t) \in \{y(t-1), y(t-2), u(t-1), u(t-2)\}$ for $j$=1,2,3, $0 \le i_j \le 3$ and $0 \le i_1 + i_2 + i_3 \le 3$. An OLS-ERR algorithm (Korenberg *et al.* 1988, Billings *et al.* 1989) was used for model structure detection, and the detected model terms and the corresponding parameters are listed in Table 3.

To inspect and compare the performance of the identified models listed in Table 3, an input signal $u(t) = 0.75\sin(15\times 2\pi t/300) + 0.25\sin(20\times 2\pi t/500)$ was used to drive the model (28), where the noise $\xi(t)$ and $\eta(t)$ were set to be zero. The model predicted outputs produced by the identified models given in Table 3 were compared with the noise free output data from the original model (28) over the range $3 \le t \le 1000$. Part of the result is shown in figure 1, which clearly indicates that the bootstrap identified model provides a better model predicted output compared with the ordinary LS identified model. By setting $\Delta = 0$ in (25), the NRMSEs for the LS and bootstrap identified models were 0.0123 and 0.0103, respectively, over the test data set (points 3 to 1000). If $\Delta$ was set to $\overline{y} = (1/998)\sum_{t=3}^{1000} y(t)$, the associated NRMSEs were 0.0506 and 0.0424, respectively.

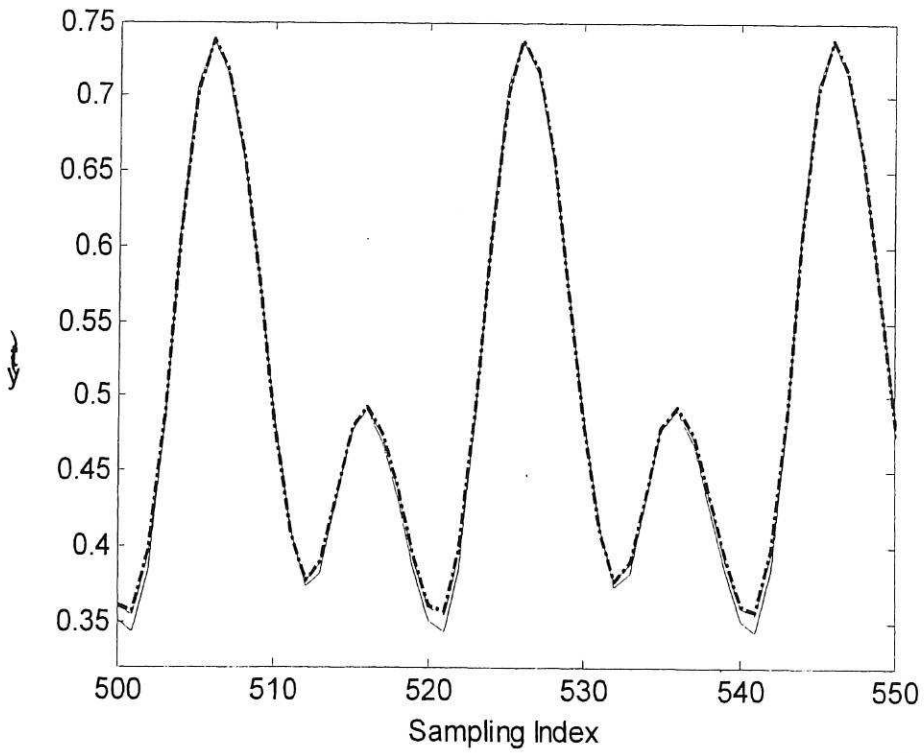



Figure 1 A comparison of the output from the LS and bootstrap identified models listed in Table 3, and from the original model (28). The solid line indicates the output from the original model (28); the bold dotted line indicates the model predicted output from the LS identified model; and the dashed line indicates the model predicted output from the bootstrap identified model.

Table 3 The identified model for the system (28) in Example 3, using least-squares algorithm and the bootstrap method.

| Index | Model Terms | Parameter Estimates | |
|---|---|---|---|
| | | LS | Bootstrap |
| 1 | y(t-1) | 0.5494 | 0.5461 |
| 2 | y(t-2) | 1.0121 | 0.9919 |
| 3 | $y^2$(t-1) | 0.1555 | 0.1980 |
| 4 | y(t-1)y(t-2) | -0.8116 | -0.8539 |
| 5 | y(t-1)u(t-1) | 0.0989 | 0.0885 |
| 6 | $y^2$(t-2) | -1.2284 | -1.1490 |
| 7 | y(t-2)u(t-2) | 0.0074 | 0.0105 |
| 8 | $u^2$(t-1) | 0.4896 | 0.4947 |
| 9 | $y^3$(t-1) | -0.3348 | -0.3731 |
| 10 | $y^2$(t-1)u(t-1) | -0.0022 | 0.0034 |
| 11 | $y^2$(t-1)u(t-2) | 0.1618 | 0.1707 |
| 12 | y(t-1) $y^2$(t-2) | 0.6458 | 0.6763 |
| 13 | y(t-1) $u^2$(t-1) | -0.1328 | -0.1437 |
| 14 | $y^3$(t-2) | 0.3946 | 0.3309 |
| 15 | $y^2$(t-2)u(t-1) | 0.1046 | 0.1145 |
| 16 | y(t-2) $u^2$(t-1) | -0.1895 | -0.1902 |
| 17 | u(t-1) $u^2$(t-2) | 0.0034 | 0.0070 |
| 18 | $u^3$(t-2) | -0.0409 | -0.0465 |

### 5.4 Example 4—Paper Making Data

Whilst the results obtained from the simulated data in Examples 1 to 3 demonstrate the effectiveness of the proposed bootstrap method, it does not provide a test on a real data set for the new algorithm. To test the performance of the bootstrap estimation algorithm, a paper making data set was considered (this data set was taken from the online time series data library http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/ ).

The paper making data set consisted of 160 input-output sampling points, which were divided into two parts: the first 120 points (see figure 2) were used for model estimation and the remaining 40 points were used for model validation. The significant variables were chosen as $y(t\text{-}1)$, $u(t)$ and $u(t\text{-}1)$ using a variable selection algorithm (Wei *et al.* 2004),and these were used to form an initial polynomial model of degree 3, which contained 21 candidate terms of the form $z_1^{i_1}(t)z_2^{i_2}(t)$ , where $z_j^{i_j}(t) \in \{y(t-1), u(t), u(t-1)\}$ for $j$=1,2, $0 \le i_j \le 2$ and $0 \le i_1 + i_2 \le 2$ . An OLS-ERR algorithm (Korenberg *et al.* 1988, Billings *et al.* 1989) was used for model structure detection. The LS and bootstrap identified models were respectively given by

$$y(t) = 0.0350y^2(t-1) + 0.3955u^2(t) + 0.4364u^2(t-1) - 0.8319u(t)u(t-1) + e(t) \quad (32)$$

$$y(t) = 0.0133y^2(t-1) + 0.4055u^2(t) + 0.4465u^2(t-1) - 0.8250u(t)u(t-1) + e(t) \quad (33)$$

To inspect and compare the performance of the identified models (32) and (33), the model predicted outputs, defined as $\hat{y}_{mpo}(t) = \hat{f}(\hat{y}_{mpo}(t-1), u(t), u(t-1))$ , were calculated from the two identified models, and these are shown in figure 3, where although the difference between the two model predicted outputs is almost distinguishable from the plots, the calculation of NRMSEs shows that the bootstrap identified model (33) is better than the LS identified model (32). Setting $\Delta = 0$ in (25), the NRMSEs for LS and bootstrap identified

models (32) and (33) are 0.3026 and 0.2997, respectively, over the test data set (points from 121 to 150). When $\Delta$ was set to $\overline{y} = (1/40)\sum_{t=121}^{160} y(t)$, the associated NRMSEs are 0.3038 and 0.3009, respectively. The correlation validity tests for model (33) which are shown in figure 4, where $E\{e(t)e(t+s)\}$, $E\{u(t)e(t+s)\}$, $E\{[e(t)u(t)]e(t+s)\}$, $E\{uu(t)e(t+s)\} = E\{\overline{u}^2(t)e(t+s)\}$ and $E\{uu(t)ee(t+s)\} = E\{\overline{u}^2(t)e^2(t+s)\}$ are defined as in Billings and Voon (1986), are all satisfied for the bootstrap model.

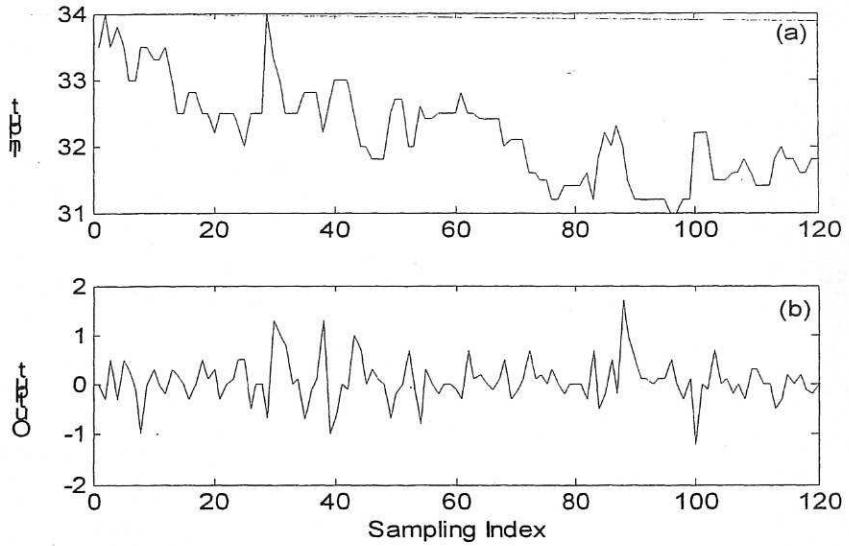



Figure 2. The paper making data. (a) Input; (b) Output.

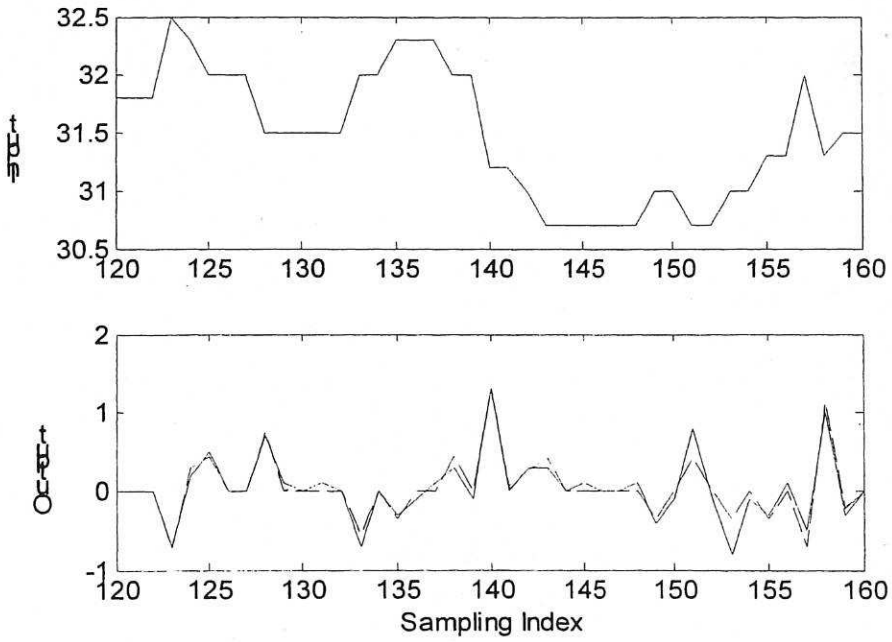



Figure 3. A comparison of the model predicted output for the paper making data from the LS and bootstrap identified models (32) and (33). The solid line indicates the real data (measurement); dotted line indicates the model predicted output from the LS identified model (32); the dashed line indicates the model predicted output from the bootstrap identified model (33).

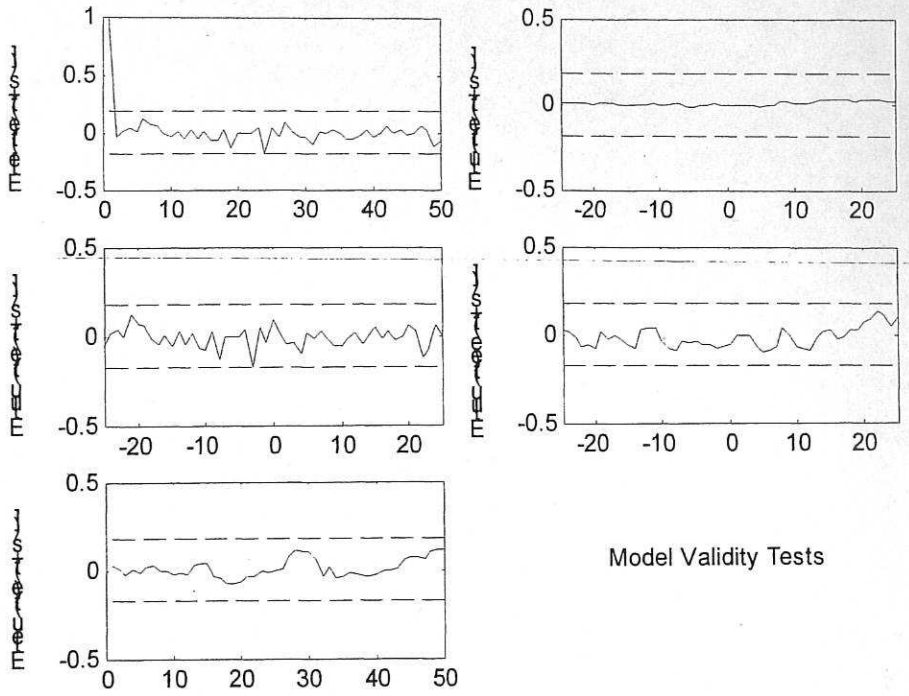



Figure 4. The model validity tests for the bootstrap identified model (33).

## 6. Conclusions

This paper has focused on improving model estimation for error-in-variables problems, where there exist measurement uncertainties, or where the systems are corrupted by noise. It has been shown that ordinary least-squares estimators for error-in-variables problems may result in biased estimates and the resulting model might lack good generalization properties. To overcome this drawback, a bootstrap algorithm has been proposed to decrease the variance of the parameter estimates associated with NARX modelling. The results obtained from the examples show that the new bootstrap algorithm works well on error-in-variables identification problems and produces improved results compared with the ordinary least-squares algorithm. The new bootstrap method can therefore be applied to improve a NARX model identified with a conventional routine, which may be biased due to neglecting the effects of necessary moving averaging terms (noise terms). It should be emphasized, however, that a prerequisite that the bootstrap method is an improvement over the ordinary least-squares algorithms is that the model structure is correctly identified. Otherwise the bootstrap algorithm may not work well. It is therefore suggested that the proposed bootstrap estimation algorithm should always be applied in conjunction with some reliable model structure detection procedure as in all the examples above.

## Acknowledgment

The authors gratefully acknowledge that this work was supported by EPSRC (UK), and the use of the data in the paper making example from http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/.

## References


T. J. Abatzoglou, J. M. Mendel, and G. A. Harada, "The constrained total least squares technique and its applications to harmonic super-regression," *IEEE Trans. Signal Processing*, 1991, 39(5), pp. 1070-1087.

A. E. Beaton, D. B. Rubin, and J. L. Barone, "The acceptability of regression solutions: another look at computational accuracy," *J. Amer. Stat. Assoc.*, 1976, 71(353), pp. 158-168.

S. A. Billings and S. Chen, "The determination of multivariable nonlinear models for dynamic systems using neural networks," in *Neural Network Systems Techniques and Applications*, C.T. Leondes, Ed. San Diego: Academic Press, 1998, pp. 231-278.

S. A. Billings, S. Chen, and M. J. Korenberg, "Identification of MIMO non-linear systems suing a forward regression orthogonal estimator," *Int. J. Control*, 1989, 49(6), pp. 2157-2189.

S. A. Billings and W. S. F. Voon, "Least squares parameter estimation algorithm for non-linear systems," *Int. J. Syst. Sci.*, 1984, 15(6), pp. 601-615.

S. A. Billings and W. S. F. Voon, "Correlation based model validity tests for nonlinear models," *Int. J. Control*, 1986, 44(1), pp. 235-244.

S. A. Billings and Q. M. Zhu, "Nonlinear model validation using correlation tests," *Int. J. Control*, 1994, 60(6), pp. 1107-1120.

S. Chatterjee and A. Hadi, *Sensitivity Analysis in Linear Regression*. New York: John Wiley & Sons, 1988.

S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, 1989, 50(5), pp. 1873-1896.

R. B. Davies and B. Hutton, "The effects of Errors in the independent variables in linear regression," *Biometrika*, 1975, 62(2), pp. 383-391.

C. Duchesne and J. F. MacGregor, "Jackknife and bootstrap methods in the identification of dynamic models," *J. Process Control*, 2001, 11(5), pp. 553-564.

B. Efron, "Bootstrap methods: another look at the jackknife," *Ann. Statist.*, 1979a, 7(1), pp. 1-26.

B. Efron, "Computers and the theory of statistics: thinking the unthinkable," *SIAM Review*, 1979b, 21(4), pp. 460-480.

B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.

G. H. Golub and C.F. Van Loan, "An analysis of the total least-squares problems," *SIAM J. Numer. Anal.*, 1980, 17(6), pp. 883-893.

G. H. Golub, G.H. and C. F. Van Loan, *Matrix Computations*. Baltimore: John Hopkins University Press, 1989.

S. D. Hodges and P. G. Moore, Data uncertainty and least squares regression. *Applied Statistics*, 1972, 21(2), pp. 185-195.

M. Korenberg, S. A. Billings, Y. P. Liu, and P. J. McIlroy, "Orthogonal parameter estimation algorithm for non-linear stochastic systems," *Int. J. Control*, 1988, 48(1), pp. 93-210.

S. L. Kukreja, H. L. Galiana, and R. E. Kearney, "A bootstrap method for structure detection of narmax models," *Int. J. Control*, 2004, 77(2), pp. 132-143.

I. J. Leontaritis. and S. A. Billings, "Input-output parametric models for non-linear systems (part I: deterministic non-linear systems; part II: stochastic non-linear systems)," *Int. J. Control*, 1985, 41(2), pp. 303-344.

B. De Moor, "Total least squares for affinely structured matrices and the noisy realization problems," *IEEE Trans. Signal Processing*, 1994, 42(11), pp. 3104-3113.

T. Soderstrom and P. Stoica, *System Identification*. London: Prentice Hall, 1989.

H. D. Vinod and A. Ullah, *Recent Advances in Regression Methods*. New York: Marcel Dekker, 1981.

H. L. Wei, S. A. Billings, and J. Liu, "Term and variable selection for nonlinear system identification," *Int. J. Control*, 2004, 77(1), pp. 86-110.

Q. M. Zhu and S. A. Billings, "Fast orthogonal identification of nonlinear stochastic models and radial basis function neural networks," *Int. J. Control*, 1996, vol. 64, pp. 871-886.