15th Triennial World Congress, Barcelona, Spain

NON-ASYMPTOTIC QUALITY ASSESSMENT OF THE LEAST SQUARES ESTIMATE

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

[1] *CSSIP, Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia.*
Email: {skoo, e.weyer}@ee.mu.oz.au
[2] *Department of Electrical Engineering and Automation, University of Brescia, Via Branze 38, 25123 Brescia, Italy.*
Email: campi@bsing.ing.unibs.it

Abstract: In any real-life identification problems, only a finite number of data points are available. On the other hand, almost all results in stochastic identification pertain to asymptotic properties, that is they tell us what happens when the number of data points tend to infinity. In this paper, we consider the problem of assessing the quality of non-asymptotic estimates obtained using least squares identification methods. The type of results needed in order to be useful for computing the quality of non-asymptotic estimates are first discussed. It turns out that the nature of non-asymptotic results has to be different from that of asymptotic results, since in finite time certain issues show up that disappear in the limit because of stochastic convergence. Then, we develop a method for the assessment of the estimate quality based on differences between partial estimates. If the partial estimate differences are within a small region around zero then, as it is intuitive, the estimate quality is good. On the other hand, we will have low confidence in the estimate if the differences between partial estimates are spread all over the place. The method is illustrated through a very simple example able to point out its main aspects in a clear-cut way.

Keywords: System identification, model validation, least squares, finite samples properties, confidence sets.

1. INTRODUCTION

The purpose of system identification is to obtain a mathematical model for a dynamic system. In order to give the user confidence in the obtained model, a quality assurance should be delivered together with the model. If there is no quality tag attached, the user will not know how to properly use the model. Hence, quality assessment is important for correct usage of the model.

In this paper we consider the problem of assessing the quality of the estimate obtained using least squares (LS) system identification with N data points. The quality of the estimate is judged by the distance between the estimate obtained, $\hat{\theta}_N$ and a so-called 'best' estimate, θ_0. It is known that the mismatch between the true plant and the model consists of two components, bias error and variance error. The cause of the bias error is that the model class considered is not rich enough to contain the 'true' plant. In this work we only consider the variance error, which is due to that the best model within the model class considered has not been found. In other words, the variance error is due to the difference between the estimated model, represented by $\hat{\theta}_N$ and the 'best' model available in the model class, represented by θ_0.

1.1 *Requirements to a non-asymptotic model quality measure*

In order for a result to be useful for computing the quality of the estimate, it must be uniform with re-

spect to the data generating mechanism. That is, if we assume that the true system is in a given class, then the quality measure must be valid for all systems in that class. Furthermore, it must be possible to compute the quality measure based on a priori information about the true system and a finite number of observed data points, and finally it must provide a rigorous result valid for a finite number of data points.

The asymptotic convergence properties of the estimate are well understood, (see e.g. (Ljung, 1999) or (Söderström and Stoica, 1988)). Under natural conditions $\hat{\theta}_N$ converges to θ_0 w.p. 1, and $\sqrt{N}(\hat{\theta}_N - \theta_0)$ is asymptotically normally distributed with zero mean and variance, P_{θ_0} as $N \to \infty$. This result gives a quality tag to the estimates, and it is useful for gaining insight in the properties of system identification methods. However, not much can be said about the quality of the estimate with a finite number of data points. Moreover, the expression for P_{θ_0} involves expected values, and hence it is dependent on the true system, which is unknown. Hence, asymptotic results do not provide a rigorous quality measure valid for a finite number of data points.

However, there is a way around the problem caused by the system dependent quantities in the asymptotic results. These quantities can be estimated from observed data, but using these estimated quantities in the asymptotic variance results (as is commonly done) introduces two types of errors. The first type of error is that, as mentioned, $\sqrt{N}(\hat{\theta}_N - \theta_0)$ normally distributed with zero mean and variance P_{θ_0} is an asymptotic results, and it is not valid for a finite number of data points in general. In addition, replacing P_{θ_0} with its estimate, $\hat{P}_{\theta_0}$ causes an extra error. These two errors may cancel each other. However, uncritically substituting system dependent quantities with their data dependent estimates leads to conceptual errors. This can be clearly seen in the following simple example.

Suppose the true system is $y(t) = \theta_0 u(t) + w(t)$, where $u(t) = 1$ for all t, and $w(t)$ is i.i.d. Gaussian with zero mean and variance σ^2. No upper bound on σ^2 is available. The model used is $\hat{y}(t, \theta) = \theta u(t)$. The LS estimate is $\hat{\theta} = 1/N \sum_{t=1}^{N} y(t)$ which is a Gaussian random variable with mean θ_0 and variance σ^2/N. If σ^2 is known, we obtain a confidence interval (quality measure) of the type $Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\} = \delta(\sigma, N, \varepsilon)$.

However, σ is unknown, but it can estimated. If we substitute an estimate $\hat{\sigma}$, we find a bound for the quality of $\hat{\theta}_N$ of the kind, $Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\} < \delta(\hat{\sigma}, N, \varepsilon)$. However, this kind of bound is wrong on the conceptual level since the right hand side is data dependent and hence stochastic. At the same time it should be noted that a data independent claim such as $Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\} < \delta$ is not uniform with respect to the true system since $\sup_\sigma Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\} = 1$ for all N and ε. Due to the simplicity of this particular example, the Student-t distribution can be used to obtain a bound on $\hat{\theta}_N - \theta_0$, but the confidence interval, ε is then data dependent. However, the Student-t distribution can not be used if we want to say something about the quality of the estimate with fixed ε.

Since the claim $Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\} < \delta(\hat{\sigma}, N, \varepsilon)$ is stochastic it must be qualified with a second level of probability telling us the probability that the claim is true. Hence, in order to be rigorous we need a result of the type $\sup_\sigma Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\} < \delta(\hat{\sigma}, N, \varepsilon)$ holds true with probability at least $1 - \mu$. Such a result is uniform with respect to both σ and θ_0.

1.2 *Quality assessment by computing partial estimates*

In order to obtain a result of the type sought after in the previous section, we propose to assess the quality of the LS estimate by bounding the difference between the estimate, $\hat{\theta}_N$ and the 'best' parameter θ_0, using partial estimates. The estimate is given by:

$$\hat{\theta}_N = (\phi^T \phi)^{-1} (\phi^T Y) \tag{1}$$

where Y is the outputs, $[y_1, ..., y_N]^T$ and ϕ is the regressors, $[\varphi_1, ..., \varphi_N]^T$. The two partial estimates, $\theta' = (\phi_1^T \phi_1)^{-1}(\phi_1^T Y_1)$ and $\theta'' = (\phi_2^T \phi_2)^{-1}(\phi_2^T Y_2)$ are computed using the first and second half of the data set, where $Y_1 = [y_1, ..., y_{N/2}]^T$, $\phi_1 = [\varphi_1, , ..., \varphi_{N/2}]^T$, $Y_2 = [y_{(N/2)+1}, ..., y_N]^T$, and $\phi_2 = [\varphi_{(N/2)+1}, ..., \varphi_N]^T$. Furthermore, swapping of data points (y_i, φ_i) between $[Y_1, \phi_1]$ and $[Y_2, \phi_2]$ in all possible combinations will provide $\binom{N}{N/2}$ sets of swapped data, and hence $\binom{N}{N/2}$ pairs of partial estimates.

The idea is to judge the quality of the estimate by the difference between partial estimates. If all values of $\theta' - \theta''$ are within a small region around zero, we have, as is intuitive, a good estimate, $\hat{\theta}_N$, since there is little variation in the partial estimates. On the other hand, we have a low confidence in $\hat{\theta}_N$ if the values of $\theta' - \theta''$ are spread all over the place, since there is large variability in the partial estimates. Intuitively we will take this as an indication that the variability due to noise, unmodelled dynamics, etc. have not been sufficiently averaged out, hence we do not place much confidence in the estimate.

Swapping of data in order to obtain information about the system based on a finite number of data points has been used in other areas, for example permutation and randomization tests used for statistical testing or data analysis (see e.g. (Edgington, 1995); (Good, 1999); (Good, 2000)). Swapping of data points is also at the core of the proof of many uniform convergence results in learning theory, see e.g. (Vidyasagar, 1997) or (Vapnik, 1998). It is also closely related to Rademacher processes, which has been used for bounding the risk in function learning, (Koltchinskii and Panchenko, 2000).

Loosely speaking, the idea is to obtain a result that relates $Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\}$ to $Pr\{|\theta' - \theta''| \geq \varepsilon\}$ and estimate the latter from the observed data. In the general case, this is a difficult problem. As a starting point, we

consider the simple first order FIR system described above. The main result (see equation (5)) states that $Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\} \leq \sqrt{2Pr\{|\theta' - \theta''| \geq 2\varepsilon\}}$. However, we can only compute an empirical value of $Pr\{|\theta' - \theta''| \geq 2\varepsilon\}$ using observed data. This empirical value is itself a random variable since it is data dependent. We quantify the discrepancy between $Pr\{|\theta' - \theta''| \geq 2\varepsilon\}$ and its estimate using Hoeffding's inequality, which leads to a second level of probability.

The developed method assesses the quality using a fixed ε rather than a fixed probability δ. One may also like to asses the quality of the estimate using a fixed probability δ. One possible approach is; given a fixed probability level, construct the empirical cumulative distribution function of the partial estimates, and estimate ε from the empirical distribution. In order to obtain a rigorous result, the discrepancy between the true ε value and its estimate has to be quantified. In the simple example considered in Section 2 and 3, the student t-distribution will provide us with a result with a fixed probability level.

Recently, finite sample properties of system identification methods have been studied in e.g. (Weyer, 2000); (Weyer and Campi, 1999); (Weyer and Campi, 2000); (Campi and Weyer, 2002). The results obtained are apriori results, since the bounds can be computed before any data are collected. This is good in the sense that we can say something about the bound on the estimates before any experiment is performed. However, due to the lack of prior knowledge this leads to results which are worst case with respect to the prior information, and the results obtained in those papers are conservative. In this paper, the properties of the estimate are studied after the data are collected, i.e. we obtain aposteriori results. In general, the aposteriori result is expected to be better because the conservativeness due to lack of prior knowledge can be reduced.

The paper is organised as follows. In Section 2, the model structures and assumptions are outlined. Then in Section 3, the main result is presented, followed by some simulation results in Section 4. Conclusions are given in Section 5.

2. MODEL STRUCTURES AND ASSUMPTIONS

As a starting point, a simple first order FIR system is considered. Even though it is a trivial example, it is computationally involved. The system is given by:

$$y(t) = \theta_0 u(t) + w(t) \tag{2}$$

where $y(t)$ is the output, θ_0 is the 'true' parameter, $u(t) = 1$ for all t and $w(t)$ is i.i.d. Gaussian random noise with zero mean and unknown variance σ^2. No upper bound exists for σ^2. If σ^2 were known, the variance of the estimate could be computed using the normal distribution formula. The Student t-distribution could have been applied to solve the problem without any knowledge of σ^2, but with a stochastic confidence interval. Here we want to say something about $Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\}$ for fixed ε. The fixed ε and the lack of an upper bound on σ^2 are what make quality assessment non-trivial even for this simple example.

The model used for identification is:

$$\hat{y}(t, \theta) = \theta u(t) \tag{3}$$

and the LS estimate is $\hat{\theta}_N = \frac{1}{N}\sum_{t=1}^{N} y(t)$.

3. MAIN RESULT: BOUND ON $\hat{\theta}_N$

Let N (even) be the number of data points, ϵ some small number, and θ' and θ'' given by

$$\theta' = \frac{1}{N/2}\sum_{t=1}^{N/2} y(t), \qquad \theta'' = \frac{1}{N/2}\sum_{t=\frac{N}{2}+1}^{N} y(t) \tag{4}$$

Lemma 1: *Given the model structure and assumptions in Section 2, we have that:*

$$Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\} \leq \sqrt{2Pr\{|\theta' - \theta''| \geq 2\varepsilon\}} \tag{5}$$

PROOF:

$$\begin{aligned} &Pr\{|\theta' - \theta''| \geq 2\epsilon\} \\ &= Pr\{|(\theta' - \theta_0) + (\theta_0 - \theta'')| \geq 2\epsilon\} \\ &\geq \frac{1}{2}[Pr\{|\theta' - \theta_0| \geq \epsilon\}]^2 \\ &\geq \frac{1}{2}[Pr\{|\hat{\theta}_N - \theta_0| \geq \epsilon\}]^2 \end{aligned}$$

The first inequality holds because θ' and θ'' are i.i.d.. The last inequality holds because $(\theta' - \theta_0)$ is Gaussian with zero mean and $(\hat{\theta}_N - \theta_0)$ is also Gaussian with zero mean and variance smaller than the variance of $(\theta' - \theta_0)$. Note that by utilising the Gaussian assumption, we can obtain the tighter bound $Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\} = Pr\{|\theta' - \theta''| \geq 2\varepsilon\}$, but the bound in the Lemma may hold in more general cases.

3.1 *Evaluation of* $Pr\{|\theta' - \theta''| \geq 2\varepsilon\}$

From (5), the quality of the estimate can be assessed by bounding $Pr\{|\theta' - \theta''| \geq 2\varepsilon\}$. The partial estimates depend on the output $y(t)$ which is stochastic with unknown variance σ^2, hence $Pr\{|\theta' - \theta''| \geq 2\varepsilon\}$ can not be bounded. It can only be estimated using observed data. In order to do that, data points are swapped between the first and the second half of the data set. We treat the swapped data set as a new data set and use it to identify θ' and θ''. Swapping is repeated a number of times, and we obtain a number of realisations of $\theta' - \theta''$ from which we can estimate $Pr\{|\theta' - \theta''| \geq 2\varepsilon\}$ by computing the frequency of $|\theta' - \theta''| \geq 2\varepsilon$. There are all together $\binom{N}{N/2}$ swapped data sets. However, considering so many data sets are usually too computationally demanding. Therefore, we need to consider swapping schemes which require fewer computations. For example, performing swapping in a random fashion or only use swapped data sets that give mutually independent $\theta' - \theta''$.

Let $\bar{\beta} := [1...1-1...-1]^T$ ($\frac{N}{2}$ of +1 and $\frac{N}{2}$ of -1) and $y := [y_1...y_N]^T$, then

$$\theta' - \theta'' = \frac{2}{N}\sum_{t=1}^{\frac{N}{2}} y_t - \frac{2}{N}\sum_{t=\frac{N}{2}+1}^{N} y_t$$
$$= \frac{2}{N}\bar{\beta}^T y = 2V_{\bar{\beta}}$$

Let $\bar{p} := Pr\{|\theta' - \theta''| \geq 2\varepsilon\} = Pr\{|V_{\bar{\beta}}| \geq \varepsilon\}$ Let $\beta_i, i = 1, \ldots, M$, be vectors of length N whose elements are either +1 or -1, and with equal number of each. If the j^{th} element of β_i is $+1$, then the j^{th} data point in the swapped data set belongs to the first subset of data, if it is negative it belongs to the second subset. Also, let $V_{\beta_i} := \frac{1}{N}\beta_i^T y$. Note that $2V_{\beta_i}$ is $\theta' - \theta''$ obtained with a swapped data set.

Since y_t is i.i.d., $\bar{p} = Pr\{|V_{\bar{\beta}}| \geq \varepsilon\} = Pr\{|V_{\beta_i}| \geq \varepsilon\}$, for $i = 1, ..., M$. Therefore:

$$\bar{p} = \frac{1}{M}\sum_{i=1}^{M} Pr\{|V_{\beta_i}| \geq \varepsilon\}$$
$$= E\left[\frac{1}{M}\sum_{i=1}^{M} 1(|V_{\beta_i}| \geq \varepsilon)\right] = E[\xi]$$

where $\xi = \xi(y) = \frac{1}{M}\sum_{i=1}^{M} 1(|V_{\beta_i}| \geq \varepsilon)$ and $1(\cdot)$ is the indicator function. ξ is the estimate of $Pr\{|\theta' - \theta''| \geq 2\varepsilon\}$. We need to quantify the discrepancy between the true probability and its estimate, and this can be done using Hoeffding's inequality.

3.2 *Bound on* $Pr\{|\theta' - \theta''| \geq 2\varepsilon\}$

When we only use those β_i sequences that are mutually orthogonal, the number of swapped data sets is reduced from $\binom{N}{N/2}$ to only $N - 1$ (see Lemma 2), and, by Lemma 3, $1(|V_{\beta_i}| \geq \varepsilon)$, for $i = 1, 2..., N-1$, are independent of each other.

Lemma 2: *In the set of vectors of length $N = 2^l$, $l = 1, 2, 3, ...$, whose elements are $+1$ or -1, with an equal number of $+1$'s and -1's, there are $N - 1$ mutually orthogonal vectors.*

PROOF: *See Appendix A.*

Lemma 3: *Given mutually orthogonal β_i, $i = 1, 2...$, $N - 1$. Then $1(|V_{\beta i}| \geq \varepsilon)$, $i = 1, 2..., N - 1$ are independent of each other.*

With orthogonal β_i's, we can use Hoeffding's inequality to find a bound for $Pr\{|\xi(y) - \bar{p}| \geq \rho\}$. By Lemma 3 and Hoeffding's inequality (see e.g. (Vidyasagar, 1997)), we have:

$$Pr\{|\xi(y) - \bar{p}| \geq \rho\} \leq 2e^{(-2(N-1)\rho^2)} \quad (6)$$

where

$$\xi(y) = \frac{1}{N-1}\sum_{i=1}^{N-1} 1(|V_{\beta_i}| \geq \varepsilon) \quad (7)$$

is computable, and the above result does not depend on the value of θ_0 or σ^2. Hence, it is uniform with respect to σ and θ_0.

Putting Lemma 1 and equation (6) together we obtain

Theorem 1: *Given the system, model structure and assumptions as in section 2, and ξ given by equation* (7) *then the statement*

$$Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\} \leq \sqrt{2(\xi + \rho)} \quad (8)$$

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

In words, we claim that the probability of the estimate to be more than ε apart from the 'best' estimate is less than $\sqrt{2(\xi + \rho)}$. This claim is itself probabilistic since $\sqrt{2(\xi + \rho)}$ is a random variable. The second probability tells us that the claim is true with probability not less than $1 - 2e^{(-2(N-1)\rho^2)}$.

4. SIMULATIONS

Simulations are performed to check the result obtained (see equation (8)). For easier readability and result keeping let $p = Pr\{|\hat{\theta}_N - \theta| \geq \varepsilon\}$ and $\mu_h = 1 - 2e^{(-2(N-1)\rho^2)}$. N (= 128) data points are generated using the FIR system in equation (2) with a true parameter $\theta_0 = 0.9$ (since we consider a first order FIR model, the best parameter is the true parameter), $u(t) = 1\ \forall t$ and $w(t)$ is white noise ($\sim N(0, \sigma^2)$) with $\sigma^2 = 0.09$. The parameter $\hat{\theta}_N$ is estimated using the LS method. $(\hat{\theta}_N - \theta_0)$ is Normally distributed with 0 mean and variance $\frac{\sigma^2}{N}$, and $Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\}$ can be computed for fixed ε and σ.

ε is fixed to be 0.05 and an estimate of $Pr\{|\theta' - \theta''| \geq 2\varepsilon\}$, $\xi = \frac{1}{M}\sum_{i=1}^{M} 1(|V_{\beta_i}| \geq \varepsilon)$ is obtained from the simulated data. As, we only consider those sequences that are mutually orthogonal, we have $N - 1$ swapped data sets. The first test is to check the tightness of the bound $Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\} \leq \sqrt{2(\xi + \rho)}$. (Note that $p = E[\xi]$, hence the bound, $\sqrt{2(\xi + \rho)}$ for p will be conservative.) In order to obtain several realisations, the simulation process is repeated 199 times. A value for ρ is chosen, and the tightness of the probability $1 - 2e^{(-2(N-1)\rho^2)}$ is checked by the number of times $Pr\{|\hat{\theta}_N - \theta| \geq \varepsilon\} \leq \sqrt{2(\xi + \rho)}$ holds.

4.1 *Results*

The $\sqrt{2\xi} = \gamma$ values obtained from the simulations are plotted in the histogram together with the value of p (= 0.05935) denoted by a cross in Figure 1. We have also plotted the value of $\theta' - \theta''$ for one particular realisation together with the value of $\hat{\theta}_N - \theta_0$ in Figure 2.

From Figure 1, we can see that with $N = 128$, $Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\} \leq \sqrt{2\xi} = \gamma$ is a loose bound, since all γ values are much larger than p. The simulation is repeated with $N = 512$.

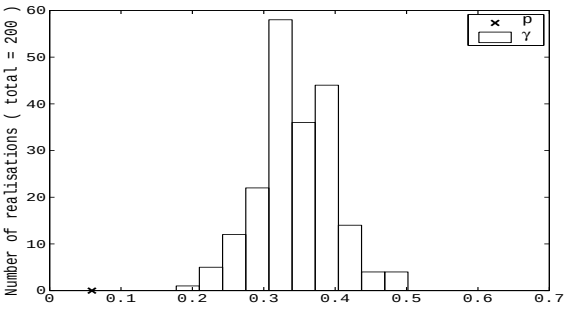

Fig. 1. Histogram of $\sqrt{2\xi} = \gamma$. ($N = 128$)

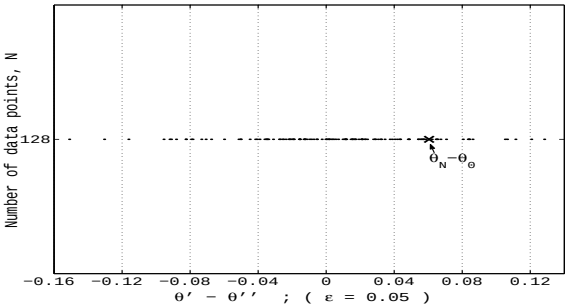

Fig. 2. $\theta' - \theta''$ values. ($N = 128$)

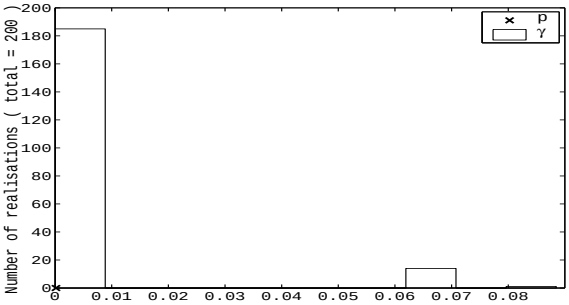

Fig. 3. Histogram of $\sqrt{2\xi} = \gamma$. ($N = 512$)

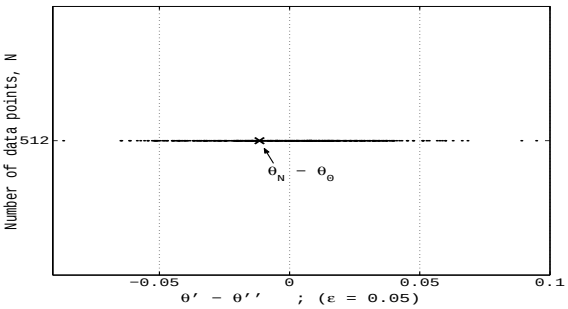

Fig. 4. $\theta' - \theta''$ values. ($N = 512$)

Comparing the result obtained with $N = 128$ and $N = 512$, we observe that as N increases, ξ decreases. This is as expected, since, as we get more and more data points, the quality of the estimate will get better, and so does the quality of the partial estimates.

For $N = 512$, 185 realisations out of the 200 gave $\sqrt{2\xi} = 0$, and the rest gave values around 0.063; which corresponds to the smallest non-zero ξ value ($= 1/511$) and 0.088; which corresponds to the second smallest non-zero ξ value ($= 2/511$). Hence, $p > \sqrt{2\xi}$ in most of the realisations since $p = 1.6 \times 10^{-4}$. According to p the event will happen in about one out of 6000 outcomes, and since we only have 512 data points, it is not likely that we will observe it. Hence, the bound is too tight. However, we bound p using $\sqrt{2(\xi + \rho)}$. Since ρ is a design variable, we can choose ρ to be small, but this causes μ_h to be small too, and hence we have low confidence in this particular claim about the quality of the estimate.

From the scatter plots of $\theta' - \theta''$, based on one simulation it is observed that the partial estimates form a region around zero, and $\hat{\theta}_N - \theta_0$ is within this region. This shows that the quality of the estimate $\hat{\theta}_N$ can be measured by the variability in the partial estimates in this simple example.

In order to check the tightness of the second probability, μ_h is computed for different values of N and ρ, and the results are shown in Table 1 together with the frequency μ_s in the 200 experiments that $Pr\{|\hat{\theta}_N - \theta| \geq \varepsilon\} \leq \sqrt{2(\xi + \rho)}$.

Table 1. μ_h with different N and ρ

N	ρ	μ_h	μ_s
128	0.06	0.20	1
	0.08	0.61	1
512	0.06	0.950	1
	0.08	0.997	1

From Table 1, with $N = 128$, we do not place much confidence in the obtained statement about the quality of the estimate (μ_h low) even though it holds true for all the simulations ($\mu_s = 1$). However, this bound is much tighter for the simulation with $N = 512$.

Naturally, we want to have μ_h close to 1 since we want high confidence in the claim about the quality of the estimate. As $\mu_h = 1 - 2e^{(-2(N-1)\rho^2)}$, its value can be increased by increasing N with fixed ρ as shown in Table 1. However usually the number of data points are fixed and we can not increase N. The other way to increase μ_h is to increase ρ. However, this leads to a more conservative bound on $Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\}$. ρ can not be larger than 0.5, otherwise $\sqrt{2(\xi + \rho)}$ will be larger than 1, and the claim about the quality of the estimate in equation (8) contains no information. Therefore, there is a natural trade-off between the conservativeness in the bound on the probability of $|\hat{\theta}_N - \theta_0| \geq \varepsilon$ and the confidence in the claim about the quality.

As mentioned, the final result contains two probabilities. The first probability p, or rather its empirical upper bound $\sqrt{2(\xi + \rho)}$ is used as a measure of the quality of $\hat{\theta}_N$, the smaller the probability the better the estimate. The second probability which is due to bounding $Pr\{|\theta' - \theta''| \geq 2\varepsilon\}$ by its empirical counterpart, gives the confidence level in the quality claim about $\hat{\theta}_N$. The higher the probability, the higher confidence we have in the claim. Therefore, in general we aim for a result with a low value of p and a value of μ_h close to 1, meaning that we obtain a statement saying that the estimate is good and that we have a high level of confidence in the statement.

For an arbitrary large value of ε and an arbitrary small positive value of ρ, there are values of N and σ^2 such that both p and μ_h are arbitrary close to 1. Seemingly this would lead to a result saying that we have high confidence in that the estimate is bad. However, p is unknown and is bounded from above by $\sqrt{2(\xi+\rho)}$. If $\sqrt{2(\xi+\rho)}$ is large, we can not say that the estimate is bad (even if it is) since $\sqrt{2(\xi+\rho)}$ is only an upper bound on p.

5. CONCLUSIONS

In this paper we have discussed requirements to a non-asymptotic measure for model quality, and proposed a technique for assessing the quality of the LS estimate by computing a bound on $\hat{\theta}_N$ using partial estimates. The main principle is to measure the quality of the LS estimate in terms of the variation in partial estimates. If $\theta' - \theta''$ is small, the estimate, as it is intuitive, is good. On the other hand if $\theta' - \theta''$ is spread all over the place, then we do not place much confidence in the estimate. As a starting point, we have investigated this approach using a first order FIR model with a constant input signal, and a Gaussian noise signal with zero mean and variance σ^2. The probability $Pr\{|\hat{\theta}_N - \theta_0| \geq \varepsilon\}$ is bounded in terms of $Pr\{|\theta' - \theta''| \geq 2\varepsilon\}$, and the latter is estimated from the observed data. However, just replacing $Pr\{|\theta' - \theta''| \geq 2\varepsilon\}$ by its estimate leads to a conceptual error since the estimate is stochastic. A second probability is therefore needed in order to assert the probability with which the claim holds. This second probability is estimated using Hoeffding's inequality. The results obtained are uniform with respect to the data generating mechanism, hence they are valid for all true systems in the model class and all values of σ^2.

Admittedly the example studied is very simple and more work is needed in order to extend the method and the results to more general settings. This is a topic for current research, and a particularly challenging problem in this respect is to generalise the computation of the outer probability to non-Gaussian and non iid noise sequences.

A PROOF OF LEMMA 2

The proof is by induction. Let $N = 2^l$; $l = 1, 2, ...$

(1) For $l = 1$, the vector $a_{1,N-1} = a_{1,1} = [-1\ 1]$ satisfies the claim.

(2) Now assume that the claim in the Lemma is true for $N = 2^n$ for some $n \geq 1$ and label the sequences as $a_{n,1}, a_{n,2}, ..., a_{n,2^n-1} = a_{n,N-1}$.

(3) Then for $N = 2^{n+1}$ the following $N-1$ vectors are mutually orthogonal

$$
\begin{aligned}
a_{n+1,1} &= \left[1 \ldots 1\ -1 \ldots -1\right] \\
a_{n+1,2} &= \left[a_{n,1} a_{n,1}\right] \\
a_{n+1,3} &= \left[a_{n,2} a_{n,2}\right] \\
&\vdots \\
a_{n+1,N-2^n} &= \left[a_{n,2^n-1} a_{n,2^n-1}\right] \\
a_{n+1,N-2^n+1} &= \left[-a_{n,1} a_{n,1}\right] \\
&\vdots \\
a_{n+1,N-2^n+(n+1)} &= \left[-a_{n,2^n-1} a_{n,2^n-1}\right] \\
&= a_{n+1,N-1}
\end{aligned}
$$

(4) By induction, for all vectors with $N = 2^l$ ($l = 1, 2, ...$) elements there are at least $N-1$ vectors that are mutually orthogonal.

Let a be a vector which is a linear combinations of vectors which have an equal number of 1's and -1's. Then the sum of the elements of a is zero, and hence there can not be N mutual orthogonal vectors consisting of an equal number of 1's and -1's, as they would have formed a basis for R^N. ■

6. REFERENCES

Campi, M. C. and E. Weyer (2002). Finite sample properties of system identification methods. Accepted for publication in *IEEE Trans on Automatic Control*.

Edgington, E. (1995). *Randomization Tests*. 3rd ed.. Marcel Dekker.

Good, P. (1999). *Resampling methods: A practical guide to data analysis*. Birkhäuser.

Good, P. (2000). *Permutation Tests: A practical guide to resampling methods for testing hypotheses*. 2nd ed.. Springer-Verlag.

Koltchinskii, V. and D. Panchenko (2000). Rademacher processes and bounding the risk of function learning. In: *High Dimensional Probability II* (E.Gine, D.Mason and J.Wellner, Eds.). pp. 443–459. Birkhäuser.

Ljung, L. (1999). *System Identification: Theory For The User*. 2nd ed.. Prentice Hall.

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

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York.

Vidyasagar, M. (1997). *A Theory of Learning and Generalization*. Springer-Verlag.

Weyer, E. (2000). "Finite sample properties of system identification of ARX models under mixing conditions". *Automatica* **36**, 1291–1299.

Weyer, E. and M. Campi (1999). "Finite sample properties of system identification methods". *Proceedings of the 38th IEEE CDC, Phoenix, Arizona, USA* pp. 510–515.

Weyer, E. and M. Campi (2000). "Non-asymptotic confidence ellipsoids for the least squares estimate". *Procedings of the 39th IEEE CDC, Sydney, Australia* pp. 2688–2693. Extended version to appear in Automatica.