L_2 Model Reduction and Variance Reduction

Fredrik Tjärnström, [Lennart Ljung](#)

Division of Automatic Control

E-mail: fredrikt@iys.liu.se, ljung@isy.liu.se

25th June 2007

Report no.: [LiTH-isy-R-2801](#)

Accepted for publication in Automatica, 2002

Address:

Department of Electrical Engineering

Linköpings universitet

SE-581 83 Linköping, Sweden

WWW: <http://www.control.isy.liu.se>

AUTOMATIC CONTROL
REGLERTEKNIK
LINKÖPINGS UNIVERSITET

Abstract

In this contribution we examine certain variance properties of model reduction. The focus is on L_2 model reduction, but some general results are also presented. These general results can be used to analyze various other model reduction schemes. The models we study are finite impulse responses (FIR) and output error (OE) models. We compare the variance of two estimated models. The first one is estimated *directly* from data and the other is computed by *reducing* a high order model by L_2 model reduction. In the FIR case, we show that it is *never* better to estimate the model directly from data, compared to estimating it via L_2 model reduction of a high order FIR model. For OE models we show that the reduced order model has the same variance as the directly estimated one if the reduced model class used contains the true system.

Keywords: identification, model reduction, variance reduction

L_2 Model reduction and variance reduction[☆]

F. Tjärnström*, L. Ljung

Department of Electrical Engineering, Linköpings Universitet, SE-581 83 Linköping, Sweden

Received 11 June 2001; received in revised form 4 January 2002; accepted 16 April 2002

Abstract

In this contribution we examine certain variance properties of model reduction. The focus is on L_2 model reduction, but some general results are also presented. These general results can be used to analyze various other model reduction schemes. The models we study are finite impulse response (FIR) and output error (OE) models. We compare the variance of two estimated models. The first one is estimated *directly* from data and the other one is computed by *reducing* a high order model, by L_2 model reduction. In the FIR case we show that it is *never* better to estimate the model directly from data, compared to estimating it via L_2 model reduction of a high order FIR model. For OE models we show that the reduced model has the same variance as the directly estimated one if the reduced model class used contains the true system. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Model reduction; Identification; Variance reduction

1. Introduction

There are many methods available for model reduction, e.g., balanced reduction (Moore, 1981), Hankel-norm model reduction (Glover, 1984), and L_2 model reduction (Spanos, Milman, & Mingori, 1992). The main objective, using any of these methods, is to compress a given representation of a system into a less complex one, without losing much information. One of the most extreme examples of this is the actual identification phase, where the “model” consisting of input–output data, Z^N , is mapped into an n th ($N \gg n$) order parameterized one. In the standard setting (see Section 2) this corresponds to finding the best L_2 approximation of data (given a model class). Irrespectively how the reduction phase is performed (Moore, 1981; Glover, 1984; Spanos et al., 1992), it will make it possible to keep track of the bias errors that the reduction step gives rise to. There has, however, been little discussion on how the variance of the high order estimated model maps over to the low order one. Since the variance error strongly affects the use and interpretation of the reduced model it is in many cases at least as important as the bias error. In this paper, we discuss this topic, or

more precisely, the problem of computing the variance of the reduced model.

We start by introducing notation and discussing some facts about system identification in Section 2. Some inspiration about the L_2 model reduction problem is given in Section 3. In Section 4 related approaches on estimating the variance of the reduced model are discussed. General formulas describing the covariance of the low order model are presented in Section 5. In Section 6 we explicitly compute the covariance matrix when the reduced models are of finite impulse response (FIR) type. Section 7 states the main results, i.e., that variance of the reduced model is the same as the variance of the directly estimated model. This is proved in Section 8. A simulation example is presented in Section 9, and some conclusions are given in Section 10.

2. Prediction error methods

Throughout the paper, we denote the input signal by $u(t)$, the output signal by $y(t)$, and N is the total number of measured data. We assume that $y(t)$ is generated according to

$$y(t) = G_0(q)u(t) + v(t), \quad v(t) = H_0(q)e(t), \quad (1)$$

where $G_0(q)$ is a linear time-invariant system, usually referred to as the “true system”, q is the discrete-time shift operator, i.e., $qu(t) = u(t+1)$. Furthermore, we assume that

[☆] This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Brett Ninness under the direction of Editor Torsten Söderström.

* Corresponding author.

E-mail address: fredrikt@isy.liu.se (F. Tjärnström).

the additive noise, $v(t)$, is independent of the input, $u(t)$, and that it is a filtered version of an independent and identically distributed noise sequence $e(t)$ with variance λ . The noise filter

$$H_0(q) = \sum_{i=0}^{\infty} h_i q^{-i}, \quad h_0 = 1, \quad (2)$$

is assumed to be monic and inversely stable.

The models we fit to data are parameterized by a d -dimensional real-valued parameter vector θ , i.e.,

$$y(t) = G(q, \theta)u(t) + v(t). \quad (3)$$

More specifically we study FIR and output error (OE) models. These are parameterized by

$$G(q, \theta) = \frac{B(q, \theta)}{F(q, \theta)},$$

$$B(q, \theta) = b_1 q^{-n_k} + \dots + b_{n_b} q^{-n_k - n_b + 1},$$

$$F(q, \theta) = 1 + f_1 q^{-1} + \dots + f_{n_f} q^{-n_f},$$

$$\theta = (b_1 \dots b_{n_b} f_1 \dots f_{n_f})^T, \quad (4)$$

where $F(q, \theta) = 1$ in the FIR case.

We define a loss function as the mean of the squared sum of the prediction errors (in this case the output errors)

$$V_N(\theta) = \frac{1}{N} \sum_{t=1}^N \epsilon^2(t, \theta), \quad (5)$$

$$\epsilon(t, \theta) = y(t) - \hat{y}(t|\theta) = y(t) - G(q, \theta)u(t). \quad (6)$$

The estimate of θ is taken as the minimizer of (5)

$$\hat{\theta}_N = \arg \min_{\theta} V_N(\theta) = \arg \min_{\theta} \frac{1}{N} \sum_{t=1}^N \epsilon^2(t, \theta), \quad (7)$$

i.e., we use prediction error methods (PEM). The basic result is then (Ljung, 1999, Chapter 8) that under weak conditions

$$\hat{\theta}_N \rightarrow \theta^* = \arg \min_{\theta} E \epsilon^2(t, \theta) \quad \text{as } N \rightarrow \infty. \quad (8)$$

That is, $\hat{\theta}_N$ converges to the best model provided by the model class. If the true system belongs to the model class, $\hat{\theta}_N$ converges to the “true parameter vector”, θ_0 , that satisfies $G(e^{i\omega}, \theta_0) = G_0(e^{i\omega})$ for almost all ω . If the minimizer is not unique, $\hat{\theta}_N$ converges to some value in the set of minimizers. To avoid lack of uniqueness one can regularize the loss function. This means that (5) is replaced by

$$W_N^\delta(\theta) = V_N(\theta) + \frac{\delta}{2} \|\theta - \bar{\theta}\|_2^2 \quad (9)$$

for a $\delta > 0$ and some $\bar{\theta}$ minimizing $V_N(\theta)$. See also Ljung (1999, pp. 221–222).

The expression for the distribution of the estimate is based on the central limit theorem, assuming global identifiability and some other weak conditions (see Ljung (1999,

Chapter 9)). We present it together with a general expression for the covariance of the parameter estimates assuming that the output error model (3) is used, see Kabaila (1983) and Ljung (1999, Chapter 9).

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \in AsN(0, P_\theta), \quad (10)$$

$$P_\theta = \lambda [E \Psi(t, \theta_0) \Psi^T(t, \theta_0)]^{-1} \\ \times [E \tilde{\Psi}(t, \theta_0) \tilde{\Psi}^T(t, \theta_0)] [E \Psi(t, \theta_0) \Psi^T(t, \theta_0)]^{-1}, \quad (11)$$

$$\Psi(t, \theta_0) = - \frac{d}{d\theta} \epsilon(t, \theta) \Big|_{\theta=\theta_0}, \quad (12)$$

$$\tilde{\Psi}(\tau, \theta_0) = \sum_{i=0}^{\infty} h_i \Psi(\tau + i, \theta_0). \quad (13)$$

When the noise, $v(t)$, actually is white, the covariance expression simplifies to

$$P_\theta = \lambda [E \Psi(t, \theta_0) \Psi^T(t, \theta_0)]^{-1}. \quad (14)$$

The regularized versions of (11) and (14) are

$$P_\theta = \lambda [E \Psi(t) \Psi^T(t) + \delta I]^{-1} [E \tilde{\Psi}(t) \tilde{\Psi}^T(t)] \\ \times [E \Psi(t) \Psi^T(t) + \delta I]^{-1}, \quad (15)$$

$$P_\theta = \lambda [E \Psi(t) \Psi^T(t) + \delta I]^{-1} [E \Psi(t) \Psi^T(t)] \\ \times [E \Psi(t) \Psi^T(t) + \delta I]^{-1}, \quad (16)$$

respectively (in somewhat shorthand notation).

The calculation of the distributions for other statistics is based on a linear approximation of the mapping from the parameter distribution given by (10) to the statistic of interest. This mapping is usually referred to as Gauss’ approximation formula. It states that if $\hat{\theta}_N$ is sufficiently close to $\theta^* = E \hat{\theta}_N$, we can make the approximation

$$\text{Cov } f(\hat{\theta}) \approx [f'(\theta^*)] P_\theta [f'(\theta^*)]^T \\ \approx [f'(\hat{\theta}_N)] P_\theta [f'(\hat{\theta}_N)]^T. \quad (17)$$

The quality of this approximation increases as the size of $\theta^* - \hat{\theta}_N$ decreases. Furthermore, if $\hat{\theta}_N$ is asymptotically Gaussian distributed, so is $f(\hat{\theta}_N)$.

3. Model reduction

To estimate a low order model, $G(e^{i\omega}, \eta)$, of a system, several possibilities exist. The most obvious one is to directly estimate a lower order model from data (7). As known from, e.g., Ljung (1999), the prediction/output error estimate automatically gives models that are L_2 approximations of the true system in a frequency-weighted norm, determined

by the input spectrum and noise model

$$\hat{\eta}_N^d \rightarrow \eta^* = \arg \min_{\eta} \int_{-\pi}^{\pi} |G_0(e^{i\omega}) - G(e^{i\omega}, \eta)|^2 \times \Phi_u(\omega) d\omega \quad \text{as } N \rightarrow \infty, \quad (18)$$

where $\Phi_u(\omega)$ is the input spectrum. This is just a restatement of (8). A second possibility is to estimate a high order model which is then subjected to model reduction to the desired order. See, e.g., Wahlberg (1989). For the model reduction step, a variety of methods could be applied, like truncating balanced state–space realizations, or applying L_2 norm reduction. The latter method means that the low order model, parameterized by η is determined as

$$\hat{\eta}_N^r = \arg \min_{\eta} \int_{-\pi}^{\pi} |G(e^{i\omega}, \hat{\theta}_N) - G(e^{i\omega}, \eta)|^2 W(\omega) d\omega. \quad (19)$$

Here, $G(q, \hat{\theta}_N)$ is the high order (estimated) model, and $W(\omega)$ is a weighting function.

An important question is whether this reduction step also implies a reduction of variance, i.e., if the variance of $G(e^{i\omega}, \hat{\eta}_N^r)$ (viewed as random variable through its dependence of the estimate $G(e^{i\omega}, \hat{\theta}_N)$) is lower than that of $G(e^{i\omega}, \hat{\theta}_N)$. A second question is how this variance compares with the one obtained by the direct identification method, i.e., $G(e^{i\omega}, \hat{\eta}_N^d)$.

The somewhat surprising answer is that (19) may in some cases give a lower variance than (7). Let us consider a simple, but still illustrating example. Note that throughout the paper, the expectation is taken over both u and e .

Example 1. Consider the true system

$$y(t) = u(t-1) + 0.5u(t-2) + e(t), \quad (20)$$

where the input u is white noise with variance μ , and e is white noise with variance λ . We compare two ways of finding a first-order model of this system. First, estimate b^d in the FIR model directly from data

$$\hat{y}(t|b^d) = b^d u(t-1).$$

This gives the estimate (using least squares) $\hat{b}_N^d$, with $\hat{b}_N^d \rightarrow E\hat{b}_N^d = 1$ as $N \rightarrow \infty$.

The variance of $\hat{b}_N^d$ is computed as

$$E(\hat{b}_N^d - 1)^2 = E \left(\frac{\sum_{t=1}^N u(t-1)(0.5u(t-2) + e(t))}{\sum_{t=1}^N u^2(t-1)} \right)^2 \approx \frac{\lambda + 0.25\mu}{N\mu}.$$

Note here that expectation is taken over both u and e . This is essential for the results of this contribution and is used in the rest of this paper.

The second method is to estimate a high order model (in this example second order)

$$\hat{y}(t|\theta) = \hat{y}(t|b_1, b_2) = b_1 u(t-1) + b_2 u(t-2).$$

This gives the estimated transfer function

$$G(q, \hat{\theta}_N) = \hat{b}_{1,N} q^{-1} + \hat{b}_{2,N} q^{-2}$$

with $\hat{b}_{i,N}^d$ tending to their true values, and each having an asymptotic variance of $\lambda/(N\mu)$. Now, subjecting $G(e^{i\omega}, \hat{\theta}_N)$ to the L_2 model reduction (19) to an FIR(1) model with $W(\omega) \equiv 1$ gives the reduced model

$$G(q, \hat{\eta}_N^r) = \hat{b}_{1,N}^r q^{-1} = \hat{b}_{1,N} q^{-1}.$$

The variance of the directly estimated first-order model is

$$\text{Var } \hat{b}_N^d \approx \frac{\lambda + 0.25\mu}{N\mu},$$

while the L_2 reduced model has

$$\text{Var } \hat{b}_N^r = \text{Var } \hat{b}_{1,N} \approx \frac{\lambda}{N\mu},$$

i.e., it is strictly smaller.

The prediction error methods are efficient in these cases (assuming that e is white and normal), i.e., their variances meet the Cramér–Rao bound if the model structure contains the true system (and the measurement noise is white and Gaussian). In those cases no other estimation method can beat the direct estimation method. Still, in this example it was strictly better to estimate the low order model, both in terms of variance and mean square error, by reducing a high order model than to estimate it directly from data. This somewhat unexpected result can clearly only happen if the low order model structure does not contain the true system.

4. Other approaches

Before going into the actual calculation we discuss some related approaches. Some contributions that take into account that the high order model is obtained through an identification experiment when performing model reduction are Porat and Friedlander (1985), Porat (1986), Söderström, Stoica, and Friedlander (1991), Stoica and Söderström (1989), Zhu and Backx (1993, Chapter 7), Wahlberg (1987, 1989), Tjärnström and Ljung (2001), Tjärnström (2002), and Hsia (1977, Chapter 7). The contributions by Porat and Friedlander study ARMA parameter estimation via covariance estimates. These papers contain similar tools as the ones presented in Section 5. However, the ideas only apply to time series models. The following contributors deal with models having input signals. These approaches are briefly summarized in this section.

Söderström et al. (1991) look at nested model structures. In particular they look for structures that can be embedded

in larger structures which are easy to estimate, such as ARX structures. After estimating the high order structure they reduce the estimate to the low order structure in a weighted non-linear least-squares sense. The method is called an *indirect prediction error method*. We illustrate the idea using the generalized least-squares structure.

Assume that the low order structure is of ARARX type, i.e.,

$$A(q)y(t) = B(q)u(t) + \frac{1}{D(q)}e(t), \quad (21)$$

where the polynomials $A(q)$, $B(q)$, and $D(q)$ are of orders n_a , n_b and n_d , respectively. The structure is parameterized by

$$\eta = (a_1 \dots a_{n_a} \ b_1 \dots b_{n_b} \ d_1 \dots d_{n_d}). \quad (22)$$

Now, rewrite this structure as a high order ARX structure by multiplying with $D(q)$, i.e.,

$$A(q, \eta)D(q, \eta)y(t) = B(q, \eta)D(q, \eta)u(t) + e(t) \quad (23)$$

$$\Leftrightarrow R(q, \theta)y(t) = S(q, \theta)u(t) + e(t), \quad (24)$$

where

$$\theta = (r_1 \dots r_{n_r} \ s_1 \dots s_{n_s}).$$

Note here that $\dim \eta = n_a + n_b + n_d < \dim \theta = n_a + n_b + 2n_d$. The relation between η and θ is a non-linear mapping given by (23) and (24), i.e., $\eta = F_1^{-1}(\theta)$. Now, θ can be estimated using standard least squares and η is found by minimizing

$$\hat{\eta} = \arg \min_{\eta} (F_1(\eta) - \hat{\theta}_N)^T \hat{P}_{\theta}^{-1} (F_1(\eta) - \hat{\theta}_N),$$

where $\hat{P}_{\theta}$ is an estimate of the covariance of θ . It is shown that the statistical properties of this indirect method are the same as for standard PEM, but does in many cases use fewer computations to come up with the final estimate.

Wahlberg (1987) uses an approach similar to the one in Söderström et al. (1991). First an n th-order FIR model parameterized by θ is estimated and is then reduced to a lower order model $G(q, \eta)$ subject to

$$\hat{\eta} = \arg \min_{\eta} (F_2(\eta) - \hat{\theta}_N)^T R_N (F_2(\eta) - \hat{\theta}_N),$$

where

$$F_2(\eta) = R_N^{-1} \sum_{t=1}^N G(q, \eta)u(t)\varphi(t),$$

$$\varphi(t) = (u(t-1) \dots u(t-n))^T,$$

$$R_N = \sum_{t=1}^N \varphi(t)\varphi^T(t).$$

It is shown that the estimate of η is asymptotically efficient, i.e., its covariance matrix meets the Cramér–Rao bound as

the FIR order, n , tends to infinity (in case of white Gaussian noise).

Note that both of these approaches (Wahlberg, 1987; Söderström et al., 1991) can coincide with L_2 model reduction, e.g., if η is a linear function of θ . This is the case when both η and θ parameterize an FIR model.

Zhu and Backx (1993) use another approach. They start by estimating a high order ARX model of order 20–40. This model $(\hat{A}_N^n, \hat{B}_N^n)$ is asymptotically unbiased in model order and data, with a variance equal to the noise to signal ratio multiplied by the model order, n , divided by the number of data, i.e.,

$$\text{Var } \hat{G}_N^n(e^{i\omega}) \approx \frac{n}{N} \frac{\Phi_v(\omega)}{\Phi_u(\omega)}. \quad (25)$$

See Ljung (1999, 1985), Zhu (1989). Using the estimate $\hat{A}_N^n, \hat{B}_N^n$ a new input and output sequence is generated from the old input, $u(t)$, according to

$$u_f(t) = \hat{A}_N^n(q)u(t), \quad y_f(t) = \frac{\hat{B}_N^n(q)}{\hat{A}_N^n(q)}u_f(t). \quad (26)$$

A low order OE model is then estimated from the simulated data $\{y_f(t), u_f(t)\}_{t=1}^N$. This approach is asymptotically efficient (in model order and data).

5. The basic tools

To ease the notation, the subscript N in the estimates will be dropped from now on, i.e., we use $\hat{\theta} = \hat{\theta}_N$, $\hat{\eta}^r = \hat{\eta}_N^r$, and $\hat{\eta}^d = \hat{\eta}_N^d$.

To translate the variance of one estimate $\hat{\theta}$ to another $\hat{\eta} = f(\hat{\theta})$ we use Gauss' approximation formula (17). To use this result to compute the variance of an L_2 reduced model, we need an (asymptotic) expression for how it depends on the high order model. For this we return to (19). Let the high order model be parameterized by θ , with estimate $\hat{\theta}$. Let η parameterize a low order model and define

$$\hat{\eta}(\hat{\theta}) = \arg \min_{\eta} J(\eta, \hat{\theta}) \quad (27)$$

for some function J , that depends on the lower order model η and the high order, estimated, model $\hat{\theta}$. For L_2 -reduction we use

$$J(\eta, \hat{\theta}) = \int_{-\pi}^{\pi} |G(e^{i\omega}, \eta) - G(e^{i\omega}, \hat{\theta})|^2 W(\omega) d\omega, \quad (28)$$

but the form of J is immaterial for the moment. We assume it to be differentiable, though.

Now, since $\hat{\eta}$ minimizes $J(\eta, \hat{\theta})$, we have

$$J'_\eta(\hat{\eta}(\hat{\theta}), \hat{\theta}) = 0, \quad (29)$$

where J'_η denotes the partial derivative of J with respect to its first argument. By definition (29) holds for all $\hat{\theta}$, so

taking the total derivative with respect to $\hat{\theta}$ gives

$$0 = \frac{d}{d\hat{\theta}} J'_\eta(\hat{\eta}(\hat{\theta}), \hat{\theta}) = J''_{\eta\eta}(\hat{\eta}(\hat{\theta}), \hat{\theta}) \frac{d}{d\hat{\theta}} \hat{\eta}(\hat{\theta}) + J''_{\eta\theta}(\hat{\eta}(\hat{\theta}), \hat{\theta})$$

or

$$\frac{d}{d\hat{\theta}} \hat{\eta}(\hat{\theta}) = -[J''_{\eta\eta}(\hat{\eta}(\hat{\theta}), \hat{\theta})]^{-1} J''_{\eta\theta}(\hat{\eta}(\hat{\theta}), \hat{\theta}). \quad (30)$$

This expression for the derivative, and Gauss' approximation formula (17), now give the translation of the variance of $\hat{\theta}$ to that of $\hat{\eta}$:

$$\begin{aligned} P_\eta &= N \text{Cov } \hat{\eta} \\ &= [J''_{\eta\eta}(\eta^*, \theta^*)]^{-1} J''_{\eta\theta}(\eta^*, \theta^*) P_\theta \\ &\quad \times J''_{\theta\theta}(\eta^*, \theta^*)^T [J''_{\eta\eta}(\eta^*, \theta^*)]^{-1}, \end{aligned} \quad (31)$$

where

$$\theta^* = \lim_{N \rightarrow \infty} \hat{\theta} \quad (32)$$

and

$$\eta^* = \eta(\theta^*). \quad (33)$$

This gives us a general expression for investigating variance reduction for any reduction technique that can be written as (27). Especially it holds for L_2 reduced estimates (28).

6. The FIR case

In this section we look at systems of FIR structure. We show the perhaps surprising result that estimating a high order model followed by L_2 model reduction *never* gives higher variance than directly estimating the low order model. Note here once again that the expectation is taken over both u and e in all calculations.

Suppose that data is generated by an FIR system with $d = d_1 + d_2$ parameters, i.e.,

$$\begin{aligned} y(t) &= \sum_{k=1}^{d_1} b_k u(t-k) + \sum_{k=d_1+1}^d b_k u(t-k) + e(t) \\ &= \eta_0^T \varphi_1(t) + \xi_0^T \varphi_2(t) + e(t) = \theta_0^T \varphi(t) + e(t), \end{aligned} \quad (34)$$

where e is white noise with variance λ , and u is a stationary stochastic process, independent of e , with spectrum $\Phi_u(\omega)$. The definitions of η , ξ , θ , and $\varphi(t)$ should be immediate from (34):

$$\eta_0 = (b_1 \dots b_{d_1})^T, \quad (35)$$

$$\varphi_1(t) = (u(t-1) \dots u(t-d_1))^T, \quad (36)$$

etc. Let us also introduce the notation

$$\begin{aligned} R_{11} &= E\varphi_1(t)\varphi_1^T(t), \quad R_{12} = E\varphi_1(t)\varphi_2^T(t) = R_{21}^T, \\ R_{22} &= E\varphi_2(t)\varphi_2^T(t). \end{aligned} \quad (37)$$

Note that the true frequency function can thus be written

$$G_0(e^{i\omega}) = (e^{-i\omega} \dots e^{-di\omega})\theta_0. \quad (38)$$

We now seek the best L_2 approximation (in the frequency weighting norm $\Phi_u(\omega)$) of this system of order d_1 :

$$\begin{aligned} \eta^* &= \arg \min_{\eta} \int_{-\pi}^{\pi} |G_0(e^{i\omega}) - G(e^{i\omega}, \eta)|^2 \Phi_u(\omega) d\omega \\ &= \arg \min_{\eta} E(\theta_0^T \varphi(t) - \eta^T \varphi_1(t))^2, \end{aligned} \quad (39)$$

where the second step is Parseval's identity. Simple calculations show that the solution is

$$\begin{aligned} \eta^* &= [E\varphi_1(t)\varphi_1^T(t)]^{-1} E\varphi_1(t)\varphi^T(t)\theta_0 \\ &= R_{11}^{-1} (R_{11} \ R_{12}) \begin{pmatrix} \eta_0 \\ \xi_0 \end{pmatrix} \\ &= \eta_0 + R_{11}^{-1} R_{12} \xi_0. \end{aligned} \quad (40)$$

6.1. Direct estimate

Now, the least-squares estimate $\hat{\eta}^d$ (in the following called *the direct estimate*) of order d_1 is

$$\begin{aligned} \hat{\eta}^d &= \left[\sum_{t=1}^N \varphi_1(t)\varphi_1^T(t) \right]^{-1} \sum_{t=1}^N \varphi_1(t)y(t) \\ &= \eta_0 + \left[\sum_{t=1}^N \varphi_1(t)\varphi_1^T(t) \right]^{-1} \sum_{t=1}^N \varphi_1(t)\varphi_2^T(t)\xi_0 \\ &\quad + \left[\sum_{t=1}^N \varphi_1(t)\varphi_1^T(t) \right]^{-1} \sum_{t=1}^N \varphi_1(t)e(t), \end{aligned} \quad (41)$$

where the second step follows from (34). This gives that

$$E\hat{\eta}^d \approx \eta^* \quad (42)$$

with an approximation error of order $1/N$, cf. (72). Using η^* instead of $E\hat{\eta}^d$ in the covariance calculations results in an error of order $1/N^2$. This does not affect the results since the covariance expressions are correct of order $1/N$. Moreover, the approximation involved also concerns the indicated inverse. When N is large the law of large numbers can be applied to give the result. (A technical comment: In the definition of the estimate, one may have to truncate for close-to-singular matrices. See Appendix 9.B in

Ljung (1999) for such technicalities.) Moreover

$$\begin{aligned}
 \text{Cov } \hat{\eta}^d &= E(\hat{\eta}^d - E\hat{\eta}^d)(\hat{\eta}^d - E\hat{\eta}^d)^T \\
 &\approx E(\hat{\eta}^d - \eta^*)(\hat{\eta}^d - \eta^*)^T \\
 &= E \left(\left[\sum_{t=1}^N \varphi_1(t) \varphi_1^T(t) \right]^{-1} \sum_{t=1}^N \varphi_1(t) e(t) \right) \\
 &\quad \times \left(\left[\sum_{t=1}^N \varphi_1(t) \varphi_1^T(t) \right]^{-1} \sum_{t=1}^N \varphi_1(t) e(t) \right)^T \\
 &\quad + E \left(\left[\sum_{t=1}^N \varphi_1(t) \varphi_1^T(t) \right]^{-1} \sum_{t=1}^N \varphi_1(t) \varphi_2^T(t) \xi_0 \right. \\
 &\quad \left. - R_{11}^{-1} R_{12} \xi_0 \right) \left(\left[\sum_{t=1}^N \varphi_1(t) \varphi_1^T(t) \right]^{-1} \right. \\
 &\quad \left. \times \sum_{t=1}^N \varphi_1(t) \varphi_2^T(t) \xi_0 - R_{11}^{-1} R_{12} \xi_0 \right)^T, \\
 &\approx \frac{\lambda}{N} R_{11}^{-1} + E H_N \xi_0 \xi_0^T H_N^T,
 \end{aligned} \tag{43}$$

where

$$\begin{aligned}
 H_N &= \left[\sum_{t=1}^N \varphi_1(t) \varphi_1^T(t) \right]^{-1} \\
 &\quad \times \left[\sum_{t=1}^N \varphi_1(t) \varphi_2^T(t) \right] - [R_{11}]^{-1} R_{12}.
 \end{aligned} \tag{44}$$

6.2. Reduced estimate

Let us now turn to the model reduction case. We first estimate the full system of order d using least squares. That gives the estimate $\hat{\theta}$ with

$$E\hat{\theta} = \theta_0 \tag{45}$$

and

$$\text{Cov } \hat{\theta} \approx \frac{\lambda}{N} [E\varphi(t)\varphi^T(t)]^{-1} = \frac{\lambda}{N} \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}^{-1} \tag{46}$$

with obvious partitioning according to (37). We insert this high order estimate into (28) using a frequency weighting $W(\omega) = \Phi_u(\omega)$ and perform the model reduction (27).

Note that, by Parseval's relation, (28) also can be written

$$J(\eta, \hat{\theta}) = E(\eta^T \varphi_1(t) - \hat{\theta}^T \varphi(t))^2, \tag{47}$$

cf. (39). Here $\varphi(t)$ is constructed from u as in (34), and where u has the spectrum $W(\omega) = \Phi_u(\omega)$. In the

notation of (29) we have

$$J''_{\eta\eta}(\eta, \hat{\theta}) = E\varphi_1(t)\varphi_1^T(t) = R_{11},$$

$$J''_{\eta\hat{\theta}}(\eta, \hat{\theta}) = E\varphi_1(t)\varphi^T(t)$$

$$= E\varphi_1(t)(\varphi_1^T(t) \ \varphi_2^T(t)) = (R_{11} \ R_{12}). \tag{48}$$

From (31), (46), and (48) we now find that the covariance of the reduced estimate equals

$$\begin{aligned}
 \text{Cov } \hat{\eta}^r &\approx R_{11}^{-1} (R_{11} \ R_{12}) \frac{\lambda}{N} \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}^{-1} \begin{pmatrix} R_{11} \\ R_{21} \end{pmatrix} R_{11}^{-1} \\
 &= \frac{\lambda}{N} R_{11}^{-1},
 \end{aligned} \tag{49}$$

where the last step simply follows from the definition of an inverse matrix.

Comparing with (43) we see that *this variance is strictly smaller than that obtained by direct identification*, provided $\xi_0 \neq 0$, that is, the true system is of higher order than d_1 . However, if the true system is of order d_1 we also find that the reduced model reaches the Cramér–Rao bound (if $e(t)$ is Gaussian), i.e.,

$$\text{Cov } \hat{\eta}^r \approx \frac{\lambda}{N} R_{11}^{-1}. \tag{50}$$

The conclusion from this is that *the variance of the reduced FIR model is never higher than the variance obtained by direct estimation*.

Comments: We remark that the variance reduction is related to performing the reduction step “correctly”. If (47) is approximated by the sample sum over the same input data as used to estimate $\hat{\theta}$ it follows that the reduced estimate is always equal to the direct one. This corresponds to choosing the weighting function equal to the discrete Fourier transform of the used input sequence

$$W(\omega) = |U_N(\omega)|^2, \tag{51}$$

$$U_N(\omega) = \frac{1}{\sqrt{N}} \sum_{t=1}^N u(t) e^{-i\omega t}. \tag{52}$$

Moreover, the variance reduction can be traced to the fact that the approximation aspect of the direct estimation method depends on the finite sample properties of u over $t=1, \dots, N$. If expectation is carried out only with respect to e we have (see (40) and (41))

$$E_e \hat{\eta}^d = \eta^* + H_N \xi_0$$

and this is the reason for the increased variance in the direct method.

7. Main result

The result that it may be advantageous to use L_2 model reduction of a high order estimated model, rather than to directly estimate a low order one is intriguing. Using the basic

tools, more general situations can be investigated. Here we focus on general OE model structures. We assume that the low order model structure contains the true system, i.e., we look at the case of no undermodeling. This is somewhat simplified from the general case where undermodeled low order models are included, but necessary to complete the proof. In Tjärnström (2002) recent results on the undermodeling case are discussed.

Let the underlying system be given by

$$y(t) = G_0(q)u(t) + v(t) = \frac{B_0(q)}{F_0(q)}u(t) + v(t),$$

$$v(t) = H_0(q)e(t) \quad (53)$$

with the same assumptions on e and u as in (34). Parameterize two OE model structures $G(q, \theta)$ and $G(q, \eta)$ where $\dim \theta \geq \dim \eta$, i.e.,

$$\theta = (b_1 \dots b_{n_b} f_1 \dots f_{n_f})^T \quad (54)$$

$$\eta = (b_1 \dots b_{n_{b_0}} f_1 \dots f_{n_{f_0}})^T, \quad (55)$$

where $n_b \geq n_{b_0}$ and $n_f \geq n_{f_0}$. Furthermore, we assume the existence of some θ^* and a unique η^* such that

$$G(e^{i\omega}, \theta^*) = G(e^{i\omega}, \eta^*) = G_0(e^{i\omega}) \quad (56)$$

for almost all ω , and that no other parameterization with fewer parameters than $\dim \eta$ fulfill (56). Or in other words, the true model order is $[n_{b_0} n_{f_0}]$.

We now state the main theorem, which is proved in the next section.

Theorem 2 (Reduced model variance). *Assume that the true system is given by*

$$y(t) = G_0(q)u(t) + v(t),$$

where $v(t) = H_0(q)e(t)$, and $e(t)$ is white noise with variance λ and u is a stationary stochastic process independent of v , with known spectrum $\Phi_u(\omega)$. We assume that u and e have bounded fourth-order moments. Furthermore, we assume that $G(q, \theta)$ and $G(q, \eta)$ (with $\dim \theta \geq \dim \eta$) are two model structures of OE type (4) that both contain the true system $G_0(q)$, and that no other parameterization with fewer parameters than η contains the true system. Let $\hat{\theta}_N$ minimize the regularized loss function $W_N^\delta(\theta)$ (given by (9)) and $\hat{\eta}_N^r$ minimize

$$J(\eta, \hat{\theta}_N) = \int_{-\pi}^{\pi} |G(e^{i\omega}, \eta) - G(e^{i\omega}, \hat{\theta}_N)|^2 \Phi_u(\omega) d\omega.$$

Let the direct estimate be defined by

$$\hat{\eta}_N^d = \arg \min_{\eta} V_N(\eta),$$

where V_N is given by (5). Then the asymptotic variance of $\hat{\eta}_N^r$ tends to the variance of the direct estimate $\hat{\eta}_N^d$ as $\delta \rightarrow 0$, i.e.,

$$\lim_{\delta \rightarrow 0} \lim_{N \rightarrow \infty} N \text{Cov} \hat{\eta}_N^r = \lim_{N \rightarrow \infty} N \text{Cov} \hat{\eta}_N^d.$$

Moreover, we find that the reduced model meets the Cramér–Rao bound if the measurement noise is white and Gaussian.

8. Proof of the main result

In this section we present the proof of Theorem 2. First we prove the theorem in the case that the measurement noise is white, i.e., $H_0(q) = 1$. After that we prove the result for general H_0 .

Note from (54) and (55) that the parameters η form a subset of θ . This can be written as

$$S_0^T \theta = \eta, \quad (57)$$

where

$$S_0 = (I_1 \dots I_{n_{b_0}} I_{n_b+1} \dots I_{n_b+n_{f_0}}) \quad (58)$$

and I_j is the j th column of the $(n_b + n_f) \times (n_b + n_f)$ identity matrix.

The gradients of $\hat{y}(t, \theta)$ and $\hat{y}(t, \eta)$ equal (see (12))

$$\Psi(t, \theta) = \frac{d}{d\theta} G(q, \theta)u(t) = \frac{d}{d\theta} \frac{B(q, \theta)}{F(q, \theta)} u(t)$$

$$= \begin{pmatrix} q^{-n_{k_0}} \\ \vdots \\ q^{-n_{k_0}-n_b+1} \\ -q^{-1}G(q, \theta) \\ \vdots \\ -q^{-n_f}G(q, \theta) \end{pmatrix} \frac{1}{F(q, \theta)} u(t) \quad (59)$$

and

$$\Psi(t, \eta) = \frac{d}{d\eta} G(q, \eta)u(t) = \frac{d}{d\eta} \frac{B(q, \eta)}{F(q, \eta)} u(t)$$

$$= \begin{pmatrix} q^{-n_{k_0}} \\ \vdots \\ q^{-n_{k_0}-n_{b_0}+1} \\ -q^{-1}G(q, \eta) \\ \vdots \\ -q^{-n_{f_0}}G(q, \eta) \end{pmatrix} \frac{1}{F(q, \eta)} u(t). \quad (60)$$

By observing that

$$\frac{B(q, \theta^*)}{F(q, \theta^*)} = G_0(q), \quad (61)$$

we find that

$$B(q, \theta^*) = B_0(q)L(q) \quad \text{and} \quad F(q, \theta^*) = F_0(q)L(q). \quad (62)$$

Here $L(q)$ is a monic FIR filter of length $r + 1$ and

$$r = \min(n_b - n_{b_0}, n_f - n_{f_0}), \quad (63)$$

i.e.,

$$L(q) = 1 + l_1 q^{-1} + \dots + l_r q^{-r} = \sum_{k=0}^r l_k q^{-k}, \quad (64)$$

where we use the convention that $l_0 = 1$. We also obviously have that

$$\frac{B(q, \eta^*)}{F(q, \eta^*)} = G_0(q). \quad (65)$$

Putting (59), (61), and (62) together gives

$$\Psi(t, \theta^*) = \begin{pmatrix} q^{-n_{k_0}} \\ \vdots \\ q^{-n_{k_0} - n_{b_0} + 1} \\ -q^{-1} G_0(q) \\ \vdots \\ -q^{-n_f} G_0(q) \end{pmatrix} \frac{1}{L(q)F_0(q)} u(t). \quad (66)$$

In the same way we get from (60), and (65)

$$\Psi(t, \eta^*) = \begin{pmatrix} q^{-n_{k_0}} \\ \vdots \\ q^{-n_{k_0} - n_{b_0} + 1} \\ -q^{-1} G_0(q) \\ \vdots \\ -q^{-n_{f_0}} G_0(q) \end{pmatrix} \frac{1}{F_0(q)} u(t). \quad (67)$$

From these two expressions and utilizing (57) we get the important relation

$$\Psi(t, \eta^*) = S_0^T L(q) \Psi(t, \theta^*). \quad (68)$$

Let us now consider (28) with $W(\omega) = \Phi_u(\omega)$:

$$\begin{aligned} J(\eta, \hat{\theta}) &= \int_{-\pi}^{\pi} |G(e^{i\omega}, \eta) - G(e^{i\omega}, \hat{\theta})|^2 \Phi_u(\omega) d\omega \\ &= E[(G(q, \eta) - G(q, \hat{\theta}))u(t)]^2 \\ &= E\gamma^2(t, \eta, \hat{\theta}) \end{aligned} \quad (69)$$

with obvious definition of $\gamma^2(t, \eta, \hat{\theta})$. Note that $\hat{\theta}$ should be regarded as fixed (independent of u) in this expression and that

$$\gamma(t, \eta^*, \theta^*) = 0, \quad \forall t \quad (70)$$

according to (56). Define as before

$$\hat{\eta}^r = \arg \min_{\eta} J(\eta, \hat{\theta}). \quad (71)$$

From the discussion in Ljung (1999, Appendix 9.B) it follows that the difference between $E\hat{\eta}^r$ and η^* (defined by (33)) is “small”, i.e.,

$$|\eta^* - E\hat{\eta}^r| \leq \frac{C}{N} \quad (72)$$

for some constant C according to Ljung (1999, Eq. (9B.13)). So the limiting estimate of the two-step method (estimation and reduction) gives approximately the same limiting estimate as the direct estimation method.

In order to calculate the variance of the reduced order model we need to derive the expressions for $J''_{\eta\eta}(\eta^*, \theta^*)$ and $J''_{\eta\hat{\theta}}(\eta^*, \theta^*)$ from (69):

$$J'_{\eta}(\eta, \hat{\theta}) = E\Psi(t, \eta)\gamma(t, \eta, \hat{\theta}), \quad (73)$$

$$J''_{\eta\eta}(\eta, \hat{\theta}) = E\gamma(t, \eta, \hat{\theta}) \frac{d}{d\eta} \Psi(t, \eta) + E\Psi(t, \eta)\Psi^T(t, \eta), \quad (74)$$

$$J''_{\eta\hat{\theta}}(\eta, \hat{\theta}) = -E\Psi(t, \eta)\Psi^T(t, \hat{\theta}). \quad (75)$$

According to (70) the first term in (74) vanishes in (η^*, θ^*) . Evaluating the last two expressions at (η^*, θ^*) gives

$$\begin{aligned} J''_{\eta\eta}(\eta^*, \theta^*) &= J''_{\eta\eta}(\eta, \hat{\theta}) \Big|_{\eta=\eta^*, \hat{\theta}=\theta^*} \\ &= E\Psi(t, \eta^*)\Psi^T(t, \eta^*), \end{aligned} \quad (76)$$

$$J''_{\eta\hat{\theta}}(\eta^*, \theta^*) = -E\Psi(t, \eta^*)\Psi^T(t, \theta^*). \quad (77)$$

Next the covariance function of the gradient $\Psi(t, \theta^*)$ is defined as

$$\begin{aligned} R_{\theta}(\tau) &= E\Psi(t + \tau, \theta^*)\Psi^T(t, \theta^*) \\ &= E\Psi(t, \theta^*)\Psi^T(t - \tau, \theta^*) \end{aligned} \quad (78)$$

and similarly for $\Psi(t, \eta^*)$. This allows us to write

$$\begin{aligned} [E\Psi(t, \theta^*)\Psi^T(t, \theta^*) + \delta I]^{-1} &= (R_{\theta}(0) + \delta I)^{-1} \\ &= \tilde{R}_{\theta}^{-1}(0), \end{aligned} \quad (79)$$

where the last equality is the definition of $\tilde{R}_{\theta}^{-1}(0)$.

We continue by giving a lemma regarding rank deficient matrices.

Lemma 3. Let A be a $n \times n$ -dimensional positive semidefinite symmetric matrix of rank $m \leq n$. Define $\tilde{A} = A + \delta I$ with $\delta > 0$. Then the following holds:

- (i) $\tilde{A}^{-1}A = A\tilde{A}^{-1} = I - \delta\tilde{A}^{-1}$.
- (ii) $\lim_{\delta \rightarrow 0} \delta^{1+\varepsilon}\tilde{A}^{-1} = 0, \quad \varepsilon > 0$.

Proof. (i) $I = \tilde{A}^{-1}\tilde{A} = \tilde{A}^{-1}(A + \delta I) \Leftrightarrow \tilde{A}^{-1}A = I - \delta\tilde{A}^{-1}$. The other equality follows similarly.

(ii) Since A is symmetric it follows that

$$A = UDU^T \quad (80)$$

with $D = \text{diag}(d_1, \dots, d_m, 0, \dots, 0)$ and $UU^T = U^T U = I$.

Adding δI to both sides of (80) gives

$$A + \delta I = U(D + \delta I)U^T.$$

Inverting both sides gives (since $U^{-1} = U^T$)

$$\tilde{A}^{-1} = U(D + \delta I)^{-1}U^T.$$

Hence we get

$$\delta^{1+\varepsilon} \tilde{A}^{-1} = U \tilde{D} U^T,$$

$$\tilde{D} = \text{diag}\left(\frac{\delta^{1+\varepsilon}}{d_1 + \delta}, \dots, \frac{\delta^{1+\varepsilon}}{d_m + \delta}, \delta^\varepsilon, \dots, \delta^\varepsilon\right).$$

From this it follows that

$$\begin{aligned} \lim_{\delta \rightarrow 0} \delta^{1+\varepsilon} \tilde{A}^{-1} &= U \lim_{\delta \rightarrow 0} \tilde{D} U^T \\ &= U 0 U^T = 0, \quad \varepsilon > 0. \quad \square \end{aligned}$$

Before presenting the next lemma we extend the definition of S_0 in (58) to

$$S_k = (I_{k+1} \dots I_{k+n_b} I_{k+n_b+1} \dots I_{k+n_b+n_{f_0}}). \quad (81)$$

Lemma 4. Let $\Psi(t, \theta^*)$, $R_\theta(\tau)$, and $R_\eta(\tau)$ be given by (66) and (78). Then it holds that:

- (i) $\Psi^T(t - k, \theta^*) S_0 = \Psi^T(t, \theta^*) S_k$, $0 \leq k \leq r$.
- (ii) $R_\eta(\tau) = \sum_{m=0}^r \sum_{k=0}^r l_m l_k S_m^T R_\theta(\tau) S_k$.

Proof. (i) First, let $(\Psi)_j$ denote the j th element of the vector Ψ . Studying the j th, $1 \leq j \leq n_b - k$, element of $\Psi(t, \theta^*)$, where $0 \leq k \leq r$, gives

$$\begin{aligned} (\Psi^T(t - k, \theta^*))_j &= \left(q^{-n_{k_0} - k - j + 1} \frac{1}{L(q)F_0(q)} u(t) \right) \\ &= (\Psi^T(t, \theta^*))_{k+j}. \end{aligned}$$

Similarly for $n_b + 1 \leq j \leq n_b + n_{f_0} - k$ we get

$$\begin{aligned} (\Psi^T(t - k, \theta^*))_j &= \left(q^{-k-j+1} \frac{G_0(q)}{L(q)F_0(q)} u(t) \right) \\ &= (\Psi^T(t, \theta^*))_{k+j}. \end{aligned}$$

Now the multiplication $\Psi(t - k, \theta^*) S_0$ picks out the first n_{b_0} elements and elements with indices between $n_b + 1$ and $n_b + n_{f_0}$ from $\Psi(t - k, \theta^*)$, whereas $\Psi(t, \theta^*) S_k$ picks out elements shifted k steps away (relatively to S_0) from $\Psi(t, \theta^*)$. This means that we pick out exactly those elements corresponding to each other by the multiplication with S_0 and S_k .

(ii) This is proved using (68) and (i):

$$\begin{aligned} R_\eta(\tau) &= E\Psi(t, \eta^*)\Psi^T(t - \tau, \eta^*) \\ &= ES_0^T L(q)\Psi(t, \theta^*)L(q)\Psi^T(t - \tau, \theta^*)S_0 \\ &= ES_0^T \sum_{m=0}^r l_m q^{-m} \Psi(t, \theta^*) \sum_{n=0}^r l_n q^{-n} \Psi^T(t - \tau, \theta^*)S_0 \\ &= \sum_{m=0}^r \sum_{n=0}^r l_m l_n ES_0^T \Psi(t - m, \theta^*)\Psi^T(t - n - \tau, \theta^*)S_0 \\ &= \sum_{m=0}^r \sum_{n=0}^r l_m l_n ES_m^T \Psi(t, \theta^*)\Psi^T(t - \tau, \theta^*)S_n \\ &= \sum_{m=0}^r \sum_{n=0}^r l_m l_n S_m^T R_\theta(\tau) S_n. \quad \square \end{aligned}$$

We are now ready to prove Theorem 2 in the case of $H_0(q) = 1$. Estimation of the high order system $G(q, \theta)$ by minimizing $W_N^\delta(\theta)$ gives $\hat{\theta}$ with covariance

$$\begin{aligned} \text{Cov } \hat{\theta} &\approx \frac{\lambda}{N} [E\Psi(t, \theta^*)\Psi^T(t, \theta^*) + \delta I]^{-1} \\ &\quad \times [E\Psi(t, \theta^*)\Psi^T(t, \theta^*)] \\ &\quad \times [E\Psi(t, \theta^*)\Psi^T(t, \theta^*) + \delta I]^{-1} \end{aligned} \quad (82)$$

according to (16). Putting (31), (77), and (82) together we find that

$$\begin{aligned} \text{Cov } \hat{\eta}^r &\approx [E\Psi(t, \eta^*)\Psi^T(t, \eta^*)]^{-1} [E\Psi(t, \eta^*)\Psi^T(t, \eta^*)] \\ &\quad \times \frac{\lambda}{N} [E\Psi(t, \theta^*)\Psi^T(t, \theta^*) + \delta I]^{-1} \\ &\quad \times [E\Psi(t, \theta^*)\Psi^T(t, \theta^*)] \\ &\quad \times [E\Psi(t, \theta^*)\Psi^T(t, \theta^*) + \delta I]^{-1} \\ &\quad \times [E\Psi(t, \theta^*)\Psi^T(t, \eta^*)][E\Psi(t, \eta^*)\Psi^T(t, \eta^*)]^{-1} \end{aligned} \quad (83)$$

We would like to show that (83) tends to

$$\text{Cov } \hat{\eta}^d \approx \frac{\lambda}{N} [E\Psi(t, \eta^*)\Psi^T(t, \eta^*)]^{-1} \quad (84)$$

as $\delta \rightarrow 0$, which is the covariance $\hat{\eta}$ would have if it had been estimated directly from the data $\{u(t), y(t)\}_{t=1}^N$. This can equivalently be stated as

$$\begin{aligned} [E\Psi(t, \eta^*)\Psi^T(t, \eta^*)] &= \lim_{\delta \rightarrow 0} [E\Psi(t, \eta^*)\Psi^T(t, \theta^*)] \\ &\quad \times [E\Psi(t, \theta^*)\Psi^T(t, \theta^*) + \delta I]^{-1} \\ &\quad \times [E\Psi(t, \theta^*)\Psi^T(t, \theta^*)] \\ &\quad \times [E\Psi(t, \theta^*)\Psi^T(t, \theta^*) + \delta I]^{-1} \\ &\quad \times [E\Psi(t, \theta^*)\Psi^T(t, \eta^*)]. \end{aligned} \quad (85)$$

Using (68) we get

$$\begin{aligned}
 E\Psi(t, \eta^*)\Psi^T(t, \theta^*) &= ES_0^T L(q)\Psi(t, \theta^*)\Psi^T(t, \theta^*) \\
 &= \sum_{m=0}^r l_m ES_0^T \Psi(t-m, \theta^*)\Psi^T(t, \theta^*) \\
 &= \sum_{m=0}^r l_m ES_m^T \Psi(t, \theta^*)\Psi^T(t, \theta^*) \\
 &= \sum_{m=0}^r l_m S_m^T R_\theta(0),
 \end{aligned}$$

where we used Lemma 4(i). Plugging this into the right-hand side of (85) and using Lemma 3(i) gives

$$\begin{aligned}
 &[E\Psi(t, \eta^*)\Psi^T(t, \theta^*)][E\Psi(t, \theta^*)\Psi^T(t, \theta^*) + \delta I]^{-1} \\
 &\times [E\Psi(t, \theta^*)\Psi^T(t, \theta^*)][E\Psi(t, \theta^*)\Psi^T(t, \theta^*) + \delta I]^{-1} \\
 &\times [E\Psi(t, \theta^*)\Psi^T(t, \eta^*)] \\
 &= \left(\sum_{m=0}^r l_m S_m^T R_\theta(0) \right) \tilde{R}_\theta^{-1}(0) \\
 &\times R_\theta(0) \tilde{R}_\theta^{-1}(0) \left(\sum_{n=0}^r l_n R_\theta(0) S_n \right) \\
 &= \left(\sum_{m=0}^r l_m S_m^T (I - \delta \tilde{R}_\theta^{-1}(0)) \right) \\
 &\times R_\theta(0) \left(\sum_{n=0}^r l_n (I - \delta \tilde{R}_\theta^{-1}(0)) S_n \right) \\
 &= \left(\sum_{m=0}^r l_m S_m^T (R_\theta(0) - \delta (I - \delta \tilde{R}_\theta^{-1}(0))) \right) \\
 &\times \left(\sum_{n=0}^r l_n (I - \delta \tilde{R}_\theta^{-1}(0)) S_n \right) \\
 &= \sum_{m=0}^r \sum_{n=0}^r l_m l_n S_m^T (R_\theta(0) - \delta (I - \delta \tilde{R}_\theta^{-1}(0))) \\
 &\quad - \delta (I - \delta \tilde{R}_\theta^{-1}(0))^2 S_n \\
 &= \sum_{m=0}^r \sum_{n=0}^r l_m l_n S_m^T (R_\theta(0) - 2\delta I + 3\delta^2 \tilde{R}_\theta^{-1}(0)) \\
 &\quad - \delta^3 (\tilde{R}_\theta^{-1}(0))^2 S_n.
 \end{aligned}$$

Letting $\delta \rightarrow 0$, the second term vanishes and last two sums vanish according to Lemma 3(ii). Moreover, the first term equals $R_\eta(0) = E\Psi(t, \eta^*)\Psi^T(t, \eta^*)$ according to Lemma 4(ii), and the result follows.

Since the direct estimate meets the Cramér–Rao bound if the noise is white and Gaussian, we get that the reduced model also meets the Cramér–Rao bound in this case.

Before presenting the proof of the theorem in the general non-white measurement noise case, we need to state another lemma.

Lemma 5. For $R_\theta(\tau)$ defined by (78) and (66) and $\tilde{R}_\theta^{-1}(0)$ defined by (79) it holds that

$$\lim_{\delta \rightarrow 0} \delta \tilde{R}_\theta^{-1}(0) R_\theta(\tau) = 0.$$

Proof. Let

$$x(t) = \frac{1}{L(q)F_0(q)} u(t),$$

$$\tilde{x}(t) = -G_0(q)x(t),$$

then we can rewrite (66) as

$$\begin{aligned}
 \Psi(t, \theta^*) &= \begin{pmatrix} q^{-n_{k_0}} \\ \vdots \\ q^{-n_{k_0} - n_b + 1} \\ -q^{-1} G_0(q) \\ \vdots \\ -q^{-n_f} G_0(q) \end{pmatrix} \frac{1}{L(q)F_0(q)} u(t) \\
 &= \begin{pmatrix} x(t - n_{k_0}) \\ \vdots \\ x(t - n_{k_0} - n_b + 1) \\ \tilde{x}(t - 1) \\ \vdots \\ \tilde{x}(t - n_f) \end{pmatrix}.
 \end{aligned}$$

Since $G_0(q) = B_0(q)/F_0(q)$ we get

$$B_0(q)x(t) + F_0(q)\tilde{x}(t) = 0$$

or in matrix notation

$$(b_1 \dots b_{n_{b_0}} \ 1 \ f_1 \dots f_{n_{f_0}}) \begin{pmatrix} x(t - n_{k_0} - 1) \\ \vdots \\ x(t - n_{k_0} - n_{b_0}) \\ \tilde{x}(t - 1) \\ \tilde{x}(t - 2) \\ \vdots \\ \tilde{x}(t - n_{f_0} - 1) \end{pmatrix} = 0, \quad \forall t. \quad (86)$$

This can be expressed in terms of the gradient $\Psi(t, \theta^*)$ as

$$0 = \begin{pmatrix} 0, b_1, \dots, b_{n_{b_0}}, \underbrace{0, \dots, 0}_{n_b - n_{b_0} - 1}, 1, f_1, \dots, f_{n_{f_0}}, \underbrace{0, \dots, 0}_{n_f - n_{f_0} - 1} \end{pmatrix} \Psi(t, \theta^*) \quad (87)$$

i.e., w_1 is orthogonal to the gradient. Moreover, since we know that the rank deficiency of $R_\theta(0)$ equals r (see (63)) we realize that it is possible to construct a total of r time-independent vectors, $w_1, \dots, w_r$, that are orthogonal to $\Psi(t, \theta^*)$ from the relation (86). These have the same structure as w_1 in (87), but the non-zero entries are shifted “downwards”, e.g.,

$$w_r = (0, \dots, 0, b_1, \dots, b_{n_{b_0}}, 0, \dots, 0, 1, f_1, \dots, f_{n_{f_0}}).$$

Since $w_1, \dots, w_r$ are orthogonal to $\Psi(t, \theta^*)$ it follows that they are also eigenvectors to $R_\theta(\tau)$ since

$$w_k R_\theta(\tau) = w_k E \Psi(t, \theta^*) \Psi^T(t - \tau, \theta^*)$$

$$= E 0 \Psi^T(t - \tau, \theta^*) = 0, \quad k = 1, \dots, r.$$

From this it follows that the singular value decomposition (SVD) of $R_\theta(\tau)$ is of the form

$$R_\theta(\tau) = (U_{1,\tau} \ U_2) \begin{pmatrix} \Sigma_\tau & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_{1,\tau}^T \\ V_2^T \end{pmatrix},$$

where

$$U_2 = (w_1 \ \dots \ w_r),$$

$$V_2 = (w_1 \ \dots \ w_r),$$

$$\Sigma_\tau = \text{diag}(\sigma_{1,\tau}, \dots, \sigma_{n_f + n_b - r, \tau}),$$

$$\sigma_{k,\tau} \geq 0, \quad k = 1, \dots, n_b + n_f - r,$$

$$\sigma_{k,0} > 0, \quad k = 1, \dots, n_b + n_f - r. \quad (88)$$

Here the subindex τ is included to indicate a possible dependency of τ . Note the strict inequality for $\sigma_{k,0}$. Now since U_2 and V_2 are independent of τ it follows that:

$$V_2^T U_{1,\tau} = V_{1,\tau}^T U_2 = 0, \quad V_2^T U_2 = I.$$

Moreover, we have that the SVD of $\tilde{R}_\theta^{-1}(0) = (R_\theta(0) + \delta I)^{-1}$ equals

$$\tilde{R}_\theta^{-1}(0) = (U_{1,0} \ U_2) \begin{pmatrix} (\delta I + \Sigma_0)^{-1} & 0 \\ 0 & \frac{1}{\delta} I \end{pmatrix} \begin{pmatrix} V_{1,0}^T \\ V_2^T \end{pmatrix}.$$

Putting all of the above together we get

$$\begin{aligned} \delta \tilde{R}_\theta^{-1}(0) R_\theta(\tau) &= (U_{1,0} \ U_2) \begin{pmatrix} \delta(\delta I + \Sigma_0)^{-1} & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} V_{1,0}^T \\ V_2^T \end{pmatrix} \\ &\times (U_{1,\tau} \ U_2) \begin{pmatrix} \Sigma_\tau & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_{1,\tau}^T \\ V_2^T \end{pmatrix} \\ &= (\delta U_{1,0}(\delta I + \Sigma_0)^{-1} \ U_2) \\ &\begin{pmatrix} V_{1,0}^T U_{1,\tau} & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \Sigma_\tau V_{1,\tau}^T \\ 0 \end{pmatrix} \\ &= \delta U_{1,0}(\delta I + \Sigma_0)^{-1} V_{1,0}^T U_{1,\tau} \Sigma_\tau V_{1,\tau}^T \rightarrow 0, \\ &\text{as } \delta \rightarrow 0, \end{aligned}$$

where the last statement follows from (88). $\square$

We are now ready to continue with the proof of Theorem 2 in the case of non-white measurement noise, i.e., $H_0(q) \neq 1$. From (11) we know that the covariance of the direct estimate, $\hat{\eta}^d$, equals

$$\begin{aligned} \text{Cov } \hat{\eta}^d &\approx \frac{\lambda}{N} [E \Psi(t, \eta^*) \Psi^T(t, \eta^*)]^{-1} \\ &\times [E \tilde{\Psi}(t, \eta^*) \tilde{\Psi}^T(t, \eta^*)] [E \Psi(t, \eta^*) \Psi^T(t, \eta^*)]^{-1} \end{aligned}$$

and the covariance of the L_2 reduced estimate, $\hat{\eta}^r$, equals (see (83))

$$\begin{aligned} \text{Cov } \hat{\eta}^r &\approx \frac{\lambda}{N} [E \Psi(t, \eta^*) \Psi^T(t, \eta^*)]^{-1} [E \Psi(t, \eta^*) \Psi^T(t, \theta^*)] \\ &\times [E \Psi(t, \theta^*) \Psi^T(t, \theta^*) + \delta I]^{-1} [E \tilde{\Psi}(t, \theta^*) \tilde{\Psi}^T(t, \theta^*)] \\ &\times [E \Psi(t, \theta^*) \Psi^T(t, \theta^*) + \delta I]^{-1} \\ &\times [E \Psi(t, \theta^*) \Psi^T(t, \eta^*)] [E \Psi(t, \eta^*) \Psi^T(t, \eta^*)]^{-1}. \end{aligned}$$

Showing equality between these two expressions as $\delta \rightarrow 0$ is the same as showing that

$$\begin{aligned} E \tilde{\Psi}(t, \eta^*) \tilde{\Psi}^T(t, \eta^*) \\ &= \lim_{\delta \rightarrow 0} [E \Psi(t, \eta^*) \Psi^T(t, \theta^*)] [E \Psi(t, \theta^*) \Psi^T(t, \theta^*) + \delta I]^{-1} \\ &\times [E \tilde{\Psi}(t, \theta^*) \tilde{\Psi}^T(t, \theta^*)] \\ &\times [E \Psi(t, \theta^*) \Psi^T(t, \theta^*) + \delta I]^{-1} [E \Psi(t, \theta^*) \Psi^T(t, \eta^*)]. \end{aligned}$$

Expressing the left- and right-hand side of this equation in terms of the covariance function gives

$$\begin{aligned} & \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} h_k h_l R_{\eta}(k-l) \\ &= \lim_{\delta \rightarrow 0} \sum_{m=0}^r l_m S_m^T R_{\theta}(0) \tilde{R}_{\theta}^{-1}(0) \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} h_k h_l R_{\theta}(k-l) \\ & \quad \times \tilde{R}_{\theta}^{-1}(0) \sum_{n=0}^r l_n R_{\theta}(0) S_n. \end{aligned} \quad (89)$$

Continuing to expand the right-hand side of (89) using Lemma 3(i) we get

$$\begin{aligned} \text{RHS} &= \sum_{m=0}^r l_m S_m^T (I - \delta \tilde{R}_{\theta}^{-1}(0)) \\ & \quad \times \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} h_k h_l R_{\theta}(k-l) \sum_{n=0}^r l_n (I - \delta \tilde{R}_{\theta}^{-1}(0)) S_n \\ &= \sum_{m=0}^r \sum_{n=0}^r l_n l_m S_m^T \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} h_k h_l R_{\theta}(k-l) S_n \\ & \quad - \delta \sum_{m=0}^r \sum_{n=0}^r l_n l_m S_m^T \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} h_k h_l \tilde{R}_{\theta}^{-1}(0) R_{\theta}(k-l) \\ & \quad - \delta \sum_{m=0}^r \sum_{n=0}^r l_n l_m \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} h_k h_l R_{\theta}(k-l) \tilde{R}_{\theta}^{-1}(0) S_n \\ & \quad + \delta^2 \sum_{m=0}^r \sum_{n=0}^r l_n l_m S_m^T \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} h_k h_l \tilde{R}_{\theta}^{-1}(0) \\ & \quad \times R_{\theta}(k-l) \tilde{R}_{\theta}^{-1}(0) S_n. \end{aligned}$$

Here the second and third term tend to zero as $\delta \rightarrow 0$ due to Lemma 5. The fourth term also tends to zero since $\delta \tilde{R}_{\theta}^{-1}(0)$ is bounded for small δ (see the proof of Lemma 3(ii)). In short

$$\begin{aligned} \lim_{\delta \rightarrow 0} \text{RHS} &= \sum_{m=0}^r \sum_{n=0}^r l_n l_m S_m^T \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} h_k h_l R_{\theta}(k-l) S_n \\ &= \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} h_k h_l \sum_{m=0}^r \sum_{n=0}^r l_n l_m S_m^T R_{\theta}(k-l) S_n \\ &= \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} h_k h_l R_{\eta}(k-l), \end{aligned}$$

where the last equality follows from Lemma 4(ii). Looking back at (89) we see that theorem is proved.

9. Example

To illustrate the results from the previous section we give a simple simulation example. The true system is given by the following OE-structure:

$$y(t) = \frac{B_0(q)}{F_0(q)} u(t) + v(t),$$

$$F_0(q) = 1 - 0.7q^{-1} + 0.52q^{-2} - 0.092q^{-3} - 0.1904q^{-4},$$

$$B_0(q) = 2q^{-1} - q^{-2}. \quad (90)$$

The system is estimated using $N = 1000$ input–output data. Different noise and input colors are used. A total of four different evaluations of the L_2 model reduction scheme are presented:

- (1) white input and white noise, i.e., $u(t) = w_1(t)$, $v(t) = w_2(t)$,
- (2) colored input and white noise, i.e., $u(t) = T_u(q)w_1(t)$, $v(t) = w_2(t)$,
- (3) white input and colored noise, i.e., $u(t) = w_1(t)$, $v(t) = T_v(q)w_2(t)$,
- (4) colored input and colored noise, i.e., $u(t) = T_u(q)w_1(t)$, $v(t) = T_v(q)w_2(t)$.

Here $w_1(t)$ and $w_2(t)$ are white Gaussian processes with variance 1, and $T_u(q)$ and $T_v(q)$ are given by

$$T_u(q) = \frac{0.5}{1 - 1.2q^{-1} + 0.7q^{-2}}, \quad (91)$$

$$T_v(q) = \frac{0.9}{1 + 0.5q^{-1}}. \quad (92)$$

The bode diagrams of T_u and T_v are displayed in Fig. 1 together with the true system, $G_0 = B_0/F_0$.

The evaluation is performed according to the following. An OE model of order 6 is estimated in each case, giving $\hat{\theta}$, and reduced in L_2 norm to the correct order, giving

Fig. 1. Bode diagram of true system, $G_0(e^{j\omega})$ (solid), noise color $T_v(e^{j\omega})$ (dashed), and input color, $T_u(e^{j\omega})$ (dash-dotted).

Fig. 2. Results from 1000 simulations. Y-axis: loss function on validation data using L_2 model reduction. X-axis: loss function on validation data using direct estimation. Every cross represents one simulation. The solid line is $y=x$, and the dashed line is the mean square estimate of a line $y = \alpha x + \beta$ from the simulations. The circle represents the mean of all simulations. (a) White noise and white input. (b) White noise and colored input. (c) Colored noise and white input. (d) Colored noise and colored input.

$\hat{\eta}^r$. The reduced model is estimated in the following way. A new input sequence, $u_s(t)$ of length $10N$ is generated with spectrum $\Phi_u(\omega)$. Then a new output sequence is simulated according to $y_s(t) = G(q, \theta)u_s(t)$. Using these input–output data the low order model $\hat{\eta}^r$ is estimated. This procedure (of simulating new data) slightly increases the variance of $\hat{\eta}^r$ (compared to performing the minimization of (28)), but this error is of order $\sim 1/10N$ and can therefore be neglected. Another OE model, $\hat{\eta}^d$, of correct order is also estimated directly from the original data. In order to avoid local minima, the estimation algorithm is initiated at the optimum.

To illustrate the results in this contribution graphically, we chose to “project” the six-dimensional co-

variance matrices down to one-dimensional scalars, namely the variance of the prediction error for each model. That is, for the two models (the directly estimate and the reduced one) the loss functions on validation data is calculated. The result of this is plotted in Fig. 2. This is repeated 1000 times (giving one cross in each figure for every estimate). Figs. 2(a)–(d) correspond to items 1–4 in the list above, respectively.

From the results presented in Fig. 2, we see that the loss function on validation data follows the straight line $y = x$ very accurately in all four cases. This gives us a good confirmation on the results in Section 7, i.e., that variance of the reduced model equals the variance of the directly estimated one (asymptotically).

10. Conclusions

The main result of this paper is that applying L_2 model reduction to an identified model gives essentially optimal reduction of the variance of that model. In particular, it follows from our results that:

- If the true system is of a certain order n , and a higher order model of output error type is first estimated and then L_2 reduced in the $\Phi_u(\omega)$ norm to order n , then the variance of that model is the same as if an n th-order output error model is directly estimated from data.
- If a high order FIR model is estimated from data in a structure that correctly can describe the system, and this model is L_2 reduced to a lower order, then we in general obtain a model with *smaller variance* than a directly estimated low order FIR model.

This implies that high order output error modeling followed by L_2 model reduction makes optimal use of the information contents in data if the measurement noise is white and Gaussian and the true system is of OE type. Then both the direct and the reduced estimates meet the Cramér–Rao lower bound. This cannot be outperformed by other model reduction techniques.

All the results are derived taking expectations over both u and e . Different results are obtained if expectation is taken only over e . Note also that the results in this paper are based on that model reduction is performed in the L_2 norm weighted in the *true* input spectrum. The results may be quite different if the weighting is chosen as an *estimate* of the input spectrum.

In general the low order model has some bias. Having arrived at the simple model by model reduction of a high order model gives an estimate of the bias as the difference between the two models. At the same time the variance of the low order model is kept small according to the results in this paper for FIR models and according to Tjörnström (2002) for general linear output error models. This gives advantages over a directly estimated low order model, which has higher variance, and a bias error which requires special measures to assess.

References

- Glover, K. (1984). All optimal hankel-norm approximations of linear multivariable systems and their L^∞ -error bounds. *International Journal of Control*, 39(6), 1115–1193.
- Hsia, T. C. (1977). *Identification: Least squares methods*. Lexington, MA: Lexington Books.
- Kabaila, P. V. (1983). On output-error methods for system identification. *IEEE Transactions on Automatic Control*, 28, 12–23.
- Ljung, L. (1985). Asymptotic variance expressions for identified black-box transfer function models. *IEEE Transactions on Automatic Control*, 30(9), 834–844.
- Ljung, L. (1999). *System identification: Theory for the user* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.
- Moore, B. (1981). Principal component analysis in linear systems: Controllability, observability and model reduction. *IEEE Transactions on Automatic Control*, 26, 17–31.
- Porat, B. (1986). On the estimation of the parameters of vector Gaussian processes from sample covariances. In *Proceedings of the 25th Conference on Decision and Control*, Athens, Greece (pp. 2002–2005).
- Porat, B., & Friedlander, B. (1985). Asymptotic accuracy of ARMA parameter estimation methods based on sample covariances. In *Preprints 7th IFAC symposium on identification and system parameter estimation*, York, UK (pp. 963–968).
- Söderström, T., Stoica, P., & Friedlander, B. (1991). An indirect prediction error method for system identification. *Automatica*, 27, 183–188.
- Spanos, J. T., Milman, M. H., & Mingori, D. L. (1992). A new algorithm for L_2 optimal model reduction. *Automatica*, 28(5), 897–909.
- Stoica, P., & Söderström, T. (1989). On reparameterization of loss functions used in estimation and the invariance principle. *Signal Processing*, 17, 383–387.
- Tjörnström, F. (2002). Variance aspects of L_2 model reduction when undermodeling—the output error case. In *Proceedings of the 15th IFAC World Congress*, Barcelona, Spain.
- Tjörnström, F., & Ljung, L. (2001). Variance properties of a two-step ARX estimation procedure. In *Proceedings of the European control conference*, Porto, Portugal (pp. 1840–1845).
- Wahlberg, B. (1987). *On the identification and approximation of linear systems*. Ph.D. thesis 163, Department of Electrical Engineering, Linköping University.
- Wahlberg, B. (1989). Model reduction of high order estimated models: The asymptotic ML approach. *International Journal of Control*, 49(1), 169–192.
- Zhu, Y. -C. (1989). Black-box identification of MIMO transfer functions: Asymptotic properties of prediction error models. *International Journal of Adaptive Control and Signal Processing*, 3, 357–373.
- Zhu, Y., & Backx, T. (1993). *Identification of multivariable industrial processes for simulation, diagnosis and control*. Berlin: Springer.

Fredrik Tjörnström was born in Örnsköldsvik, Sweden in 1973. He received the M.Sc. degree in Applied Physics and Electrical Engineering in 1997 and the Ph.D. degree in automatic control in 2002, both from University of Linköping. Currently he is a research associate in the Automatic Control group, Department of Electrical Engineering, Linköping University, Linköping, Sweden. His research topics include system identification and its connection to model reduction, bootstrap techniques and

identification of nonlinear systems.

Lennart Ljung received his Ph.D. in Automatic Control from Lund Institute of Technology in 1974. Since 1976 he is Professor of the chair of Automatic Control in Linköping, Sweden, and is currently Director of the Competence Center “Information Systems for Industrial Control and Supervision” (ISIS). He has held visiting positions at Stanford and MIT and has written several books on System Identification and Estimation. He is an IEEE Fellow and an IFAC Advisor as well as a member of the Royal Swedish Academy of Sciences (KVA), a member of the Royal Swedish Academy of Engineering Sciences (IVA), and an Honorary Member of the Hungarian Academy of Engineering. He has received honorary doctorates from the Baltic State Technical University in St Petersburg, and from Uppsala University. In 2002 he received the Quazza Medal from IFAC.