

Sequentially Normalized Least Squares Models

J. Rissanen, Teemu Roos, and Petri Myllymaki
Complex Systems Computation Group,
Helsinki Institute for Information Technology

3/1/2007

Abstract

Abstract

This paper introduces universal models based on sequentially minimized squared deviations, which are smaller than the usual sum of the least squares, which, in turn, are smaller than the squared prediction errors in the so-called 'plug-in' models. We prove that these models are asymptotically optimal for linear quadratic regression problems, where the regressor (design) matrix is either non random or determined by the observed data as in AR models, by reaching the stochastic complexity. They also provide criteria for estimation of the number of parameters and the structure where the parameters lie both for small and large amounts of data, and importantly there are no hyper parameters.

1 Introduction

In this paper we are concerned with deriving a model selection criterion for a class of normal models $f(y^n | X_n; \sigma^2, b) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2} \sum_1^n (y_t - b'\bar{x}_t)^2}$, induced by the regression equations

$$y_t = b'\bar{x}_t + \epsilon_t, \quad (1)$$

where the prime indicates transposition, $b' = (b(1), \dots, b(k))$, some k . The deviations $\{\epsilon_t\}$ are taken as an iid sequence generated by a normal distribution of zero mean and variance σ^2 . The columns $\bar{x}_t = x_{t1}, \dots, x_{tk}$ of real valued elements, defining the regressor matrices X_t , are either fixed numbers, not related to y^n , or $\bar{x}_t = \text{col}\{y_{t-1}, \dots, y_{t-k}\}$ as in AR models.

For each $t = 1, 2, \dots, n$ let $k(t)$ be the largest integer such that the least squares estimate $b'_t = (b_{t,1}, \dots, b_{t,k(t)})$ can be uniquely solved. Hence, typically $k(t) = t$ except for AR models, where $k(t) = t - 1$. Consider the three representations of data for $t = 1, 2, \dots, n$ and $k(t) \leq k$, some k

$$y_t = b'_{t-1}\bar{x}_t + e_t = \sum_{i=1}^{k(t)} b_{t-1,i}x_{t,i} + e_t, \quad (2)$$

$$y_t = b'_n\bar{x}_t + \hat{\epsilon}_t(n) = \sum_{i=1}^{k(t)} b_{n,i}x_{t,i} + \hat{\epsilon}_t(n), \quad (3)$$

$$y_t = b'_t\bar{x}_t + \hat{\epsilon}_t = \sum_{i=1}^{k(t)} b_{t,i}x_{t,i} + \hat{\epsilon}_t. \quad (4)$$

The predictor $\bar{x}'_t b_{t-1}$ of y_t in the first is sometimes called the 'plug-in' predictor, in which the parameters are calculated from the data available up to $t - 1$. The plug-in model defines a conditional normal density function for $t > m$,

$$f(y_t | y^{t-1}, X_{t-1}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_{t-1}^2}} e^{-\frac{1}{2\hat{\sigma}_{t-1}^2} e_t^2},$$

where m is the smallest integer t such that $k(t) = k$ and $y^t = y_1, \dots, y_t$. It defines the so-called *PMDL* (Predictive Minimum Description Length) criterion,

$$PMDL = \sum_{m+1}^n [\ln \hat{\sigma}_{t-1}^2 + (y_t - b'_{t-1} \bar{x}_t)^2 / \hat{\sigma}_{t-1}^2],$$

studied in [4], [2], [6], and [14]. Its special case for constant variance $\hat{\sigma}_{t-1}^2 = \sigma^2$ is the *PLS* (Predictive Least Squares) criterion, studied in [11] and [14]. Both of these are examples of Dawid's prequential models, [3].

The second representation is traditional, and it, too, has an associated model selection criterion, *BIC*, [15] or, equivalently, the early version of the *MDL* criterion, [9]. The third representation, which we are interested in, is new.

All these various criteria, including the one studied in this paper, have a common basis, which explains their good asymptotic behavior: They may be seen to be the negative logarithm of an optimal universal density model. A density function $f(y^n | X_n)$ for a class of parametric models $\mathcal{M}_k = \{f(y^n | X_n; \theta)\}$ is called *universal*, if

$$\frac{1}{n} \log \frac{f(y^n | X_n; \theta)}{f(y^n | X_n)} \rightarrow 0$$

for all parameters $\theta = \theta_1, \dots, \theta_k \in \Omega$, and *optimal universal* if the convergence is the fastest possible; the convergence is either in the mean or almost surely or both.

In case of the plug-in models, the criterion is the negative logarithm of a normal density function, whose universality follows from the consistency, and asymptotic optimality from the further analysis in the cited papers. The second representation gives a non normalized density function $f(y^n | X_n; b_n, \hat{\sigma}_n^2)$, which when normalized by $C_{n,k} = \int_{z^n \in Y} f(z^n | X_n; b_n, \hat{\sigma}_n^2) dz^n$ gives a universal *NML* (Normalized Maximum Likelihood) model, [1]. The range of integration Y requires hyper parameters. An equivalent construct as a Bayesian mixture, which also requires hyper parameters for the prior, exists, [5]. Again the universality and optimality follow from analysis.

We show in this paper that the third representation (4) also induces a conditional density function, but it is not normal. It induces the density function for all the data y^n , which we show also to be optimal universal. Because it uses more data than the plug-in model in the least squares estimate the sum of the squared deviations is smaller than in the plug-in model. Because it calculates the least squares estimates recursively, its sum of the squared deviations is smaller than even that in the traditional representation (3). All told, because it is asymptotically optimal and has the smallest fitting error without over parametrization it seems to provide an excellent criterion for model selection.

2 Linear quadratic regression models

2.1 Minimized sum of the squared deviations

For each fixed k , say for $t > m$, where m is the smallest value for t for which $k(t) = k$, the well known recursions exists, see for instance [7],

$$b_t = V_t \sum_{j=1}^t \bar{x}_j y_j = b_{t-1} + V_{t-1} \bar{x}_t (y_t - \bar{x}_t' b_{t-1}) / (1 + c_t) \quad (5)$$

$$V_t = (X_t X_t')^{-1} = V_{t-1} - V_{t-1} \bar{x}_t \bar{x}_t' V_{t-1} / (1 + c_t) \quad (6)$$

$$c_t = \bar{x}_t' V_{t-1} \bar{x}_t$$

$$d_t = \bar{x}_t' V_t \bar{x}_t$$

$$1 - d_t = 1 / (1 + c_t). \quad (7)$$

The last equality was shown in [6] and [14] with the interpretation that the quantity $1 - d_t$ is the ratio of the (Fisher) information in the first $t - 1$ observations relative to all the t observations, [14].

We derive first the sequentially minimized sum of the squared deviations. Let $T_\nu = \{1 < t(1) < t(2) < \dots < t(\nu) = n\}$ be an increasing set of indices, and consider the sum of the squared residuals

$$\hat{s}_n(T_\nu) = \hat{s}_{t(1)} + \sum_{i=1}^{\nu-1} \sum_{j=t(i)+1}^{t(i+1)} (y_j - \bar{x}_j' b_{t(i+1)})^2, \quad (8)$$

where $\hat{s}_{t(1)} = \hat{e}_1^2 + \sum_2^{t(1)} (y_j - b_j' \bar{x}_j)^2$. The minimizing set will be seen to be $T_\nu = \{1, 2, \dots, n\}$, and with the notation (4) the minimized sum of the squares for $t \leq n$, all n , is given by

$$\hat{s}_t = \sum_{j=1}^t (y_j - \bar{x}_j' b_j)^2 = \hat{s}_{t-1} + \hat{e}_t^2. \quad (9)$$

To see this write $t(i+1) = \tau'$ and $t(i) = \tau$ and

$$s_t = \sum_1^t (y_j - b_t' \bar{x}_j)^2. \quad (10)$$

Then for $\tau' - \tau > 1$

$$\begin{aligned} s_{\tau'} - s_\tau &> s_{\tau'-1} + \hat{e}_{\tau'}^2 - s_\tau \\ &> \hat{s}_{\tau'-2} + \hat{e}_{\tau'-1}^2 + \hat{e}_{\tau'}^2 - s_\tau \\ &\vdots \\ &> \hat{e}_{\tau+1}^2 + \dots + \hat{e}_{\tau'-1}^2 + \hat{e}_{\tau'}^2. \end{aligned}$$

2.2 Sequentially Normalized Model

The main objective is to get statistical information from the data with the sequentially minimized quadratic loss function. However, in order to be able to measure both the loss function and the complexity of the model in the same units, bits, the loss function is converted in the *MDL* theory into a universal density model.

Consider first the case where the variance σ^2 is fixed. Consider the non-normalized conditionals

$$f(y_t | y^{t-1}, X_t; \sigma^2, b_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - \hat{y}_t)^2}{2\sigma^2}\right), \quad (11)$$

obtained by replacing the parameter vector b in the conditional normal density function $f(y_t | y^{t-1}, X_t; \sigma^2, b)$ by the least squares estimate b_t , where

$$\hat{y}_t = \bar{x}_t' b_t.$$

If we in (11) replace the variance by the minimized variance $\hat{\sigma}_t = (1/t)\hat{s}_t$ and try to normalize the result the normalizing integral will be infinite. To make it finite would require hyper parameters. Consider instead the maximization problem

$$\max_{\sigma^2} \prod_{t=1}^n f(y_t | y^{t-1}, X_t; \sigma^2, b_t). \quad (12)$$

With the solution $\hat{\sigma}_n^2 = \frac{1}{n}\hat{s}_n$ the minimized product is $(2\pi e \hat{\sigma}_n^2)^{-n/2}$. By dropping the constant we get the non-normalized and normalized conditional density functions

$$\begin{aligned} f(y_t | y^{t-1}, X_t) &= \frac{\hat{s}_t^{-t/2}}{\hat{s}_{t-1}^{-(t-1)/2}} = \hat{s}_{t-1}^{-1/2} \left(1 + \frac{(y_t - \hat{y}_t)^2}{\hat{s}_{t-1}}\right)^{-t/2} \\ \hat{f}(y_t | y^{t-1}, X_t) &= \frac{\hat{s}_{t-1}^{-1/2} \left(1 + \frac{(y_t - \hat{y}_t)^2}{\hat{s}_{t-1}}\right)^{-t/2}}{K(y^{t-1})} \\ K(y^{t-1}) &= \hat{s}_{t-1}^{-1/2} \int_{-\infty}^{\infty} \left(1 + \frac{(y_t - \hat{y}_t)^2}{\hat{s}_{t-1}}\right)^{-t/2} dy_t. \end{aligned}$$

To get the normalizing integral we write first

$$\begin{aligned} \hat{y}_t &= \bar{x}_t' b_t = d_t y_t + \bar{y}_t \\ d_t &= \bar{x}_t' V_t \bar{x}_t \end{aligned} \quad (13)$$

$$\bar{y}_t = \bar{x}_t' V_t \sum_1^{t-1} y_i \bar{x}_i, \quad (14)$$

where \bar{y}_t does not depend on y_t . Then

$$K(y^{t-1}) = \hat{s}_{t-1}^{-1/2} \int_{-\infty}^{\infty} \left[1 + \frac{(1 - d_t)^2}{\hat{s}_{t-1}} \left(y - \frac{\bar{y}_t}{1 - d_t}\right)^2\right]^{-t/2} dy. \quad (15)$$

By change of variables

$$z = \frac{1 - d_t}{\sqrt{\hat{s}_{t-1}}} \left(y - \frac{\bar{y}_t}{1 - d_t} \right)$$

we get

$$K(y^{t-1}) = K_{t-1} = \frac{1}{1 - d_t} \int_{-\infty}^{\infty} (1 + z^2)^{-t/2} dz = \frac{\sqrt{\pi}}{1 - d_t} \Gamma\left(\frac{t-1}{2}\right) / \Gamma(t/2),$$

the last equality by the fact that z is seen to have Student's distribution.

The conditional density function is then given by

$$\begin{aligned} \hat{f}(y_t | y^{t-1}, X_t) &= \frac{\hat{s}_{t-1}^{-1/2}}{K_{t-1}} \left(1 + \frac{(1 - d_t)^2}{\hat{s}_{t-1}} \left(y_t - \frac{\bar{y}_t}{1 - d_t} \right)^2 \right)^{-t/2} \\ &= K_{t-1}^{-1} \frac{\hat{s}_t^{-t/2}}{\hat{s}_{t-1}^{-(t-1)/2}}. \end{aligned} \quad (16)$$

We see that again the predictor that maximizes the conditional density function is the plug-in predictor. By putting the initial density function $\hat{f}(y_1 | X_1) = f(y_1 | X_1; \lambda)$, where λ denotes the empty parameter and which we assume to be in the family, we get the desired parameter free density function, called *SNLS* (*Sequentially Normalized Least Squares*) model

$$\hat{f}(y^n | X_n) = q(y^m | X_m) \prod_{t=m+1}^n \hat{f}(y_t | y^{t-1}, X_t) \quad (17)$$

$$\ln 1/\hat{f}(y^n | X_n) = \frac{n}{2} \ln(\hat{s}_n) - \ln \Gamma(n/2) + \ln \prod_t \frac{\sqrt{\pi}}{1 - d_t} \quad (18)$$

$$= \frac{n}{2} \ln(2\pi e \hat{s}_n/n) + \sum_{m+1}^n \ln(1 + c_t) + \frac{1}{2} \ln n + O(1) \quad (19)$$

which is universal in the family of the normal density functions considered. The negative logarithm of the *SNLS* model (19) gives a criterion for the order selection as well as one for denoising for both small and large data sets. One of its distinguished properties is the fact that unlike the *NML* universal model it has no hyper parameters.

We conclude this subsection by a large data set behavior of the *SNLS* model, provided the regressor variables \bar{x}_t satisfy

$$\frac{1}{n} \sum_1^n \bar{x}_i \bar{x}_i' \rightarrow \Sigma. \quad (20)$$

Then

$$\frac{\ln 1/\hat{f}(y^n | X_n) - \frac{n}{2} \ln(2\pi e \hat{s}_n/n)}{\ln n} \rightarrow k + 1/2. \quad (21)$$

To obtain this we apply Stirling's approximation to the gamma function $\Gamma(n/2)$ and get then from (20) $\ln(1 + c_t) = c_t + O(1/t^2)$. By the first of the following results, derived in [11] and [14],

$$\frac{1}{\ln n} \sum_{t=1}^n c_t \rightarrow k \quad (22)$$

$$\frac{1}{\ln n} \sum_{t=1}^n d_t \rightarrow k. \quad (23)$$

we deduce (21).

We show in the rest of the paper that the *SNLS* model is also *optimal* universal in senses to be specified.

2.3 Fixed regression matrix

The first theorem shows the mean square deviations in the three representations of data (3), (2), and (4), which are of some interest. Since we need the recursive formulas (5), (??), (7) we give the results for $t > m$.

Theorem 1 *If the regressor variables are non random satisfying (20) and the data generated by (1), then*

$$\frac{1}{n-m} \sum_{t=m+1}^n E\hat{e}_t^2 = \sigma^2 \left(1 - \frac{1}{n-m} \sum_{t=m+1}^n d_t \right) \quad (24)$$

$$\frac{1}{n-m} \sum_{t=m+1}^n Ee_t^2 = \sigma^2 \left(1 + \frac{1}{n-m} \sum_{t=m+1}^n c_t \right) \quad (25)$$

$$\frac{1}{n-m} \sum_{t=m+1}^n E\hat{e}_t^2(n) = \sigma^2, \quad (26)$$

where the expectation is with the parameters b and σ .

Proof: To obtain (25) we start with $y_i = \bar{x}_i' b + \epsilon_i$

$$\begin{aligned} e_t &= y_t - \bar{x}_t' V_{t-1} \sum_{i=1}^{t-1} \bar{x}_i y_i \\ &= \epsilon_t - \bar{x}_t' V_{t-1} \sum_{i=1}^{t-1} \bar{x}_i \epsilon_i, \end{aligned}$$

and noticing that $\bar{x}_t' V_{t-1} \sum_{i=1}^{t-1} \bar{x}_i \bar{x}_i' b = \bar{x}_t' b$. Further by the fact that $\{\epsilon_t\}$ is a 0-mean iid sequence

$$Ee_t^2 = \sigma^2 \left(1 + \bar{x}_t' V_{t-1} \sum_{i=1}^{t-1} \bar{x}_i \bar{x}_i' V_{t-1} \bar{x}_t \right) = \sigma^2 (1 + c_t).$$

To derive (24) we have as above

$$\begin{aligned}
\hat{\epsilon}_t &= \epsilon_t - \bar{x}'_t V_t \sum_{i=1}^t \bar{x}_i \epsilon_i \\
E\hat{\epsilon}_t^2 &= \sigma^2[(1 - d_t)^2 + \bar{x}'_t V_t \sum_{i=1}^{t-1} \bar{x}_i \bar{x}'_i V_t \bar{x}_t] \\
&= \sigma^2[1 + d_t^2 - 2d_t + d_t - d_t^2] \\
&= \sigma^2(1 - d_t),
\end{aligned}$$

where we first added the term $d_t^2 = \bar{x}'_t V_t \bar{x}_t \bar{x}'_t V_t \bar{x}_t$ in the brackets, which made the sum term d_t , and then we subtracted the same term d_t^2 .

To prove the remaining equation (26) we have as in the previous case

$$\begin{aligned}
\hat{\epsilon}_t(n) &= \epsilon_t - \bar{x}'_t V_n \sum_{i=1}^n \bar{x}_i \epsilon_i \\
&= (1 - \mu_t) \epsilon_t - \bar{x}'_t V_n \sum_{i \neq t} \bar{x}_i \epsilon_i \\
E\hat{\epsilon}_t^2(n) &= \sigma^2[(1 - \mu_t)^2 + \bar{x}'_t V_n \sum_{i \neq t} \bar{x}_i \bar{x}'_i V_n \bar{x}_t] \\
&= \sigma^2(1 - \mu_t),
\end{aligned}$$

where $\mu_t = \bar{x}'_t V_n \bar{x}_t = \text{tr}(V_n \bar{x}_t \bar{x}'_t)$. Here we also added and subtracted the term μ_t^2 in the brackets. Then

$$\sum_{t=1}^n E\hat{\epsilon}_t^2(n) = \sigma^2(n - k).$$

The same way

$$\sum_{t=1}^m E\hat{\epsilon}_t^2(m) = \sigma^2(m - k),$$

and Equation (26) follows concluding the proof.

The next theorem shows the asymptotic optimality of the *SNLS* model in a mean sense in the case where the regressor matrix is fixed, independent from the data y^n .

Theorem 2 *Let the assumption (20) hold. Then*

$$\frac{E \ln 1/\hat{f}(y^n | X_n) - \frac{n}{2} \ln(2\pi e\sigma^2)}{\ln n} \rightarrow \frac{k+1}{2}, \quad (27)$$

for almost all parameters b and σ . Also,

$$\frac{\ln 1/\hat{f}(y^n | X_n) - \frac{n}{2} \ln(2\pi e\sigma^2)}{\ln n} \rightarrow \frac{k+1}{2}$$

almost surely.

To prove (27) take the mean in (21) and exchange the mean and the logarithm on the right hand side. We get by Jensen's inequality

$$E \ln 1/\hat{f}(y^n | X_n) \leq \frac{n}{2} \ln(2\pi e E\hat{s}_n/n) + \frac{k+1}{2} \ln n + o(\ln n).$$

Substituting (24) and applying (23) we then conclude that

$$E \ln 1/\hat{f}(y^n | X_n) \leq \frac{n}{2} \ln(2\pi e \sigma^2) + \frac{k+1+\delta}{2} \ln n, \quad (28)$$

or

$$\frac{E \ln 1/\hat{f}(y^n | X_n) - \frac{n}{2} \ln(2\pi e \sigma^2)}{\ln n} \leq \frac{k+1+\delta}{2}$$

for large enough n . By Theorem 1 in [10] the opposite inequality holds for all data generating parameters except some in a set of Lebesgue measure zero, and (27) holds.

The proof of the a.s. result is an exercise in martingales; in fact, Problem [15], page 165 in [8]. Define $\xi_t = \hat{\epsilon}_t^2 - (1-d_t)\sigma^2$, and $s_n = \sum_1^n \xi_t/t$, which is a martingale. We have

$$E s_n^2 = \sum_1^n E \xi_t^2/t^2 + 2 \sum_{i<j} \frac{E \xi_i \xi_j}{ij}.$$

Since $E \xi_i \xi_j = E E_{i|j} \xi_i \xi_j = 0$, where $E_{i|j}$ denotes the conditional expectation. Since $E \xi_t^2$ is uniformly bounded so are both $E s_n^2$ and $E|s_n|$. By Doob's martingale convergence theorem s_n converges a.s. to a finite limit. The limit is σ^2 by Kronecker's lemma. In fact,

$$\begin{aligned} S_n &= \sum_1^n \xi_t/n = \frac{\xi_1 + 2(S_2 - S_1) + \dots + n(S_n - S_{n-1})}{n} \\ &= S_n - \frac{S_1 + \dots + S_{n-1}}{n}. \end{aligned}$$

Since $S_n \leq s_n$ and s_n converges, so does S_n . Both terms in the right hand side of the last equality converge to the same limit, and hence the limit of S_n is zero.

This means that

$$\frac{1}{n} \sum_1^n \hat{\epsilon}_t^2 - \frac{\sigma^2}{n} \sum_1^n (1-d_t) \rightarrow 0$$

a.s., which we write as

$$\frac{1}{n} \sum_1^n \hat{\epsilon}_t^2 + o(1) = \frac{\sigma^2}{n} \sum_1^n (1-d_t).$$

Further

$$\ln\left[\frac{1}{n} \sum_1^n \hat{\epsilon}_t^2\right] + o(1) = \ln \sigma^2 + \ln\left(1 - \sum_1^n d_t/n\right).$$

We have by (23)

$$\sum_1^n d_t = k \ln n + O(1)$$

so that

$$\ln\left[\frac{1}{n} \sum_1^n \hat{e}_t^2\right] + o(1) = \ln \sigma^2 - \frac{k}{n} \ln n,$$

which completes the proof.

2.4 AR models

We then consider the case where the data are generated by an AR model,

$$y_t = \sum_{i=1}^k a_i y_{t-i} + \epsilon_t, \quad t \geq 1, \quad (29)$$

in which the regressor matrix is random, determined by the the data y^n , and where we write the coefficients as a_i to avoid confusing them with b_i , where the subindex refers to time i .

We have the almost sure asymptotic optimality:

Theorem 3 *Let the data be generated by an AR model (29), where the roots of the polynomial $1 - \sum_{i=1}^k a_i z^{-i}$ are inside the unit circle, and ϵ_t is an iid zero mean gaussian process with variance σ^2 . The process is also assumed to be ergodic and stationary with $E\bar{x}_t \bar{x}_t' = \Sigma$. Then for $\hat{\sigma}_n^2 = (1/n) \sum_1^n \hat{e}_i^2(n)$*

$$\ln \frac{1}{n} \hat{s}_n = \ln \hat{\sigma}_n^2 - k \frac{\ln n}{n} (1 + o(1)) \text{ a.s.} \quad (30)$$

and

$$\frac{\ln 1/\hat{f}(y^n | X_n) - \frac{n}{2} \ln(2\pi e \hat{\sigma}_n^2)}{\ln n} \rightarrow \frac{k+1}{2}$$

almost surely.

The proof, which takes advantage of the proof of the asymptotic optimality of the predictive model (2) in [14].

Proof. The proof takes advantage of the proof of the asymptotic optimality of the predictive model (2) in [14]. The beginning point is the important equality (2.6) in [14]

$$\sum_{m+1}^n e_t^2 = \sum_1^n \hat{e}_i^2(n) + \sum_{m+1}^n d_t e_t^2 - \sum_1^m \hat{e}_i^2(m), \quad (31)$$

where $\hat{e}_t(n) = y_t - \bar{x}_t' b_n$ for $b_n = \text{col}(\hat{a}_1(y^n), \dots, \hat{a}_k(y^n))$. This gives

$$\ln \frac{1}{n} \sum_{m+1}^n e_t^2 = \ln \hat{\sigma}_n^2 + \ln\left[1 + \frac{\sum_{m+1}^n d_t e_t^2}{n \hat{\sigma}_n^2} - \frac{\sum_1^m \hat{e}_i^2(m)}{n \hat{\sigma}_n^2}\right], \quad (32)$$

where $n\hat{\sigma}_n^2 = \sum_1^n \hat{\epsilon}_i^2(n)$. By Corollary 4.2.1 in [14]

$$\ln \frac{1}{n} \sum_1^n e_i^2 = \ln \hat{\sigma}_n^2 + k \frac{\ln n}{n} (1 + o(1)) \text{ a.s.} \quad (33)$$

Hence

$$\ln \left[1 + \frac{\sum_{m+1}^n d_t e_t^2}{n\hat{\sigma}_n^2} - \frac{\sum_1^m \hat{\epsilon}_i^2(m)}{n\hat{\sigma}_n^2} \right] = k \frac{\ln n}{n} (1 + o(1)) \text{ a.s.}$$

and

$$\frac{\sum_{m+1}^n d_t e_t^2}{n\hat{\sigma}_n^2} = k \frac{\ln n}{n} (1 + o(1)) \quad (34)$$

as well.

From (??),

$$\sum_{m+1}^n \hat{\epsilon}_t^2 = \sum_{m+1}^n e_t^2 - 2 \sum_{m+1}^n d_t e_t^2 + \sum_{m+1}^n d_t^2 e_t^2,$$

which with (31) and (34) gives

$$\sum_{m+1}^n \hat{\epsilon}_t^2 = n\hat{\sigma}_n^2 - \hat{\sigma}_n^2 k \ln(1 + o(1)) + \sum_{m+1}^n d_t^2 e_t^2.$$

After we show that the last term is

$$\sum_{m+1}^n d_t^2 e_t^2 = o(1) \ln n \quad (35)$$

we get the first claim, (30),

$$\sum_{m+1}^n \hat{\epsilon}_t^2 = n\hat{\sigma}_n^2 \left[1 - k \frac{\ln n}{n} (1 + o(1)) \right].$$

The second claim in the Theorem follows from (30), (19), (20), and (23).

We show first that $\bar{x}_t' \bar{x}_t \leq \alpha \ln t$ almost surely for α to be chosen later.

The density function for \bar{x}_t is gaussian

$$f(\bar{x}_t) = \frac{|\Sigma|^{1/2}}{(2\pi)^{k/2}} e^{-\frac{1}{2} \bar{x}_t' \Sigma^{-1} \bar{x}_t},$$

where by stationarity $\Sigma = E_a \bar{x}_t \bar{x}_t'$. For μ the least eigenvalue of Σ ,

$$f(\bar{x}_t) \leq \frac{|\Sigma|^{1/2}}{(2\pi)^{k/2}} e^{-\frac{\mu}{2} \bar{x}_t' \bar{x}_t}.$$

Let $A_t = \{\bar{x}_t : \bar{x}_t' \bar{x}_t \geq \alpha \ln t\}$. Then

$$P(A_t) \leq \frac{|\Sigma|^{1/2}}{(2\pi)^{k/2}} \sum_{i \geq t} \int_{B_i} e^{-\frac{\mu}{2} \bar{x}_i' \bar{x}_i} d\bar{x}_i,$$

where $d\bar{x}_i$ is the differential volume and $B_i = \{\bar{x}_t : \alpha \ln i \leq \bar{x}'_t \bar{x}_t \leq \alpha \ln(i+1)\}$. The integrand is upper bounded by $i^{-\gamma}$ for $\gamma = \alpha\mu/2$, which remains constant on the surface of the k -dimensional sphere of radius $\alpha \ln i$. Hence, the integration of the surface area over the radius difference $\alpha \ln(1 + 1/i) = O(\alpha/i)$ gives $i^{-\gamma}$ times the volume of B_i , or

$$i^{-\gamma} \int_{\alpha \ln i \leq r \leq \alpha \ln(i+1)} dr \leq O(i^{-(1+\gamma)} (\ln i)^{k-1}).$$

The sum of this from $i = t$ to $i = \infty$ is upper bounded by $O(\int_t^\infty y^{-\gamma} dy) = O(t^{1-\gamma})$ for $\gamma > 2$, and $\sum_t P(A_t) < \infty$. The claim follows by Borel-Cantelli lemma, namely, that the probability of the event that $\bar{x}'_t \bar{x}_t \geq \alpha \ln t$ infinitely often is zero.

By the ergodic theorem (20) holds, and $d_t \leq O((\ln t)/t)$ a.s.. Since $\sum_s^n d_t e_t^2 \leq O(\ln(n/s))$

$$\sum_1^n d_t^2 e_t^2 = \sum_{t=1}^{s-1} d_t d_t e_t^2 + \sum_s^n d_t d_t e_t^2 \leq O(\ln s) + O\left(\frac{\ln s}{s}\right) O\left(\ln \frac{n}{s}\right).$$

Take $s = \ln n$, which implies (35) and the proof of the theorem follows.

References

- [1] Barron, A.R., Rissanen, J., and Yu, B. (1998), 'The MDL Principle in Modeling and Coding', special issue of *IEEE Trans. Information Theory* to commemorate 50 years of information theory, Vol. **IT-44**, No. 6, October 1998, pp. 2743–2760
- [2] Davis, M.H.A. and Hemerly, E.M. (1990), 'Order Determination and Adaptive Control of ARX Models Using the PLS Criterion', *Proceedings of the Fourth Bad Honnef Conference on Stochastic Differential Systems. Lecture Notes in Control and Information Sci. (N. Christopeit, ed.)* Springer, New York
- [3] Dawid, A.P. (1984), 'Present Position and Potential Developments: Some Personal Views, Statistical Theory, The Prequential Approach', *J. Royal Stat. Soc. A*, Vol. **147**, Part 2, pp. 278–292
- [4] Hannan, E.J., Mcdougall, A.J. and POskit, D.S. (1989). 'Recursive estimation of autoregressions'. *J.Roy. Statist. Soc. Ser. B* **51**, 217-233
- [5] Hansen, M.H. and Yu, B. (2001), 'Model Selection and the Principle of Minimum Description Length', *Journal of American Statistical Association*, **96**(454), 746-774
- [6] Lai, T.L. and Wei, C.Z. (1982), 'Least Squares Estimates in Stochastic Regression Models with Applications to Identification and Control of Dynamic Systems', *Annals of Statistics*, Vol. 10, **1**, pp. 154–166

- [7] Plackett, R.L. (1950), 'Some Theorems in Least Squares', *Biometrika*, Vol. **37**, No. 1/2, pp. 149-157
- [8] Pollard, D. (2002), *A User's Guide to Measure Theoretic Probability*, Cambridge University Press, 351 pages
- [9] Rissanen, J. (1978), 'Modeling by shortest data description', *Automatica*, Vol. **14**, pp. 465-471
- [10] Rissanen, J. (1986), 'Stochastic Complexity and Modeling', *Annals of Statistics*, Vol. **14**, pp. 1080-1100
- [11] Rissanen, J. (1986), 'A Predictive Least Squares Principle', *IMA J. Math. Control Inform.* **3**, pp. 211-222
- [12] Rissanen, J. (1996), 'Fisher Information and Stochastic Complexity', *IEEE Trans. Information Theory*, Vol. **IT-42**, No. 1, pp. 40-47
- [13] Rissanen, J. (2007), *Information and Complexity in Statistical Modeling*, Springer Verlag, 142 pages
- [14] Wei, C.Z. (1992), 'On Predictive Least Squares Principles', *Annals of Statistics*, Vol. 20, **1**, pp. 1-42
- [15] Schwarz, G. (1978). 'Estimating the dimension of a model'. *Annals of Statistics*, Vol. **6**, pp. 416-464