Proceedings of the 41st IEEE
Conference on Decision and Control
Las Vegas, Nevada USA, December 2002

Non-asymptotic quality assessment of generalised FIR models

M.C. Campi[1] Su Ki Ooi[2] E. Weyer[2]

Abstract

This paper presents new results on quality assessment of models identified from a finite data sample. Suppose that we are given a finite sample of measurements coming from a plant and that we are asked to provide a model of the plant along with a certification of the model quality. The certification of the model quality is a measure of how far the identified model can be from the best model in the selected model class. At the present state of knowledge, this task is not trivial to accomplish, especially in the presence of unmodelled dynamics. In this paper we focus on least squares identification of generalised FIR models and provide new finite sample bounds for the corresponding estimation error. Our method is based on tests involving permuted data sets and bears a promise of applicability to more general settings than the one developed in this paper.

Keywords: System identification, FIR models, Non-asymptotic quality assessment, Least squares

1 Introduction

In this paper, the finite sample properties of the least squares identification method are studied. The problem we investigate can be described as follows. Suppose that a finite number of measurements have been collected from a plant and that, based on these measurements, a model has been identified using the least squares method. Then, the following questions arise naturally: what can we say about the quality of this model? and, how far is the estimated model from the best model within the selected model class?

The asymptotic properties of system identification methods have been extensively studied and are now well understood (see e.g. Ljung (1999) and Söderström and Stoica (1988)). However, the asymptotic theory can be rigorously applied only when the number of measurements tends to infinity - which is never the case in practice - and it can provide poor results when the available measurements are relatively few. Thus, in order to answer the above questions, there is a need for quality assessment methods applicable to finite measurement samples.

[1]Department of Electrical Engineering and Automation, University of Brescia, Via Branze 38, 25123 Brescia, Italy. Email: campi@bsing.ing.unibs.it

[2]CSSIP, Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia. Email: {skoo, e.weyer}@ee.mu.oz.au

In order to be credited for real applicability, finite sample results should provide tight bounds on the estimation error. In previous contributions (Campi and Weyer (2002), and Weyer and Campi (2002)), finite sample results which can be applied in a general identification setting have been developed. While these results are of conceptual interest, admittedly the corresponding bounds are not tight. The main reason is that those bounds are uniform over all possible data generating mechanisms, so that they may be conservative for the specific situation at hand.

In this paper, we concentrate on a particular - though of interest in applications - setting where the plant is identified through generalised FIR models with periodic inputs. In this context, we are able to derive model quality results that are data-driven (that is, the quality is not a-priori determined and it is evaluated based on the available data). In this way, the uniformity requirement is dropped and the bounds gain in tightness. The setting allows for the presence of unmodelled dynamics, that is, the true system is not required to belong to the model class. Moreover, many tools developed in the present paper bear a promise of applicability to more general situations.

The paper is organised as follows. In the next section we introduce the identification setting. The algorithm for assessing the model quality is introduced in Section 3, together with our main non-asymptotic result on the model quality (Theorem 3.6). The proposed method for quality assessment is illustrated on a simulation example in Section 4.

2 Identification setting

2.1 Model class and input signal

We consider generalised FIR models with predictors of the type

$$\hat{y}_{t,\theta} = \theta_1 L_1(q^{-1},\alpha)u_t + \cdots \theta_n L_n(q^{-1},\alpha)u_t \qquad (1)$$

This predictor can be written in linear regression form $\hat{y}_{t,\theta} = \phi_t^T\theta$ with $\theta = [\theta_1,\ldots,\theta_n]^T$ and $\phi_t = [L_1(q^{-1},\alpha)u_t,\ldots,L_n(q^{-1},\alpha)u_t]^T$. Here θ is the parameter vector to be estimated, $L_k(q^{-1},\alpha),\ k=1,\ldots,n$ are asymptotically stable transfer functions in the backward shift operator (i.e. $q^{-1}u_t = u_{t-1}$). The transfer functions may depend on a parameter α to be chosen by the user.

Example. A popular choice of $L_k(q^{-1},\alpha)$ is the Laguerre polynomials $L_k(q^{-1},\alpha) = \frac{1}{q-\alpha}\left(\frac{1-\alpha q}{q-\alpha}\right)^{k-1}$.

0-7803-7516-5/02/$17.00 ©2002 IEEE

Next we introduce an assumption on the input signal.

A1. The input signal is deterministic and periodic with period T, and the input has been applied since time $t = -\infty$.

To simplify notation let

$$\Phi = [\phi_1, \ldots, \phi_{NT}] = [\Phi_1, \ldots, \Phi_1]$$
$$\Phi_1 = [\phi_1, \ldots, \phi_T] = [\phi_{(i-1)T+1}, \ldots, \phi_{iT}]^T;\ i = 1, \ldots N$$

where N is the number of periods we observe. The last equality is due to the periodicity of the input. As a final assumption on the input signal we assume that:

A2. The input u_t and the transfer functions $L_k(q^{-1}, \alpha)$ are such that $\Phi\Phi^T$ is non-singular.

2.2 True system

We assume that the true system can be written as

$$y_t = h(q^{-1}, u_t) + w_t$$

which in vector form becomes

$$Y = \bar{Y} + W$$

where

$$\begin{aligned} Y &= [y_1, \ldots, y_{NT}]^T = [Y_1^T, \ldots, Y_N^T]^T \\ \bar{Y} &= [h(q^{-1}, u_1), \ldots, h(q^{-1}, u_{NT})]^T \\ &= [\bar{Y}_1^T, \ldots, \bar{Y}_N^T]^T \\ W &= [w_1, \ldots, w_{NT}]^T = [W_1^T, \ldots, W_N^T]^T \\ Y_i &= [y_{(i-1)T+1}, \ldots, y_{iT}]^T \\ \bar{Y}_i &= [h(q^{-1}, u_{(i-1)T+1}), \ldots, h(q^{-1}, u_{iT})]^T \\ W_i &= [w_{(i-1)T+1}, \ldots, w_{iT}]^T \end{aligned}$$

Here h is a time invariant, BIBO stable, causal operator. In addition we assume that

A3 If u_t is T-periodic then $h(q^{-1}, u_t)$ is T-periodic after transients have died out.

A4 $\Phi_1 W_i, i = 1, \ldots, N$ are iid and symmetrically distributed around zero.

Example. The assumption on h in **A3** is satisfied for all asymptotically stable noise free LTI systems, and we can also allow for static input and output nonlinearities. The assumption on the noise is satisfied if w_t is a sequence of symmetrically distributed zero mean random variables. Assumption **A4** allows for correlated noise as long as $\Phi_1 W_i$ is iid over blocks of data. ∎

We also remark that we do not assume that the true system is contained in the model class.

2.3 LS estimate

The LS estimate is given by

$$\hat{\theta}_{NT} = (\Phi\Phi^T)^{-1}(\Phi Y) \tag{2}$$

As the number of data points tends to infinity, $\hat{\theta}_{NT}$ converges to

$$\theta^* = (\Phi\Phi^T)^{-1}(\Phi E Y) = (\Phi\Phi^T)^{-1}(\Phi\bar{Y})$$

where E is the expectation operator.

Lemma 2.1 *Y can be written as $Y = \Phi^T\theta^* + \bar{W}$ where $\bar{W}$ has the property that $\Phi\bar{W} = \Phi W$.*

Proof: $\Phi\bar{W} = \Phi(Y - \Phi^T\theta^*) = \Phi(\bar{Y} + W - \Phi^T\theta^*) = \Phi W$. ∎

3 Algorithm for model quality evaluation

Our aim is to derive a bound on $(\hat{\theta}_{NT} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{NT} - \theta^*)$. To this end let the Singular Value Decomposition (SVD) of Φ be given by $\Phi = USV^T = U\bar{S}\bar{V}^T$, where $S = [\bar{S} \vdots 0]$, $V^T = \begin{bmatrix} \bar{V}^T \\ \cdots \\ \tilde{V}^T \end{bmatrix}$ where $\bar{S}$ is an $n \times n$ matrix containing the nonzero singular values and $\bar{V}^T$ is an $n \times NT$ matrix containing the n first rows of V^T. Due to the periodicity of Φ, $\bar{V}^T$ can be written as $\bar{V}^T = [\bar{V}_1^T, \ldots, \bar{V}_1^T]$ where $\bar{V}_1^T$ is an $n \times T$ matrix.

Lemma 3.1

$$(\hat{\theta}_{NT} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{NT} - \theta^*) = W^T\bar{V}\bar{V}^T W \tag{3}$$

Proof: $\hat{\theta}_{NT} - \theta^* = (\Phi\Phi^T)^{-1}(\Phi Y) - \theta^* = (\Phi\Phi^T)^{-1}\Phi W$. Hence, $(\hat{\theta}_{NT} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{NT} - \theta^*) = W^T\Phi^T(\Phi\Phi^T)^{-1}\Phi W$. Now, $\Phi^T(\Phi\Phi^T)^{-1}\Phi = \bar{V}\bar{S}U^T(U\bar{S}\bar{V}^T\bar{V}\bar{S}U^T)^{-1}U\bar{S}\bar{V}^T = \bar{V}\bar{S}U^T(U\bar{S}\bar{S}U^T)^{-1}U\bar{S}\bar{V}^T = \bar{V}\bar{V}^T$ where the second equality follows from properties of the SVD and the last from assumption **A2**. ∎

Notice that Lemmas 2.1 and 3.1 hold under much more general conditions than those introduced in the preceding section. In particular we do not need to assume that the input is periodic.

The right hand side of (3) depends on W which we only have information about via Y. The idea is to find matrices H_β such that the statistical properties of $\bar{V}^T W$ are the same as those of $\bar{V}^T H_\beta Y$. As $\bar{V}^T H_\beta Y = \bar{V}^T H_\beta(\bar{V}\bar{S}U^T\theta^* + \bar{W})$, the matrix H_β should be such that $\bar{V}^T H_\beta \bar{V} = 0$ and $\bar{V}^T H_\beta \bar{W}$ has the same statistical properties as $\bar{V}^T W$. Under assumption **A4** the latter will in general only happen if H_β swaps blocks of data corresponding to one period with or without changing the sign (If $w(t)$ is iid Gaussian it is sufficient that H_β is unitary.). However, due to the periodicity of $\bar{V}^T$ the requirements are in this case satisfied if H_β changes signs of half of the blocks.

Lemma 3.2 *Let $\beta = [\bar{\beta}_1, \ldots, \bar{\beta}_N]$ be an N-vector with $N/2$ elements equal to 1 and $N/2$ elements equal to -1, and let*

$$H_\beta = \begin{bmatrix} \bar{\beta}_1 I_T & 0 & \ldots & 0 \\ 0 & \bar{\beta}_2 I_T & & 0 \\ & & \ddots & \\ 0 & 0 & & \bar{\beta}_N I_T \end{bmatrix} \quad (4)$$

where I_T is the $T \times T$ identity matrix. Then $\bar{V}^T H_\beta \bar{V} = 0$ and $\bar{V}^T W$ has the same distribution as $\bar{V}^T H_\beta Y$.

Proof: Due to the periodicity of $\bar{V}^T$ we have that $\bar{V}^T H_\beta \bar{V} = \sum_{i=1}^N \bar{\beta}_i \bar{V}_1^T \bar{V}_1 = 0$ since $\sum_{i=1}^N \bar{\beta}_i = 0$. From the assumption that $\Phi\Phi^T$ is non-singular it follows that $\bar{V}^T = (U\bar{S})^{-1}\Phi$. Hence $\bar{V}^T H_\beta Y = \bar{V}^T H_\beta(\bar{V}\bar{S}U^T\theta^* + \bar{W}) = (U\bar{S})^{-1}\Phi H_\beta \bar{W} = (U\bar{S})^{-1}\sum_{i=1}^N \bar{\beta}_i \Phi_1 \bar{W}_i$. Now, $\bar{W} = Y - \Phi^T\theta^* = W + (\bar{Y} - \Phi^T\theta^*)$. From Lemma 2.1 it follows that $\Phi(\bar{Y} - \Phi^T\theta^*) = 0$ and from assumptions **A1** and **A3** $(\bar{Y} - \Phi^T\theta^*)$ and Φ are both periodic, and we must have that $\Phi_1(\bar{Y}_i - \Phi_1^T\theta^*) = 0$. Therefore $(U\bar{S})^{-1}\sum_{i=1}^N \bar{\beta}_i\Phi_1\bar{W}_i = (U\bar{S})^{-1}\sum_{i=1}^N \bar{\beta}_i\Phi_1 W_i = (U\bar{S})^{-1}\Phi H_\beta W$. From assumption **A4**, $\Phi_1 W_i$ has the same distribution as $-\Phi_1 W_i$, $i = 1, \ldots, N$ and $\Phi_1 W_i$, $i = 1, \ldots, N$ are iid and hence $(U\bar{S})^{-1}\Phi H_\beta W$ has the same distribution as $(U\bar{S})^{-1}\Phi W = \bar{V}^T W$. ∎

In view of Lemma 3.1 and 3.2 an algorithm for estimating

$$p = Pr\{(\hat{\theta}_{NT} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{NT} - \theta^*) \geq \epsilon\}$$

is to find a number of matrices H_{β_i} which change the sign of half of the data blocks and compute the frequency of the event $||\bar{V}^T H_{\beta_i}^T Y||^2 \geq \epsilon$.

Lemma 3.3 *Let $\beta_1, \ldots, \beta_M$ be N-vectors with $N/2$ elements equal to 1 and $N/2$ elements equal to -1, and let H_{β_i}, $i = 1, \ldots, M$ be the corresponding matrices given in equation (4). Furthermore, let 1_A denote the indicator function for the set A. Then*

$$\hat{p} = \frac{1}{M}\sum_{i=1}^M 1_{\{X|\ ||\bar{V}^T H_{\beta_i} X||^2 \geq \epsilon\}}(Y) \quad (5)$$

is an unbiased estimator for p.

Proof:

$$\begin{aligned} E\hat{p} &= \frac{1}{M}\sum_{i=1}^M E 1_{\{X|\ ||\bar{V}^T H_{\beta_i} X||^2 \geq \epsilon\}}(Y) \\ &= \frac{1}{M}\sum_{i=1}^M Pr\{||\bar{V}^T H_{\beta_i} Y||^2 \geq \epsilon\} \\ &= \frac{1}{M}\sum_{i=1}^M Pr\{||\bar{V}^T W||^2 \geq \epsilon\} = p \end{aligned}$$

The second last equality follows from Lemma 3.2 and the last from Lemma 3.1. ∎

Based on Lemma 3.3, an intuitive way of assessing the model quality is to replace the true probability p with its estimate $\hat{p}$. As $\hat{p}$ is an unbiased estimator, this is a non-conservative bound on the quality. However a statement like

$$Pr\{(\hat{\theta}_{NT} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{NT} - \theta^*) \geq \epsilon\} \leq \hat{p} \quad (6)$$

is wrong at the conceptual level since $\hat{p}$ is data dependent and hence stochastic. The statement (6) needs to be qualified with another level of probability, telling us the probability that the stochastic quality claim (6) is true. What is sought, is a bound on the probability that $p \leq \hat{p} + \rho$ where ρ is a margin. By choosing vectors β_i, $i = 1, \ldots, M$ that are mutually orthogonal, such a bound can be obtained under some added conditions.

Lemma 3.4 *Let $N = 2^l$. Then there exists $N - 1$ mutually orthogonal vectors whose elements are 1 and -1 with an equal number of each.*

Proof: See Ooi et al (2002) for a constructive proof. ∎

Lemma 3.5 *Let $\beta_1, \ldots, \beta_M$ be mutually orthogonal and let $H_{\beta_1}, \ldots, H_{\beta_M}$ be the corresponding matrices given by (4). Strengthen condition* **A4** *to*

A4' *$w(t)$ is a sequence of iid zero mean Gaussian random variables.*

Then

$$Pr\{p \geq \hat{p} + \rho\} \leq e^{-2M\rho^2} \quad (7)$$

Proof: (7) follows from Hoeffding's inequality (Vidyasagar (1997)) once we have proved that $\bar{V}^T H_{\beta_i} Y$, $i = 1, \ldots, M$ are independent of each other for $i \neq j$. From the proof of Lemma 3.3 $\bar{V}^T H_{\beta_i} Y = \bar{V}^T H_{\beta_i} W$, and under the strengthened assumption they are iid if $E(\bar{V}^T H_{\beta_i} W)(W^T H_{\beta_j}^T \bar{V}) = 0$ since uncorrelated Gaussian random variables are independent. Now

$$\begin{aligned} &E(\bar{V}^T H_{\beta_i} W)(W^T H_{\beta_j} \bar{V}) \\ &= \bar{V}^T H_{\beta_i} E(WW^T) H_{\beta_j} \bar{V} \\ &= \sigma^2 \bar{V}^T H_{\beta_i} H_{\beta_j} \bar{V} \\ &= \sigma^2 [\bar{\beta}_{i1}\bar{V}_1^T, \bar{\beta}_{i2}\bar{V}_1^T, \ldots, \bar{\beta}_{iN}\bar{V}_1^T] \begin{bmatrix} \bar{\beta}_{j1}\bar{V}_1 \\ \bar{\beta}_{j2}\bar{V}_1 \\ \vdots \\ \bar{\beta}_{jN}\bar{V}_1 \end{bmatrix} \\ &= \sigma^2 \beta_i \beta_j \bar{V}_1^T \bar{V}_1 = 0 \end{aligned}$$

since $\beta_i\beta_j = 0$. ∎

Combining Lemma 3.3 and 3.5 we obtain the following theorem

Theorem 3.6 *Given a model class and a true system as in section 2.1 and 2.2. Let the number of observed periods of*

data be $N = 2^l$, and let the LS estimate be given by (2). Let $\hat{p}$ be given by (3.3) with $M = N - 1$ and $\beta_1, \ldots, \beta_M$ mutually orthogonal, and assume that **A4**$'$ *is satisfied, then the statement*

$$Pr\{(\hat{\theta}_{NT} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{NT} - \theta^*) \geq \epsilon\} \leq \hat{p} + \rho$$

holds true with probability at least $1 - e^{-2(N-1)\rho^2}$.

The above theorem involves two levels of probability. Firstly, it claims that the probability that $\hat{\theta}_{NT}$ and θ^* are more than ϵ apart is less than $\hat{p} + \rho$. This claim is itself stochastic since $\hat{p}$ is a random variable. The second level of probability tells us that the quality claim is true with probability at least $1 - e^{-2(N-1)\rho^2}$. For example, if we would like the statement to be true with probability 0.95 and $\rho = 0.05$, the required number of periods is $N = 600$.

3.1 Discussion

Computing partial estimates. Under assumption **A1** and **A3** the proposed approach for model quality evaluation is equivalent to comparing the differences between partial estimates as used in Ooi et al (2002). To see this we note that

$$\begin{aligned}
&||\bar{V}^T H_\beta Y||^2 = ||(U\bar{S})^{-1}\Phi H_\beta Y||^2 \\
&= ||(U\bar{S})^{-1}(\textstyle\sum_{\beta_i=1}\Phi_1 Y_i - \sum_{\beta_i=-1}\Phi_1 Y_i)||^2 \\
&= \left|\left|\tfrac{1}{2}(\bar{S}U^T)(\theta' - \theta'')\right|\right|^2 = \tfrac{1}{4}(\theta' - \theta'')^T(\Phi\Phi^T)(\theta' - \theta'')
\end{aligned}$$

where θ' and θ'' are the LS estimates computed using only data blocks corresponding to $\beta_i = 1$ and $\beta_i = -1$ respectively. Hence

$$\begin{aligned}
&Pr\{(\hat{\theta}_{NT} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{NT} - \theta^*) \geq \epsilon\} \\
&= Pr\{||\bar{V}^T H_\beta Y||^2 \geq \epsilon\} \\
&= Pr\{(\theta' - \theta'')^T(\Phi\Phi^T)(\theta' - \theta'') \geq 4\epsilon\}
\end{aligned}$$

The idea of using partial estimates for estimating the variance dates back a long time, see e.g. McCarthy (1969).

Subsampling. The approach is also strongly connected with subsampling methods (Politis et. al. 1999) where the variability in the estimate is assesed by comparing the estimate to estimates computed on subsets of the total data set. This can be seen by noting that $\theta' - \theta'' = 2(\hat{\theta}_{NT} - \theta'')$. Hartigan (1969) has also used subsamples to compute exact confidence intervals for a scalar parameter, and his use of balanced sets is similar to our use of orthogonal β-vectors.

Fisher distribution. If in addition to the assumptions in Theorem 3.6 it is assumed that the true system belongs to the model class, one could derive a confidence ellipsoid with a data dependent ϵ using the fact that

$$\frac{(\hat{\theta}_{NT} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{NT} - \theta^*)}{n\hat{\sigma}^2_{NT}} \tag{8}$$

has a $F(n, NT - n)$ distribution (Ljung(1999), Appendix II) where $\hat{\sigma}^2_{NT} = \frac{1}{NT-n}\sum_{t=1}^{NT}(y_t - \phi_t^T\hat{\theta}_{NT})^2$. This would remove the need for a second level of probability. However, if the true system does not belong to the model class (8) does not have a Fisher distribution.

Relationship to bootstrap. The technique of swapping blocks of output data and changing sign on half of them bears some resemblances with the resampling technique used in bootstrap. However, in our approach the swapping and sign changes are done in a systematic and deterministic fashion, and unlike bootstrap there is no random sampling from an empirical distribution. One way to view our proposed method is that we have replaced the original problem of estimating $Pr\{(\hat{\theta}_{NT} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{NT} - \theta^*) \geq \epsilon\}$ with the problem of estimating $Pr\{||\bar{V}^T H_\beta Y||^2 \geq \epsilon\}$ where the latter problem is "easier" since we have $N - 1$ iid realisations of $\bar{V}^T H_\beta Y$ at hand.

The proposed approach has advantages over basic implementations of the bootstrap method, in the sense that we do not have to model the data generating mechanism in detail as the following example shows.

Suppose the assumption in Theorem 3.6 are satisfied and that the signals have period 4, with $U_1 = [2, 0, -1, 0]$ and $\bar{Y}_1 = [4, 6, -2, 6]$. As a model we use $y_t = \theta u_t$. $\theta^* = 2$ and the prediction errors over one period are $[0, 6, 0, 6]$ plus white Gaussian noise. Assuming the estimate $\hat{\theta}_{NT} \approx \theta^*$ the empirical prediction errors are $z_t = a_t + w_t$ where a_i is around 6 half of the times and around zero the other half. A new data set is generated using the same periodic input signal and the signal model $y_t = \hat{\theta}_{NT}u_t + z_t$ for the output, where z_t is drawn randomly from the empirical distribution of the prediction errors. This leads to output values $Y_B \approx [7 \pm 3, 3 \pm 3, 1 \pm 3, 3 \pm 3]$ over one period. Average values of an estimate of θ using the resampled data is approximately $\hat{\theta}_B \approx 2.6$, and evaluating the variation in $\hat{\theta}_{NT}$ by computing $\hat{\theta}_B - \hat{\theta}_{NT}$ would lead to an erroneous result. One may argue that this is not the fault of bootstrap, but due to our naive implementation. However, in order to use bootstrap correctly, we would have to go through more complicated computations, and these computations would in general require more detailed knowledge of the true system.

Similarities with other methods. In addition to the connection with subsampling the proposed approach has similarities with permutation tests in statistical hypothesis testing (Lehmann (1986)) and the use of Rademacher processes in learning theory (Koltchinskii (2001)). Moreover, the selection of matrices H_β such that $\bar{V}^T H_{\beta_i} Y$ are iid is similar to the main idea in multitaper power spectrum estimation (Manolakis et al (2000)).

4 Simulations

Simulations are performed in order to illustrate the obtained results. A deterministic periodic input signal with period $T = 4$ and values $[u_1, u_2, u_3, u_4] =$

$[2.330, 1.254, 0.150, 0.703]^T$ is generated. The number of periods is $N = 512$ and $NT = 2048$ data points are generated using the following third order FIR system with static output nonlinearity:

$$\bar{y}_t = b_1 u_{t-1} + b_2 u_{t-2} + b_3 u_{t-3}$$
$$y_t = \bar{y}_t^2 + w_t$$

The parameter is $\theta_0 = [b_1, b_2, b_3] = [0.9, 0.3, 0.6]^T$ and $w(t)$ is white noise ($\sim N(0, \sigma^2)$) with $\sigma^2 = 0.09$. The parameter $\hat{\theta}_{NT}$ is estimated using the LS method. We know that $(\hat{\theta}_N - \theta^*)$ is Normally distributed with 0 mean and covariance matrix $\sigma^2(\Phi\Phi^T)^{-1}$, hence $\frac{1}{\sigma^2}(\hat{\theta}_{NT} - \theta^*)^T\Phi\Phi^T(\hat{\theta}_{NT} - \theta^*) \in \chi^2(n)$, where n is the model order (see e.g. Ljung(1999)). The result in Ljung(1999) is derived under the assumption that the model is in the model class, but the result also holds in the setting considered here. Therefore, $p = Pr\{(\hat{\theta}_{NT} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{NT} - \theta^*) \geq \varepsilon\}$ can be computed for fixed ε and σ using the Chi-Square distribution table (see e.g. Richmond(1964)) as the level of significance corresponding to $\chi^2(3) = \frac{\varepsilon}{\sigma^2}$.

ε is fixed to be 0.70335 and an estimate, $\hat{p}$ (see equation (5)) of p is obtained from the simulated data. As we only consider β-sequences which are mutually orthogonal, we have $M = N - 1$ swapped data sets. In order to obtain several realisations, the simulation process is repeated 199 times. First we compare $\hat{p}$ with p. Then, a value for ρ is chosen, and the probability $1 - e^{(-2(N-1)\rho^2)}$ is compared to the number of times $p \leq \hat{p} + \rho$ holds.

4.1 Results

In order to observe the estimation error in each component of the parameter vector, we decompose Φ into $U\bar{S}V^T$ using singular value decomposition. Hence, $\Phi\Phi^T = U\bar{S}\bar{S}U$, and $(\theta' - \theta'')^T(\Phi\Phi^T)(\theta' - \theta'') = (\theta' - \theta'')^T(U\bar{S}\bar{S}U^T)(\theta' - \theta'')$. We plotted the scatter plot of $\bar{S}U^T(\theta' - \theta'')$ for one particular realisation together with the value of $\bar{S}U^T(\hat{\theta}_N - \theta_0)$; where $\hat{\theta}_{NT} = [2.0289, 0.9203, 0.7968]^T$ and $\theta^* = [2.0272, 0.9352, 0.7910]^T$, denoted by a 'cross', see Figure 1. For easier readability, the three dimension plot can be projected onto the (x, y), (x, z) and (y, z)-plane as shown in Figure 2 ((x, y)-plane only). From these scatter

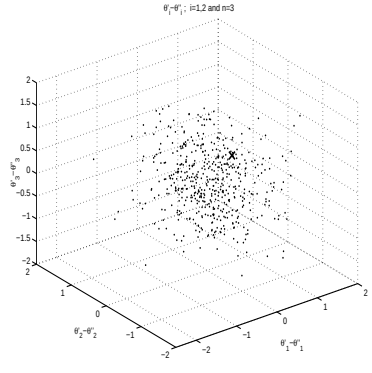

Figure 1: Plot of $\bar{S}U^T(\theta' - \theta'')$ with $N = 512$ and $T = 4$.

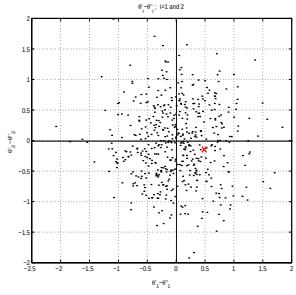

Figure 2: Projection of $\bar{S}U^T(\theta' - \theta'')$ onto (x, y)-plane.

plots, based on one simulation, it is observed that the partial estimates form a region around zero, and $\bar{S}U^T(\hat{\theta}_{NT} - \theta^*)$ is within this region. Out of the 511 realisations of $(\theta' - \theta'')^T(\Phi\Phi^T)(\theta' - \theta'')$, 88 are larger than $4(\hat{\theta}_{NT} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{NT} - \theta^*)$ and 423 are smaller. This show that the idea of assessing the quality of the estimate, $\hat{\theta}_{NT}$ using swapped data sets is feasible in this setting. The weighting $U\bar{S}\bar{S}U^T$ is comparatively large, where

$$U = \begin{bmatrix} -0.542 & 0.707 & 0.454 \\ -0.642 & 0 & -0.767 \\ -0.542 & -0.707 & 0.454 \end{bmatrix}$$

$$\bar{S} = diag[89.72, 50.87, 30.15] \quad (9)$$

Hence, with $\varepsilon = 0.70335$ the actual estimation error in each component of the parameter vector is small. In the situation that the estimation error is due to the first parameter only, the actual estimation error in the first parameter is 0.014 when $\varepsilon = 0.70335$. Due to the periodicity assumption, $\Phi\Phi^T$ have the same diagonal elements. Hence, if the estimation error is due only to either the second or third parameter, the actual estimation error in that parameter is also equal to 0.014. $\bar{S}\bar{S}$ increases linearly with N, and when the value of ε is chosen, $\bar{S}$ should be taken into account. For example, if the diagonal elements of the $\bar{S}\bar{S}$ matrix are of the order of magnitude of 1000, in order to have the actual estimation error in each of the parameter around 0.01, ε should be of the order of magnitude of $0.01 \times 1000 \times 0.01 = 0.1$.

The $\hat{p}$ values obtained from the 200 realisations are plotted in the histogram together with the value of $p = 0.05$ in Figure 3. From Figure 3, we can see that $\hat{p}$ is a good estimate of p, since the histogram is concentrated around the middle bar and the probability computed using the χ^2 distribution, $p = 0.05$ falls right in the middle of the histogram. Note that we can only compute p because this is a simulation and we know σ^2. p cannot be computed using the χ^2 distribution without the knowledge of σ^2. However, in order to obtain a quality bound on the estimate using our method, no knowledge of σ^2 is required. When $\varepsilon = 0.56259$ and from the χ^2 distribution, $p = 0.1$, a similar result is obtained (histogram not shown), i.e. $\hat{p}$ is a good estimate of p.

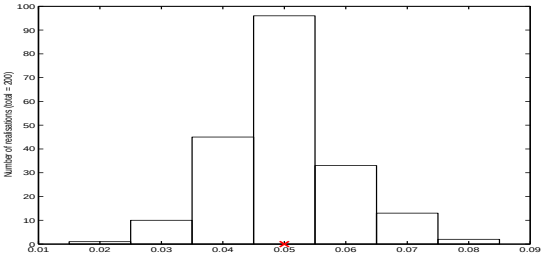

Figure 3: Histogram of $\hat{p}$ ($N = 512$, $T = 4$, $\varepsilon = 0.70335$).

Let $\mu_h = 1 - e^{(-2(N-1)\rho^2)}$. μ_h is computed for $N = 512$, $T = 4$ and different values of ρ and ε (see Table 1), and compared with the frequency that $p \leq \hat{p} + \rho$ denoted by μ_s. From Table 1, when $\rho = 0.03$ or 0.04, we do not

Table 1: μ_h and μ_s for different values of ρ and ε

N	T	ε	ρ	μ_h	μ_s
512	4	0.56259	0.03	0.601	0.985
			0.04	0.805	1
			0.05	0.922	1
512	4	0.70335	0.04	0.805	1
			0.05	0.922	1

place very high confidence in the obtained statement about the quality of the estimate (μ_h is low from a statistical point of view) even though it holds true for most or all of the simulations. However, this bound is much tighter with $\rho = 0.05$. In fact, for $\varepsilon = 0.70335$, $p = 0.05$ and the claim about the quality of the estimate will hold true all the time when $\rho \geq 0.05$. For a fixed ρ, μ_h increases with N. Since we want to have high confidence in the claim about the quality of the estimate, we would like to have μ_h close to 1. This can be done by increasing the number of data points, but in practise, the number of data points is often fixed. However, μ_h can also be increased by increasing ρ. The price to pay is a more conservative bound: $Pr\{(\hat{\theta}_{NT} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{NT} - \theta^*) \geq \varepsilon\} \leq \hat{p} + \rho$. We see that there is a natural trade off between the additional margin ρ in the bound on $Pr\{(\hat{\theta}_{NT} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{NT} - \theta^*) \geq \varepsilon\}$ and the confidence in the claim about the quality of the estimate. Note that $\hat{p}$ is a non-conservative estimate of p, and any conservativeness is introduced when we express the confidence in our quality claim.

5 Conclusion

In this paper we have presented new results on model quality assessment of system identification models. We have considered LS estimation of generalised FIR models with periodic inputs. Importantly, we have not assumed that the true system belongs to the model class. The probability $p = Pr\{(\hat{\theta}_{NT} - \theta^*)^T(\Phi\Phi^T)(\hat{\theta}_{NT} - \theta^*) \geq \epsilon\}$ is estimated using an unbiased estimator based on permutation of the data set. As the estimate of p is stochastic, a second probability is needed in order to assert the probability with which the stochastic quality claim is true. This second probability is obtained using Hoeffding's inequality. The bound on the quality of the parameter estimate as stated in Theorem 3.6 provides a rigorous result valid for a finite number of data points. Simulation results have shown that the proposed method works well, and moreover it bears a promise of applicability to more general settings than the one considered in this paper.

References

[1] Campi, M.C. and E. Weyer (2002). "Finite sample properties of system identification methods" *IEEE Trans on Automatic Control*, Vol. 47, no. 8 pp. 1329-1334.

[2] Hartigan, J.A. (1969). "Using Subsample Values as Typical Values" *Journal of the American Statistical Association*, Vol. 64, Iss. 328 pp. 1303-1317.

[3] Koltchinskii V. (2001). "Rademacher penalties and structural risk minimization" *IEEE Trans. on Information Theory*, Vol. 47, no. 5 pp. 1902-1914.

[4] Lehmann E.L. (1986). *Testing Statistical Hypotheses*. Second Edition. John Wiley. (Reprinted (1994) edition by Chapman and Hall.)

[5] Ljung L. (1999). *System Identification - Theory for the User*. Second edition. Prentice Hall.

[6] Manolakis D.G., V.K. Ingle and S.M. Kogon (2000). *Statistical and Adaptive Signal Processing*. McGraw Hill.

[7] McCarthy P.J. (1969). "Pseudo-Replication: Half Samples" *Review of The International Statistical Institute*, Vol. 37, pp. 239-263.

[8] Ooi S.K., M.C. Campi, and E. Weyer (2002). "Non-asymptotic quality assessment of the least squares estimate" *Proceedings of the 15th IFAC World Congress*, Barcelona, Spain.

[9] Politis D.N., J.P. Romano, and M. Wolf (1999). *Subsampling*. Springer Verlag.

[10] Richmond S.B. (1964). *Statistical Analysis*. Second edition. The Ronald Press Company.

[11] Söderström, T. and P. Stoica (1988). *System Identification*. Prentice Hall.

[12] Vidyasagar M. (1997). *A theory of Learning and Generalization*. Springer Verlag.

[13] Weyer, E. and M.C. Campi (2002). "Non-asymptotic confidence ellipsoids for the least squares estimate" *Automatica*, Vol. 38, no. 9 pp. 1529-1547.