

CONSISTENT LOW-COMPLEXITY ESTIMATION OF ACTIVE PARAMETERS IN LARGE LINEAR REGRESSIONS¹

Fredrik Gustafsson

Department of Electrical Engineering, Linköping University, S-581 83 Linköping

Abstract. Some important practical signals and systems can be modeled by very large linear regression models where it is reasonable that most of the parameters are zero. We give an efficient method to solve this combined estimation and structure determination problem. It is related to Akaike-like criteria, and is based on one LMS filter and thus it is of low complexity. Asymptotic analysis shows that the method is consistent for finite impulse response models. A recursive algorithm is derived, which can be applied to time-varying systems as well. An example shows the efficiency of the approach.

1. INTRODUCTION

We will consider linear regression models of the form

$$y(t) = \varphi(t)^T \theta + e(t), \quad (1)$$

where $\varphi(t)$ is a regression vector and θ the corresponding parameter vector. Here $e(t)$ is assumed to be white Gaussian noise with variance σ^2 . It is implicitly assumed that the number of parameters $d = \dim \theta$ is large (e.g. $d > 100$). The assumption in this contribution is that only a small number n of parameters (e.g. $n < 10$) are active

$$\begin{aligned} \theta(k_i) &\neq 0, \quad i = 1, 2, \dots, n \\ \theta_i &= 0, \quad \text{otherwise} \end{aligned} \quad (2)$$

By k^n we denote the set of active parameters k_1, k_2, \dots, k_n . We point out two important applications where this is a realistic assumption.

1. Multipath signal propagation. In telephone communication, echoes can deteriorate the speech quality severely. In 4-wire loop telephony, the echoes come from circuit echo paths (Sondhi and Berkley 1980), while in mobile radio channels they are caused by room acoustic echo paths. The effect can be removed by equalization once a channel model has been identified.

The signal can be written as

$$y(t) = \sum_{i=1}^n \theta(k_i) u(t - k_i) + e(t) \quad (3)$$

where u is the interesting speech signal. This finite impulse response (FIR) model is commonly used in communication applications. The indices k_i for active coefficients correspond to the time delay in the echo path. The same problem occurs in sonar applications (Burdic

1984), where these echoes are caused by reflections at the surface and the bottom of the sea and also in geophysical signal processing (E.A.Robinson and T.S.Durrani 1986).

2. Approximation using basis functions. In system identification, the use of (orthonormal) basis functions has become a popular approach recently (Wahlberg 1991, Ninness 1993, Van den Hof *et al.* 1993). Here, the system is modeled by

$$y(t) = \sum_{i=1}^n \theta(k_i) \psi_{k_i} * u(t) + e(t) \quad (4)$$

with $\psi_{k_i} * u(t)$ denoting the convolution of the basis function $\psi_{k_i}(t)$ and the system input $u(t)$. A similar problem occurs in function approximation using, for instance, polynomial basis functions. In this case, $u(t)$ is the function to be approximated.

Note that both (3) and (4) are linear regressions. The problem is to estimate the number of active coefficients n , their positions k^n and their values $\theta(k_i)$. The upper bound d on the number of parameters is assumed to be known (it might be taken as the number of data). Any conceivable method that claims to be optimal for this problem has to examine all possible combinations of k^n . If n was known, there would be $\binom{n}{d}$ different combinations to examine. Here, where n is unknown, there are $\sum_{n=0}^d \binom{n}{d} = 2^d$ different combinations, the latter expression coming from the fact that each coefficient can be either active or inactive.

Basically, the described problem is a classical model structure selection one, and there is a large number of proposed methods, see for instance (Veres 1991, Gustafsson and Hjalmarsson 1995). Another approach might be to estimate the full parameter vector and apply a hypothesis test to each component. However, because of the large complexity of the problem under consideration, these are not to recommend. In the area of function approximation using wavelet basis, a method as simple as

¹ Submitted to SYSID'97

the one obtained here is proposed in (D.L.Donoho 1992).

This contribution is an extension of (?), where no consistency proof was given.

2. NOTATION

Consider the model (1) with the assumption (2). Assume that we for each possible combination $k^n = (k_1, k_2, \dots, k_n)$ minimize the sum of squared residuals,

$$V_N(k^n) = \frac{1}{N} \sum_{t=1}^N [y(t) - \varphi^T(t; k^n) \hat{\theta}^{RLS}(k^n)]^2 \quad (5)$$

$$= \min_{\theta} \frac{1}{N} \sum_{t=1}^N [y(t) - \varphi^T(t; k^n) \theta(k^n)]^2. \quad (6)$$

Here $\varphi(t; k^n)$ denotes the regression vector where only the elements k^n are kept. The least squares estimate can be written

$$\hat{\theta}_N^{RLS}(k^n) = R_N(k^n)^{-1} f_N(k^n) \quad (7)$$

where

$$f_N(k^n) = \frac{1}{N} \sum_{t=1}^N \varphi(t; k^n) y(t) \quad (8)$$

$$R_N(k^n) = \frac{1}{N} \sum_{t=1}^N \varphi(t; k^n) \varphi^T(t; k^n) \quad (9)$$

Although the least squares estimate is here expressed in off-line notation, the tag RLS (recursive least squares) will be used with the implicit assumption that the estimate can (and will) be computed on-line when used in applications.

The loss function can now be rewritten in a standard manner

$$V_N(k^n) = \frac{1}{N} \sum_{t=1}^N (y(t) - \varphi(t; k^n)^T \hat{\theta})^2 \quad (10)$$

$$= \frac{1}{N} \sum_{t=1}^N y^2(t) - f_N^T(k^n) R_N^{-1}(k^n) f_N(k^n) \quad (11)$$

$$= \hat{\sigma}_y^2 - \frac{1}{N} f_N^T(k^n) \hat{\theta}_N^{RLS}(k^n) \quad (12)$$

where $\hat{\sigma}_y^2$ is the variance of the observed output $y(t)$.

The simplified algorithm we propose is based on the LMS (least mean square) estimate with a rather non-standard stepsize. The LMS estimate is here defined as

$$\hat{\theta}_N^{LMS}(k^n) = D_N(k^n)^{-1} f_N(k^n) \quad (13)$$

where

$$D_N(k^n) = \text{diag}(\hat{\sigma}_\varphi^2(k_1), \hat{\sigma}_\varphi^2(k_2), \dots, \hat{\sigma}_\varphi^2(k_n)) \quad (14)$$

$$\hat{\sigma}_\varphi^2(k) = \frac{1}{N} \sum_{t=1}^N \varphi(t; k)^2 \quad (15)$$

Note that the only difference between the RLS and LMS estimates here is that the off-diagonal elements in R_N are set to 0 when computing the LMS estimate so the inversion of D_N is done elementwise on the diagonal elements.

3. OUTLINE AND SUMMARY

The contribution in this paper is to convert a standard test with known properties, but infeasible to compute for this problem, to a very simple test and analyse its properties. Including two intermediate steps, we will discuss the following tests:

Original test:

$$\hat{k}^n = \arg \min_{k^n} V_N(k^n) + n\gamma(N) \quad (16)$$

Equivalent test:

$$\hat{k}^n = \arg \min_{k^n} -\frac{1}{N} f_N^T(k^n) \hat{\theta}_N^{RLS}(k^n) + n\gamma(N) \quad (17)$$

Simplified test:

$$\hat{k}^n = \arg \min_{k^n} -\frac{1}{N} f_N^T(k^n) \hat{\theta}_N^{LMS}(k^n) + n\gamma(N) \quad (18)$$

Equivalent simplified test: Keep the coefficients $\theta(k)$ for which

$$\hat{\theta}_N^{LMS}(k) > \sqrt{\frac{\gamma(N)}{\hat{\sigma}_\varphi^2(k)}} \quad (19)$$

The first test is standard in the area of model structure selection, where there is an inherent trade-off between model fit (first term) and model complexity (second penalty term). This is also known as the *parsimonious principle* and *Ockham's razor*. Depending on the choice of $\gamma(N)$ different criteria are obtained. For instance, $\gamma(N) = \log(N)/N$ yields Akaike's BIC (Information Criterion B) (Akaike 1981) which coincides with Rissanen's MDL (minimum description length) (Rissanen 1978).

However, this test is infeasible for large linear regression models because of the following two problems:

- The test requires the computation of the least squares solution to problems of very high dimension.
- We need to compare 2^d different model structures and apparently there is no way to avoid computing the least squares estimate to each of them.

The remedy we propose to the first problem is to replace the least squares estimate in (17) by the least mean square estimate in (18). Fortunately, this test can be rewritten in the equivalent form (19) which resolves the second problem.

The remaining question is how to choose the penalty term $\gamma(N)$. The conditions for which these tests are consistent are summarized below:

Original test consistent if and only if:

$$\begin{aligned} \lim_{N \rightarrow \infty} \gamma(N) &= 0 \\ \lim_{N \rightarrow \infty} N\gamma(N) &= \infty \end{aligned} \quad (20)$$

Simplified test consistent for FIR models if:

$$\begin{aligned} \lim_{N \rightarrow \infty} \gamma(N) &= 0 \\ \lim_{N \rightarrow \infty} \sqrt{N}\gamma(N) &= \infty \end{aligned} \quad (21)$$

Since the penalty term must be less for the simplified criteria, we trade off computational complexity with convergence time.

The analysis also gives a clue on the actual form of the penalty term. We propose $\gamma(N) = \hat{\sigma}_y^2 \log(N) / \sqrt{N}$. The algorithm then becomes:

Algorithm: Keep the coefficients $\theta(k)$ for which

$$\hat{\theta}_N^{LMS}(k) > \frac{\hat{\sigma}_y}{\hat{\sigma}_{\varphi(k)}} \sqrt{\frac{\log(N)}{\sqrt{N}}} \quad (22)$$

This test can be interpreted as the following *hypothesis test*:

$$H_0(k) : \theta_k = 0 \quad (23)$$

$$H_1(k) : \theta_k \neq 0 \quad (24)$$

The hypothesis test would here look like

$$|\hat{\theta}_N(k)| > c_\alpha \sqrt{P_{kk}}$$

One problem is that we do not know the variance and distribution of the LMS estimate. Another one is that we face a multiple hypothesis test but design d independent tests, so we do not know the total confidence level.

4. PROOF OF CONSISTENCY

The consistency results stated in the previous section follow from the three theorems in this section:

- Consistency for the standard model structure selection test for this problem.
- A quantification of the error introduced when replacing the RLS estimate with the LMS estimate.
- Consistency of the proposed test.

The proof of the first theorem follows well-known lines, resembling for instance the proof in (Söderström and Stoica 1989). It is included because the proof of the third theorem is based on this one. The proof of the second theorem is the core of the analysis.

Theorem 1. Suppose the signal can be described by (1) for a particular $k^n = \Omega_0$ in (2). Then, the test (16) is consistent if and only if the two conditions in (20) are satisfied.

Proof: The proof is based on comparing two different sets of k^n , Ω_1 and Ω_2 , and distinguishing the cases of over-modeling and under-modeling. In this way we show that asymptotically, the estimate will be neither over-modeled nor under-modeled. Define

$$W_N^{RLS}(\Omega_i) = V_N(\Omega_i) + n_i \gamma(N)$$

where n_i is the number of elements k_i in Ω_i .

Case 1. Suppose $\Omega_0 \subset \Omega_2$ and $\Omega_0 \not\subset \Omega_1$. That is, the second set includes the active parameters and possibly some more but the first one does not. Then, it is well-known that the least squares loss function has expected value

$$EV_N(\Omega_2) = \sigma_e^2 \left(1 - \frac{n_2}{N}\right). \quad (25)$$

On the other hand, the other set gives rise to a non-vanishing bias term b so

$$EV_N(\Omega_1) = \sigma_e^2 + b^2.$$

Comparing the two quantities that we want to minimize in (16) gives

$$\begin{aligned} E(W_N^{RLS}(\Omega_1) - W_N^{RLS}(\Omega_2)) \\ = \sigma_e^2 \frac{n_2}{N} + b^2 + (n_1 - n_2)\gamma(N) \\ \rightarrow b^2 > 0 \end{aligned}$$

as N tends to infinity if the condition $\gamma(N) \rightarrow 0$ holds. That is, asymptotically under-modeling is avoided. Note that this condition is not only sufficient but also necessary. Now, we know that $NV_N(\Omega_i)$ is $\chi^2(N - n_i)$ distributed (non-central for Ω_1) and thus the variance of $V_N(\Omega_i)$ tends to 0 as $O(N^{-1})$, so

$$(W_N^{RLS}(\Omega_1) - W_N^{RLS}(\Omega_2)) \rightarrow b^2 > 0$$

with probability one.

Case 2. Suppose $\Omega_0 \subset \Omega_2$ and $\Omega_0 \subset \Omega_1$. Without loss of generality, we can assume that $\Omega_2 = \Omega_0$. From equation II.84(b) in (Ljung 1987) we have

$$N \frac{V_N^{RLS}(\Omega_0) - V_N^{RLS}(\Omega_1)}{\sigma_e^2} \rightarrow \chi^2(n_1 - n_0) > 0 \quad (26)$$

This positive term with N -independent variance will be killed by the penalty term with probability one if

$$(n_0 - n_1)N\gamma(N) \rightarrow -\infty$$

as N tends to infinity. Thus, the condition is that $N\gamma(N) \rightarrow \infty$. That is, asymptotically over-modeling is avoided. Again, note that this condition is not only sufficient but also necessary. \square

Theorem 2. The error when replacing the RLS estimate in (7) by the LMS estimate in (13) (when computing the projection on f_N defined in (8)), is bounded as

$$\begin{aligned} \mathbb{E} \lim_{N \rightarrow \infty} \frac{1}{\sqrt{N}} |f_N^T(k^n) \hat{\theta}_N^{RLS}(k^n) - f_N^T(k^n) \hat{\theta}_N^{LMS}(k^n)| \\ \leq \hat{\sigma}_y^2 \sqrt{n(n-1)} < \hat{\sigma}_y^2 n \end{aligned}$$

where $\hat{\sigma}_y^2$ is the variance of the observed output $y(t)$.

Remark 3. When $n = 1$, RLS and LMS coincide and the upper bound becomes 0.

Proof: In the proof, we will let the dependence of N and k^n be implicit. Using matrix notation $Y = (y(1), \dots, y(N))^T$, $\Phi = (\varphi(1), \dots, \varphi(N))^T$, we have

$$\begin{aligned} & |f_N^T(k^n) \hat{\theta}_N^{RLS}(k^n) - f_N^T(k^n) \hat{\theta}_N^{LMS}(k^n)| \\ &= |Y^T \Phi (R^{-1} - D^{-1}) \Phi^T Y| \\ &\leq \|Y\|_2^2 \|\Phi (R^{-1} - D^{-1}) \Phi^T\|_2 \\ &= N \hat{\sigma}_y^2 \|\Phi (R^{-1} - D^{-1}) \Phi^T\|_2 \\ &\leq N \hat{\sigma}_y^2 (\text{tr}(\Phi (R^{-1} - D^{-1}) \Phi^T \Phi (R^{-1} - D^{-1}) \Phi^T))^{1/2} \\ &= N \hat{\sigma}_y^2 (\text{tr}(\Phi^T \Phi (R^{-1} - D^{-1}) \Phi^T \Phi (R^{-1} - D^{-1})))^{1/2} \\ &= N \hat{\sigma}_y^2 (\text{tr}((I - RD^{-1})(I - RD^{-1})))^{1/2} \\ &= N \hat{\sigma}_y^2 (\text{tr}(I - 2RD^{-1} + RD^{-1}RD^{-1}))^{1/2} \\ &= N \hat{\sigma}_y^2 (\text{tr}(RD^{-1}RD^{-1} - I))^{1/2} \end{aligned}$$

where we have used Cauchy-Schwartz inequality, $\|A\|_2 \leq \sqrt{\text{tr}(A^T A)}$, $\text{tr}AB = \text{tr}BA$ and $\text{tr}RD^{-1} = \text{tr}I = n$ (the diagonal elements of RD^{-1} are all one), respectively. Now the result follows from Lemma 4 below, linearity of the trace operator and Jensen's inequality; the square-root is a strictly concave function so $E(\sqrt{X}) < \sqrt{E(X)}$. \square

Lemma 4. Consider the matrices D_N in (14) and R_N in (9). Assume a FIR model (3) is used with a quasi-stationary input, so

$$\begin{aligned} R_N^{ij} &= r(i-j) = \frac{1}{N} \sum_{t=1}^N u(t)u(t+j-i), \quad i, j = 1, 2, \dots, n \\ D_N^{ii} &= r(0) = \frac{1}{N} \sum_{t=1}^N u^2(t), \quad i = 1, 2, \dots, n \end{aligned}$$

will all converge. Then we have

$$\mathbb{E} \lim_{N \rightarrow \infty} N \text{tr}(R_N D_N^{-1} R_N D_N^{-1} - I) = n(n-1), \quad N \rightarrow \infty \quad (27)$$

and the variance of $\lim_{N \rightarrow \infty} N \text{tr}(R_N D_N^{-1} R_N D_N^{-1} - I)$ is bounded.

Remark 5. The definition of R_N assumes pre- and post-windowing in the least squares method.

Proof: Omitting the index N , we have

$$RD^{-1} = \begin{pmatrix} 1 & \frac{r(1)}{r(0)} & \dots & \frac{r(n-1)}{r(0)} \\ \frac{r(1)}{r(0)} & 1 & \dots & \frac{r(n-2)}{r(0)} \\ \vdots & & \ddots & \vdots \\ \frac{r(n-1)}{r(0)} & \dots & \frac{r(1)}{r(0)} & 1 \end{pmatrix} \quad (28)$$

and the diagonal elements on its square are given by

$$\text{diag}(RD^{-1}RD^{-1}) = \begin{pmatrix} 1 + \frac{r^2(1)}{r^2(0)} + \dots + \frac{r^2(n-1)}{r^2(0)} \\ \frac{r^2(1)}{r^2(0)} + 1 + \dots + \frac{r^2(n-2)}{r^2(0)} \\ \vdots \\ \frac{r^2(n-1)}{r^2(0)} + \dots + \frac{r^2(1)}{r^2(0)} + 1 \end{pmatrix}$$

Now a standard result, see (Söderström and Stoica 1989) equation (11.9), says that

$$N \frac{r^2(k)}{r^2(0)} \rightarrow \chi^2(1) \quad (29)$$

in distribution. That is,

$$\mathbb{E} \lim_{N \rightarrow \infty} N(\text{tr}(RD^{-1}RD^{-1} - I)) = n(n-1)$$

The distribution of the expression above is a weighted sum of χ^2 distributions and its variance will be a polynomial in the mean and thus bounded in N which concludes the proof. \square

Theorem 6. Suppose the observed signal can be described by an FIR model (3), with a quasi-stationary input, for a particular $k^n = \Omega_0$ in (2). Then, the test (18) is consistent if the two conditions in (21) are satisfied.

Proof: Let

$$W_N^{LMS}(\Omega_i) = -\frac{1}{N}f_N^T(\Omega_i)\hat{\theta}_N^{LMS}(\Omega_i) + n_i\gamma(N)$$

First, the comparison of two tests is rewritten into the following form:

$$\begin{aligned} & W_N^{LMS}(\Omega_1) - W_N^{LMS}(\Omega_2) \\ &= W_N^{RLS}(\Omega_1) - W_N^{RLS}(\Omega_2) \\ &+ \underbrace{\frac{1}{N} \left(-f_N^T(\Omega_1)\hat{\theta}_N^{LMS}(\Omega_1) + f_N^T(\Omega_1)\hat{\theta}_N^{RLS}(\Omega_1) \right)}_{r_1(N)} \\ &- \underbrace{\frac{1}{N} \left(-f_N^T(\Omega_2)\hat{\theta}_N^{LMS}(\Omega_2) + f_N^T(\Omega_2)\hat{\theta}_N^{RLS}(\Omega_2) \right)}_{r_2(N)} \end{aligned} \quad (30)$$

where (12) has been used. The first term is exactly the one analyzed in the proof of Theorem 1, and the last two ones can be bounded by using Theorem 2. The two cases in the proof of Theorem 1 need to be modified slightly:

Case 1. Denote the last terms in (31) by $r_1(N)$ and $r_2(N)$. From Theorem 2 these are both $O(N^{-1/2})$. That is, the bias term will still dominate asymptotically:

$$\begin{aligned} & W_N^{LMS}(\Omega_1) - W_N^{LMS}(\Omega_2) \\ &= EV_N(\Omega_1) + n_1\gamma(N) + r_1(N) - EV_N(\Omega_2) - n_2\gamma(N) - r_2(N) \\ &= \sigma_e^2 \frac{n_2}{N} + b^2 + (n_1 - n_2)\gamma(N) + O(N^{-1/2}) \\ &\rightarrow b^2 > 0 \end{aligned}$$

Case 2. Now we get

$$\begin{aligned} & \sqrt{N}(W_N^{LMS}(\Omega_1) - W_N^{LMS}(\Omega_2)) \\ &= O(1) + (n_1 - n_0)\sqrt{N}\gamma(N) \\ &\rightarrow \pm\infty \end{aligned}$$

where the $O(1)$ term comes from $\sqrt{N}r_i(N)$ using Theorem 2, if $\sqrt{N}\gamma(N)$ tends to infinity as N tends to infinity. Note that $\sqrt{N}(V_N(\Omega_0) - V_N(\Omega_1))$ tends to zero with probability one from (26). \square

5. IMPLEMENTATION

In the test (18) we have

$$\begin{aligned} \frac{1}{N}f_N^T(k^n)\hat{\theta}^{LMS}(k^n) &= \frac{1}{N}(\hat{\theta}^{LMS}(k^n))^T D_N(k^n)\hat{\theta}^{LMS}(k^n) \\ &= \sum_{i=1}^n \hat{\sigma}_\varphi^2(k_i)(\hat{\theta}^{LMS}(k_i))^2 \end{aligned}$$

Since we want to make this sum as large as possible and at the same time with as few parameters as possible the strategy is clear. Start with the k_i which gives the largest contribution and keep iterating. It is clear that the iterations should continue as long as the condition in test (19) is satisfied.

According to the condition for consistency, the decay of $\gamma(N)$ should be a function “between” 1 and $1/\sqrt{N}$. Inspired by BIC/MDL we propose the function $\log(N)/\sqrt{N}$. Since the role of the penalty term is to kill the approximation errors (r_1 and r_2) in the simplified test, Theorem 2 suggests the scaling $\hat{\sigma}_y^2$. That is, the penalty term $n\gamma(N) = n\hat{\sigma}_y^2 \log(N)/\sqrt{N}$ in algorithm (22) seems logical.

Also note that conditioned on a particular k^n , there is generally no need to re-estimate the parameters using RLS. If the distance of the active parameters are large, or the input is white, the inputs corresponding to active parameters will be approximately uncorrelated, and the LMS estimates using (13) coincide approximately with the RLS ones.

The total algorithm is given below.

Algorithm 1. Compute recursively for $t = 1, 2, \dots$ and for $k = 1, 2, \dots, d$

$$f_t(k) = \frac{t-1}{t}f_{t-1}(k) + \frac{1}{t}\varphi(t; k)y(t) \quad (32)$$

$$D_t(k) = \frac{t-1}{t}D_{t-1}(k) + \frac{1}{t}\varphi(t; k)^2 \quad (33)$$

$$\hat{\sigma}_{y,t}^2 = \frac{t-1}{t}\hat{\sigma}_{y,t-1}^2 + \frac{1}{t}y^2(t) \quad (34)$$

$$\hat{\theta}_t(k) = \frac{f_t(k)}{D_t^{-1}(k)} \quad (35)$$

Keep the coefficients for which

$$\hat{\theta}_t(k) > \frac{\hat{\sigma}_{y,t}}{D_t(k)} \sqrt{\frac{\log(t)}{\sqrt{t}}} \quad (36)$$

This algorithm is trivially extended in a natural but *ad-hoc* way to time-varying systems, where both the parameter values and active parameter positions may change in time. The motivation is that the systems in telephone and sonar applications are time-varying and there is a strong need for recursive estimation.

Algorithm 2. Choose a forgetting factor $\lambda < 1$ and compute recursively

$$f_t(k) = \lambda f_{t-1}(k) + (1 - \lambda)\varphi(t; k)y(t) \quad (37)$$

$$D_t(k) = \lambda D_{t-1}(k) + (1 - \lambda)\varphi(t; k)^2 \quad (38)$$

$$\hat{\sigma}_{y,t}^2 = \lambda \hat{\sigma}_{y,t-1}^2 + (1 - \lambda)y^2(t) \quad (39)$$

$$\hat{\theta}_t = \frac{f_t(k)}{D_t^{-1}(k)} \quad (40)$$

Keep the coefficients for which

$$\hat{\theta}_t(k) > \frac{\hat{\sigma}_{y,t}}{D_t(k)} \sqrt{-\log(1 - \lambda)\sqrt{1 - \lambda}} \quad (41)$$

Note that N is replaced by $1/(1 - \lambda)$ here.

6. APPLICATION TO ECHO DETECTION

We here describe an example similar to one in (J.Homer *et al.* 1994). The underlying model is (3). The channel's impulse response has three active taps:

$$y(t) = 6u(t - 10) - 5u(t - 51) - 1.5u(t - 71) + e(t).$$

The input and measurement noise are independent Gaussian white noises with variances 1 and 16, respectively. This implies an SNR of $E((y - e)^2)/E(e^2) = 63.25/16 \approx 4$.

Figure 1 shows the result of Algorithm (22) averaged over 100 simulations. The correct taps were found in all simulations. The upper plot shows the estimated active taps. They converge quickly to the correct values 6, 5 and 1.5. Also the number of estimated active taps converges quickly to the correct number 3 as shown in the lower plot.

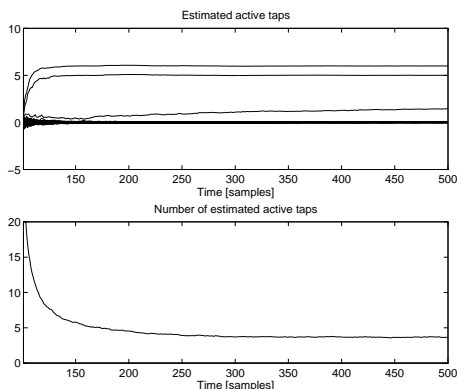


Fig. 1. Estimated active taps (upper plot) and estimated number of active taps (lower plot)

7. CONCLUSIONS

We have investigated a combined structure determination and parameter estimation problem for linear re-

gression models where most of the parameters are expected to be zero. Most conceivable methods like the BIC method lead to minimization of a criterion including the least squares loss function that must be evaluated a huge number of times. We have here proposed a way to simplify the loss function by using the least mean square estimate instead of the least squares one. The method was proved to be consistent for finite impulse response systems, which is often used in communication problems. The price paid for computational complexity is a slower convergence. A recursive algorithm was pointed out, that can be applied to time-varying systems. The algorithm was tested on an echo equalization problem.

8. REFERENCES

- Akaike, H. (1981). Modern development of statistical methods. In: *Trends and Progress in System Identification* (P. Eykoff, Ed.). Pergamon Press. Oxford.
- Burdic, W.S. (1984). *Underwater Acoustic System Analysis*. Prentice-Hall. Englewood Cliffs, NJ.
- D.L.Donoho (1992). De-noising by soft-thresholding. Technical Report 409. Dept. of Statistics. Stanford University.
- E.A.Robinson and T.S.Durrani (1986). *Geophysical Signal Processing*. Prentice-Hall. Englewood Cliffs, NJ.
- Gustafsson, F. and H. Hjalmarsson (1995). 21 ML estimators for model selection. *Automatica* **31**(10), 1377–1392.
- J.Homer, B.Wahlberg, F.Gustafsson, I.Mareels and R.Bitmead (1994). LMS estimation of sparsely parameterized channels via structural detection. In: *Proc. on the CDC, 1994*. Florida, USA. pp. 257–262.
- Ljung, L. (1987). *System Identification, Theory for the User*. Prentice Hall. Englewood Cliffs, New Jersey.
- Ninness, B.M. (1993). Stochastic and Deterministic Modelling. PhD thesis. University of Newcastle.
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica* **14**, 465–471.
- Söderström, T. and P. Stoica (1989). *System Identification*. Prentice Hall. New York.
- Sondhi, M.M. and D.A. Berkley (1980). Silencing echoes on the telephone network. *Proceedings of the IEEE* **68**, 948–963.
- Van den Hof, P.M.J., P.S.C. Heuberger and J. Bokor (1993). Identification with generalized orthonormal basis functions—statistical analysis and error bounds. *Selected Topics in Identification Modelling and Control* **6**, 39–48.
- Veres, S.M. (1991). *Structure Selection of Stochastic Dynamic Systems, the Information Criterion Approach*. Stochastic monographs. Gordon and Breach Science Publishers. New York.
- Wahlberg, B. (1991). System identification using Laguerre models. *IEEE Transactions on Automatic Control*.