Optimal model order reduction with the Steiglitz-McBride method for open loop data ⋆

Niklas Everitt [a], Miguel Galrinho [a], Håkan Hjalmarsson [a]

[a] *ACCESS Linnaeus Center, School of Electrical Engineering, KTH - Royal Institute of Technology, Sweden*

Abstract

In system identification, it is often difficult to find a physical intuition to choose a noise model structure. The importance of this choice is that, for the prediction error method (PEM) to provide asymptotically efficient estimates, the model orders must be chosen according to the true system. However, if only the plant estimates are of interest and the experiment is performed in open loop, the noise model may be over-parameterized without affecting the asymptotic properties of the plant. The limitation is that, as PEM suffers in general from non-convexity, estimating an unnecessarily large number of parameters will increase the chances of getting trapped in local minima. To avoid this, a high order ARX model can first be estimated by least squares, providing non-parametric estimates of the plant and noise model. Then, model order reduction can be used to obtain a parametric model of the plant only. We review existing methods to perform this, pointing out limitations and connections between them. Then, we propose a method that connects favorable properties from the previously reviewed approaches. We show that the proposed method provides asymptotically efficient estimates of the plant with open loop data. Finally, we perform a simulation study, which suggests that the proposed method is competitive with PEM and other similar methods.

Key words: System identification, Steiglitz-McBride, High order ARX-modeling, maximum likelihood.

1 Introduction

The prediction error method (PEM) is a well-know approach for estimation of parametric models [5]. If the model orders are chosen correctly, a quadratic cost function provides asymptotically efficient estimates when the noise is Gaussian. The drawback is that, in general, PEM requires solving a non-convex optimization problem, which can converge to minima that are only local. Alternative methods, such as subspace [15] or instrumental variable [10] methods, do not suffer from non-convexity, being useful to provide initialization points for PEM.

Other methods first estimate a high order (non-parametric) model. In general, this is an ARX model, for which the global minimum of the prediction error cost function can be found by least squares. Because it is high order, this estimate will have high variance. However, it can be reduced to a parametric model description of low order. If the model reduction step is performed according to an exact maximum likelihood (ML) criterion, the low order estimates are asymptotically efficient [16]. This approach still requires, in general, solving a non-convex optimization problem.

Another possibility to perform model order reduction from a high order non-parametric model is with the weighted null-space fitting (WNSF) method [2]. Although it can be motivated by an exact ML criterion, this criterion is not minimized explicitly. Rather, it is interpreted as a weighted least squares problem by fixing the parameters in the weighting.

One problem with estimation of parametric models is the choice of model orders. If this choice can sometimes be based on physical intuition for the plant, the noise model order is usually a more abstract concept. This has been observed in [8], where a frequency domain method is proposed to estimate a parametric model of the plant and a non-parametric noise model. Because this approach does not require a noise model order selection, it can be seen as more user friendly.

If the data are obtained in open loop, the asymptotic

⋆ This work was supported by the Swedish Research Council under contract 2015-05285 and by the European Research Council under the advanced grant LEARN, contract 267381. The material in this paper was not presented at any conference.

Email addresses: everitt@kth.se (Niklas Everitt), galrinho@kth.se (Miguel Galrinho), hjalmars@kth.se (Håkan Hjalmarsson).

properties of the plant and noise model estimates obtained with PEM are uncorrelated, if the two transfer functions are independently parametrized [5]. Therefore, when a parametric noise model estimate is not of interest, asymptotically efficient estimates of the plant can be obtained as long as the noise model order is chosen high enough for the system to be in the model set. The limitation of choosing the noise model order arbitrarily large with PEM is that, as more parameters are estimated, the complexity of the problem increases, and it is more difficult to find the global minimum.

However, if a non-parametric ARX model is estimated, there are no issues with local minima, while the order is arbitrarily large. Then, for the model reduction step, an approximate asymptotic ML criterion allows separating the estimation of the plant and noise model [16]. This allows obtaining asymptotically efficient estimates of the plant in open loop without the high order structure of the noise model affecting the difficulty of the problem. Nevertheless, the model reduction step still requires solving a non-convex optimization problem. The ASYM method [17] is based on this approach.

Another approach that does not require a parametric noise model is the BJSM method [18]. This method uses a non-parametric ARX model estimate to pre-filter the input and output data, creating a pre-filtered data set for which the output noise is approximately white. Then, a noise model is no longer required when estimating the plant based on the pre-filtered data set. Instead of explicitly minimizing a non-convex function, BJSM applies the Steiglitz-McBride to the pre-filtered data set. In [18], it is shown that this procedure is asymptotically efficient in open loop. However, there are two limitations. First, even if the true noise model is known exactly, a non-parametric estimate is still required to achieve efficiency. Second, although the method does not apply local non-linear optimization techniques, the number of Steiglitz-McBride iterations needs to tend to infinity to obtain a consistent estimate.

Our contributions are the following. First, we make a connection between ASYM and BJSM, and propose a method—termed Model Order Reduction Steiglitz-McBride (MORSM)—connecting ideas from both. Second, we show that MORSM is asymptotically efficient in open loop with one iteration. Third, we perform a simulation study, where we observe that MORSM has better finite sample convergence properties than BJSM, and that it is a viable alternative to PEM.

2 Preliminaries

Assumption 2.1 (True system) *The system has scalar input u_t, scalar output y_t and is subject to scalar noise e_t. The relationship between these signals is given by*

$$y_t = G^\circ(q)u_t + H^\circ(q)e_t, \tag{1}$$

where $G^\circ(q)$ and $H^\circ(q)$ are rational functions in the time shift operator q^{-1} ($q^{-1}x_t := x_{t-1}$) according to

$$G^\circ(q) = \frac{L^\circ(q)}{F^\circ(q)} = \frac{l_1^\circ q^{-1} + \cdots + l_{m_l}^\circ q^{-m_l^\circ}}{1 + f_1^\circ q^{-1} + \cdots + f_{m_f}^\circ q^{-m_f^\circ}},$$
$$H^\circ(q) = \frac{C^\circ(q)}{D^\circ(q)} = \frac{1 + c_1^\circ q^{-1} + \cdots + c_{m_c}^\circ q^{-m_c^\circ}}{1 + d_1^\circ q^{-1} + \cdots + d_{m_d}^\circ q^{-m_d^\circ}}.$$

The transfer functions G°, H°, and $1/H^\circ$ are assumed to be stable. The polynomials L° and F°—as well as C° and D°—do not share common factors.

Let the input sequence $\{u_t\}$ be a realization of a stochastic process generated by a random sequence $\{w_t\}$. Also, let $\mathcal{F}_{t-1}$ be the σ-algebra generated by $\{e_s, w_s, s \leq t-1\}$. Then, the following assumption applies for the input signal.

Assumption 2.2 (Input) *The sequence $\{u_t\}$ is defined by*

$$u_t = F_u(q)w_t,$$

where $F_u(q)$ is a stable and inversely stable finite dimensional filter, where $\{w_t\}$ is independent of $\{e_t\}$, satisfying

$$\mathrm{E}\left[w_t|\mathcal{F}_{t-1}\right] = 0, \quad \mathrm{E}\left[w_t^2|\mathcal{F}_{t-1}\right] = \sigma_\circ^2, \quad |w_t| \leq C, \forall t$$

for some finite positive finite constant C.

Assumption 2.2 implies that the system is operating in open loop. Also, F_u can be interpreted as the stable minimum phase spectral factor of the input spectrum.

For the noise, the following assumption applies.

Assumption 2.3 (Noise) *$\{e_t\}$ is a stochastic process that satisfies*

$$\mathrm{E}\left[e_t|\mathcal{F}_{t-1}\right] = 0, \quad \mathrm{E}\left[e_t^2|\mathcal{F}_{t-1}\right] = \sigma_\circ^2, \quad |e_t|^{10} \leq C, \forall t$$

for some positive finite constant C.

3 The Prediction Error Method

The idea of the prediction error method (PEM) is to minimize a cost function of the prediction errors. In this section, we discuss how PEM can be used to estimate a model of the system (2.1). First, we consider a Box-Jenkins (BJ) model, and then a high order ARX model.

3.1 Box-Jenkins model

In a Box-Jenkins model, $G(q)$ and $H(q)$ are rational transfer functions parameterized independently, according to

$$y_t = G(q,\theta)u_t + H(q,\alpha)e_t, \tag{2}$$

where

$$G(q,\theta) = \frac{L(q,\theta)}{F(q,\theta)} = \frac{l_1 q^{-1} + \cdots + l_{m_l} q^{-m_l}}{1 + f_1 q^{-1} + \cdots + f_{m_f} q^{-m_f}},$$
$$H(q,\alpha) = \frac{C(q,\alpha)}{D(q,\alpha)} = \frac{1 + c_1 q^{-1} + \cdots + c_{m_c} q^{-m_c}}{1 + d_1 q^{-1} + \cdots + d_{m_d} q^{-m_d}},$$

and

$$\theta = \begin{bmatrix} f_1 & \dots & f_{m_f} & l_1 & \dots & l_{m_l} \end{bmatrix}^\top, \tag{3}$$
$$\alpha = \begin{bmatrix} c_1 & \dots & c_{m_c} & d_1 & \dots & d_{m_d} \end{bmatrix}^\top. \tag{4}$$

We assume that $H^\circ(q)$ is in the model set defined by $H(q,\alpha)$ (i.e., $m_c \geq m_c^\circ$ and $m_d \geq m_d^\circ$). Moreover, the order of the polynomials of $G^\circ(q)$ are assumed to be known (i.e., $m_f = m_f^\circ$ and $m_l = m_l^\circ$). For simplicity of notation only, we also assume that $m := m_f = m_l$.

The one step ahead prediction errors of the BJ model (2) are given by

$$\varepsilon_t(\theta,\alpha) = \frac{D(q,\alpha)}{C(q,\alpha)}\left[y_t - \frac{L(q,\theta)}{F(q,\theta)}u_t\right].$$

The parameter estimates using PEM with a quadratic cost function are determined by minimizing the loss function

$$V_N(\theta,\alpha) = \frac{1}{N}\sum_{t=1}^{N}\varepsilon_t^2(\theta,\alpha), \tag{5}$$

where N is the number of data samples. We denote by $\hat{\theta}_N^{\mathrm{PEM}}$ the estimate of θ obtained by minimizing (5). Moreover, $\theta_\circ$ corresponds to the vector θ evaluated at the coefficients of $F^\circ(q)$ and $L^\circ(q)$.

Since the system operates in open loop (Assumption 2.2), it is well known that, when PEM is applied to the model (2), the asymptotic covariance matrix of the parameter estimate $\hat{\theta}_N^{\mathrm{PEM}}$ is given by [5]

$$\lim_{N\to\infty} N\mathrm{E}\left[(\hat{\theta}_N^{\mathrm{PEM}} - \theta_\circ)(\hat{\theta}_N^{\mathrm{PEM}} - \theta_\circ)^\top\right] = \sigma_\circ^2 M_{\mathrm{CR}}^{-1},$$

where (we omit the argument of the transfer functions for brevity)

$$M_{\mathrm{CR}} = \frac{1}{2\pi\sigma_\circ^2}\int_{-\pi}^{\pi}\begin{bmatrix} -\frac{G^\circ}{F^\circ H^\circ}\Gamma_m \\ \frac{1}{F^\circ H^\circ}\Gamma_m \end{bmatrix}\begin{bmatrix} -\frac{G^\circ}{F^\circ H^\circ}\Gamma_m \\ \frac{1}{F^\circ H^\circ}\Gamma_m \end{bmatrix}^* \Phi_u \,\mathrm{d}\omega,$$

with $\Gamma_m(q) = \begin{bmatrix} q^{-1} & \dots & q^{-m} \end{bmatrix}^\top$ and Φ_u the spectrum of the input $\{u_t\}$.

When $\{e_t\}$ is Gaussian, PEM with a quadratic cost function is asymptotically efficient, meaning that M_{CR}^{-1} corresponds to the Cramér-Rao lower bound—the smallest possible asymptotic covariance matrix for a consistent estimator [5]. Again, we recall that only the orders of $G^\circ(q)$ need to be chosen correctly to achieve efficiency, while $H(q,\alpha)$ only needs to include $H^\circ(q)$. Thus, if only a model for $G^\circ(q)$ is of interest, and the order of $H^\circ(q)$ is unknown, m_c and m_d can be let grow to infinity (guaranteeing that $H^\circ(q)$ is in the model set) without asymptotically affecting the estimate of θ.

An important remark is that minimizing the loss function (5) is a non-convex optimization problem. Therefore, a good initialization point is required to converge to the global minimum. For Box-Jenkins models, an initialization point that is sufficiently close to the global minimum is particularly challenging to obtain. Moreover, the problem becomes yet more challenging if we want to let the order of the noise model $H(q,\alpha)$ be arbitrarily large, as PEM will have increased problems with local minima.

3.2 High order ARX model

To circumvent the limitations of solving a non-convex optimization problem, we consider the following approach. Note that the system (1) can be represented as

$$A^\circ(q)y_t = B^\circ(q)u_t + e_t, \tag{6}$$

where

$$A^\circ(q) := \frac{1}{H^\circ(q)} =: 1 + \sum_{k=1}^{\infty} a_k^\circ q^{-k},$$
$$B^\circ(q) := \frac{G^\circ(q)}{H^\circ(q)} =: \sum_{k=1}^{\infty} b_k^\circ q^{-k}$$

are stable transfer functions (by Assumption 2.1).

Consider also the ARX model

$$A(q,\eta^n)y_t = B(q,\eta^n)u_t + e_t,$$

where

$$A(q,\eta^n) = 1 + \sum_{k=1}^{n} a_k q^{-k}, \quad B(q,\eta^n) = \sum_{k=1}^{n} b_k q^{-k}, \tag{7}$$

and

$$\eta^n = \begin{bmatrix} a_1 & \dots & a_n & b_1 & \dots & b_n \end{bmatrix}^\top.$$

Here, we assumed, without loss of generality, that $A(q)$ and $B(q)$ are both modeled with n coefficients. Note that (6) is not in the model set defined by (7) due to the truncation by n coefficients. Nevertheless, the stability assumption on $G^\circ(q)$ and $1/H^\circ(q)$ implies that $\{a_k^\circ\}$ and $\{b_k^\circ\}$ are sequences converging to zero. Thus, if n is chosen large enough, (7) can model (6) with good accuracy.

An advantage of ARX models is that they are linear in the model parameters. In particular, the PEM estimate of η^n is obtained by minimizing the cost function

$$V_N(\eta^n) = \frac{1}{N}\sum_{t=1}^{N}\left[A(q,\eta^n)y_t - B(q,\eta^n)u_t\right]^2, \quad (8)$$

which can be done by linear least squares. Thus, it can be solved as follows. Write (7) as

$$y_t = (\varphi_t^n)^\top \eta^n + e_t,$$

where

$$\varphi_t^n = \begin{bmatrix} -y_{t-1} & \dots & -y_{t-n} & u_{t-1} \dots & u_{t-n} \end{bmatrix}^\top. \quad (9)$$

Then, the least squares estimate of η^n is given by

$$\hat{\eta}_N^{n,ls} := [R_N^n]^{-1} r_N^n, \quad (10)$$

where

$$R_N^n = \frac{1}{N}\sum_{t=1}^{N}\varphi_t^n(\varphi_t^n)^\top, \quad r_N^n = \frac{1}{N}\sum_{t=1}^{N}\varphi_t^n y_t.$$

In the analysis, we will use the slightly modified estimate

$$\hat{\eta}_N^n := [R_{N,\mathrm{reg}}^n]^{-1} r_N^n, \quad (11)$$

where

$$R_{N,\mathrm{reg}}^n = \begin{cases} R_N^n & \text{if } \left\|[R_N^n]^{-1}\right\|_2 < 2/\delta \\ R_N^n + \frac{\delta}{2} I_{2n} & \text{otherwise} \end{cases},$$

for some small $\delta > 0$. The reason is that $\hat{\eta}_N^n$ is easier to analyze statistically, while the first and second order statistical properties of $\hat{\eta}_N^{n,ls}$ and $\hat{\eta}_N^n$ are asymptotically identical [6].

It follows from Assumption 2.2 and Assumption 2.3 (see [6] for details),

$$\hat{\eta}_N^n \to \bar{\eta}^n := [\bar{R}^n]^{-1}\bar{r}^n,$$

where $\bar{R}^n$ and $\bar{r}^n$ are the limits of R_N^n and r_N^n w.p.1, respectively.

To guarantee that the true system (6) is asymptotically in the model set defined by the ARX model (7), n should be allowed to grow to infinity. Accordingly, we let the model order depend on the sample size N. For our theoretical results, we use the following assumption.

Assumption 3.1 (Model order) *It holds that*

$$n(N) \to \infty, \quad N \to \infty$$
$$n(N)^{4+\delta}/N \to 0, \quad N \to \infty$$

for some $\delta > 0$.

Introduce the notation $\hat{\eta}_N := \hat{\eta}_N^{n(N)}$ and, for future reference,

$$\eta_\circ^n := \begin{bmatrix} a_1^\circ & \dots & a_n^\circ & b_1^\circ & \dots & b_n^\circ \end{bmatrix}^\top, \quad (12)$$

$$\eta_\circ := \begin{bmatrix} a_1^\circ & a_2^\circ & \dots & b_1^\circ & b_2^\circ & \dots \end{bmatrix}^\top. \quad (13)$$

The asymptotic properties of $\hat{\eta}_N$ have been established in [6]. We will need the following result on the rate of convergence of the ARX model.

Lemma 3.1 *Assume that Assumptions 2.1, 2.2, 2.3 and 3.1 hold. Then with probability 1,*

$$\sup_\omega \left\| \begin{bmatrix} A(e^{j\omega}, \hat{\eta}_N) - A^\circ(e^{j\omega}) \\ B(e^{j\omega}, \hat{\eta}_N) - B^\circ(e^{j\omega}) \end{bmatrix} \right\|_2 = \mathcal{O}(m(N)),$$

where

$$m(N) = n(N)\sqrt{\log N/N}(1 + d(N)) + d(N)$$

and

$$d(N) := \sum_{k=n(N)+1}^{\infty} |a_k^\circ| + |b_k^\circ| \le \bar{C}\rho^{n(N)}, \quad (14)$$

for some $\bar{C} < \infty$ and $\rho < 1$.

PROOF. See Appendix A.

Lemma 3.1 implies that, as N tends to infinity, the coefficients of $A(q,\hat{\eta}_N)$ converge to those of $A^\circ(q) = 1/H^\circ(q)$, and the coefficients of $B(q,\hat{\eta}_N)$ converge to those of $B^\circ(q) = G^\circ(q)/H^\circ(q)$. Therefore, $B(q,\hat{\eta}_N)/A(q,\hat{\eta}_N)$ can be used as a high order estimate of $G^\circ(q)$, and

$1/A(q,\hat{\eta}_N)$ as a high order estimate of $H^\circ(q)$. We thus define these high order estimates by

$$G(q,\hat{\eta}_N) := \frac{B(q,\hat{\eta}_N)}{A(q,\hat{\eta}_N)}, \quad H(q,\hat{\eta}_N) := \frac{1}{A(q,\hat{\eta}_N)}. \quad (15)$$

Despite the simplicity of ARX models, they are not appropriate to model (2.1) for most practical uses. As the order n is required to be arbitrarily large, the estimated model will have unacceptably high variance.

Nevertheless, the high order ARX model estimate can be used to obtain a model of low order, reducing the variance. This can be done efficiently without re-using the data. The reason is that the estimate $\hat{\eta}_N$ and its covariance are asymptotically a sufficient statistic for our problem. To observe this, consider the infinite order ARX model

$$y_t = \varphi_t^\top \eta + e_t, \quad (16)$$

where

$$\varphi_t := \begin{bmatrix} -y_{t-1} & -y_{t-2} & \dots & u_{t-1} & u_{t-2} & \dots \end{bmatrix}^\top, \quad (17)$$

$$\eta := \begin{bmatrix} a_1 & a_2 & \dots & b_1 & b_2 & \dots \end{bmatrix}^\top. \quad (18)$$

Then, the probability density function of $y^N := \{y_t\}_{t=1}^N$ given η is

$$\begin{aligned} f(\theta; y^N) &= \prod_{t=1}^N \frac{1}{\sqrt{2\pi\sigma_\circ^2}} e^{-\frac{y_t - \varphi_t^\top \eta}{2\sigma_\circ^2}} \\ &= C e^{-\frac{1}{2\sigma_\circ^2}\left(\eta^\top \sum_{t=1}^N \varphi_t \varphi_t^\top \eta + 2\sum_{t=1}^N \varphi_t^\top y_t \eta\right)} e^{-\frac{1}{2\sigma_\circ^2}\sum_{t=1}^N y_t^2}. \end{aligned} \quad (19)$$

where we treat $\sigma_\circ^2$ as a known constant (in this case, because it is a scalar, it does not influence the estimation of the parameters of interest). Then, it follows from [4] that $R_N := \sum_{t=1}^N \varphi_t \varphi_t^\top$ and $r_N := \sum_{t=1}^N \varphi_t^\top y_t$ form a sufficient statistic for the data y^N. Alternatively, since $\hat{\eta}_N = R_N^{-1} r_N$, we can say that $\hat{\eta}_N$ and R_N are the sufficient statistic. However, when n is finite, there is a bias error induced by the truncation of the parameter sequences $\{a_k\}$ and $\{b_k\}$. If that error is assumed to be small, the estimate $\hat{\eta}_N^n$ will contain practically the same information about the system dynamics as the data. If the order n is allowed to tend to infinity as a function of the sample size N, according to Assumption 3.1, then the estimate $\hat{\eta}_N$ is, asymptotically, a sufficient statistic. Thus, the data could in principle be disregarded, and $\hat{\eta}_N$ alone be used to obtain an estimate of a lower order model that is asymptotically efficient.

4 Model Reduction

Having estimated a high order ARX model, we are interested in using this estimate to obtain a low order estimate $G(q,\theta)$. In this section, we discuss available approaches to do so.

4.1 Exact Maximum Likelihood

Being a sufficient statistic, $\hat{\eta}_N$ and its covariance can be used to obtain an estimate of θ that is asymptotically efficient. This can be done using an exact ML criterion [16]. Let $\eta^n(\theta,\alpha)$ be the parameter vector η^n obtained from θ and α, satisfying the relations

$$A(q,\eta) = \frac{1}{H(q,\alpha)}, \quad B(q,\eta) = \frac{G(q,\theta)}{H(q,\alpha)}. \quad (20)$$

This procedure consists in minimizing

$$\left[\hat{\eta}_N - \eta^n(\theta,\alpha)\right]^\top \left[\operatorname{cov}(\hat{\eta}_N)\right]^{-1} \left[\hat{\eta}_N - \eta^n(\theta,\alpha)\right], \quad (21)$$

where $\operatorname{cov}(\hat{\eta}_N)$ denotes the covariance of the estimated vector $\hat{\eta}_N$. Since this covariance matrix is in general unknown, in practice the cost function (21) requires an approximation. We consider two possibilities that do not affect the asymptotic properties of the obtained estimates.

One possibility consists in replacing $[\operatorname{cov}(\hat{\eta}_N)]^{-1}$ by a consistent estimate—for example, R_N^n [16]. In this case, we minimize

$$\left[\hat{\eta}_N - \eta^n(\theta,\alpha)\right]^\top R_N^n \left[\hat{\eta}_N - \eta^n(\theta,\alpha)\right], \quad (22)$$

which yields asymptotically efficient estimates of $G(q,\theta)$ and $H(q,\alpha)$. Because $\eta^n(\theta,\alpha)$ is nonlinear in general, minimizing (22) is a non-convex optimization problem.

Another possibility is to write the covariance matrix as function of the low order parameters θ and α—denoted $R^n(\theta,\alpha)$ (see [16] for details). In this case, we minimize the criterion

$$\left[\hat{\eta}_N - \eta^n(\theta,\alpha)\right]^\top R^n(\theta,\alpha) \left[\hat{\eta}_N - \eta^n(\theta,\alpha)\right], \quad (23)$$

Although minimizing (23) seems, at first sight, more complicated than minimizing (22), it is observed in [16] that the cost function (23) can be approximated by an asymptotic ML criterion that allows separating the estimation of $G(q,\theta)$ and $H(q,\alpha)$, while still providing asymptotically efficient estimates.

4.2 Asymptotic Maximum Likelihood (ASYM)

As shown in [16], minimizing (23) is asymptotically the same as minimizing

$$\int_0^{2\pi} \left|G(e^{i\omega},\hat{\eta}_N) - G(e^{i\omega},\theta)\right|^2 \frac{\Phi_u(e^{i\omega})}{\left|H(e^{i\omega},\hat{\eta}_N)\right|^2} d\omega + \frac{\hat{\sigma}^2}{2\pi}\int_0^{2\pi} \frac{\left|H(e^{i\omega},\hat{\eta}_N) - H(e^{i\omega},\alpha)\right|^2}{\left|H(e^{i\omega},\hat{\eta}_N)\right|^2} d\omega, \quad (24)$$

where $\hat{\sigma}^2$ is a consistent estimate of $\sigma_\circ^2$. Because the first term in (24) is only dependent on $G(q,\theta)$ and the second term on $H(q,\alpha)$, $G(q,\theta)$ can be estimated by minimizing the first term. Then, the minimization problem we are interested in becomes

$$V_N(\theta) = \int_0^{2\pi} \left|G(e^{i\omega},\hat{\eta}_N^n) - G(e^{i\omega},\theta)\right|^2 \frac{\Phi_u(e^{i\omega})}{\left|H(e^{i\omega},\hat{\eta}_N)\right|^2} d\omega. \quad (25)$$

The idea of the ASYM method [17] is to minimize the time domain equivalent to (25) for finite sample size:

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N}\left[\left(\frac{B(q,\hat{\eta}_N)}{A(q,\hat{\eta}_N)} - G(q,\theta)\right) A(q,\hat{\eta}_N)u_t\right]^2. \quad (26)$$

Minimizing (26) is still a non-convex optimization problem. However, it is pointed out in [17] that this minimization problem has an advantage over directly estimating $G(q,\theta)$ using PEM, which makes the method numerically more reliable. Because the output is not used explicitly in (26), and the noise contribution is only present indirectly through the high order estimates, the influence of the disturbance is reduced.

4.3 BJSM method

In alternative to using local non-linear optimization techniques, the BJSM method uses the Steiglitz-McBride iterations. The idea of BJSM is to first estimate a high order ARX model and then apply the Steiglitz-McBride method [12] to a data set pre-filtered by the ARX model estimate. The estimates obtained are asymptotically efficient in open loop. Because BJSM uses the Steiglitz-McBride, we start by reviewing the latter.

4.3.1 Steiglitz-McBride

The setting for the Steiglitz-McBride algorithm is when the transfer function $H^\circ(q)$ equals one (i.e., $C^\circ(q) = D^\circ(q) = 1$). The objective is to estimate $L(q,\theta)$ and $F(q,\theta)$.

Consider the following three steps. First, an ARX model

$$F(q,\theta)y_t = L(q,\theta)u_t + e_t$$

is estimated using least squares, providing an initilialization estimate $\hat{\theta}_N^0$. Second, the output and input are filtered by

$$y_t^f = \frac{1}{F(q,\hat{\theta}_N^1)}y_t, \qquad u_t^f = \frac{1}{F(q,\hat{\theta}_N^1)}u_t.$$

Third, least squares is applied to the ARX model

$$F(q,\theta)y_t^f = L(q,\theta)u_t^f + e_t,$$

providing a new estimate—$\hat{\theta}_N^1$. Then, we can continue to iterate by repeating Steps 2 and 3. We define the estimate obtained at iteration k by $\hat{\theta}_N^k$.

Notice that, since the true system has an OE structure, and we are estimating an ARX model, we are actually minimizing, in Step 1, the function

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N}\left[F(q,\theta)y_t - L(q,\theta)u_t\right]^2, \quad (27)$$

which, evaluated at the true parameter $\theta_\circ$, equals

$$V_N(\theta_\circ) = \frac{1}{N}\sum_{t=1}^{N}\left[F(q,\theta_\circ)e_t\right]^2. \quad (28)$$

From (28), we observe the true parameter $\theta_\circ$ does not correspond to the cost function of a white sequence. Consequently, the initialization estimate $\hat{\theta}_N^0$ is not consistent. However, at iteration k we have, evaluated at $\theta = \theta_\circ$,

$$V_N(\theta_\circ) = \frac{1}{N}\sum_{t=1}^{N}\left[\frac{F(q,\theta_\circ)}{F(q,\hat{\theta}_N^k)}e_t\right]^2. \quad (29)$$

So, assuming convergence to the true parameters (i.e., $\hat{\theta}_N^k \to \theta_\circ$, as $k \to \infty$ and $N \to \infty$), (29) asymptotically corresponds to (5) for an OE model structure.

Convergence of the Steiglitz-McBride has been studied in [14], where it is shown that the method is locally convergent when the additive output noise is white. Moreover, it will be globally convergent if the signal-to-noise ratio is sufficiently large. Assuming convergence, the estimates are asymptotically Gaussian distributed. However, in general, the covariance of the estimated parameters does not asymptotically attain M_{CR}^{-1}.

The Steiglitz-McBride is thus an attempt to minimize (5), but it only does so consistently with additive white noise, and even then it is not asymptotically efficient.

4.3.2 BJSM

In [18], the Box-Jenkins Steiglitz-McBride (BJSM) algorithm is introduced. This algorithm copes with two limitations of the Steiglitz-McBride. First, it is consistent for systems with BJ structure, instead of only OE. Second, it is asymptotically efficient for open loop data.

The method uses the following procedure. First, an ARX model (7) is estimated with least squares. Second, the original data set is pre-filtered by $A(q, \hat{\eta}_N)$. Third, the Steiglitz-McBride algorithm is applied to the pre-filtered data set.

Recall that, to be convergent, the Steiglitz-McBride algorithm requires that $H^\circ(q) = 1$. The main idea of BJSM is thus to use $A(q, \hat{\eta}_N)$ as an estimate of $[H^\circ(q)]^{-1}$ and pre-filter the data according to

$$y_t^{\mathrm{pf}} = A(q, \hat{\eta}_N) y_t, \qquad u_t^{\mathrm{pf}} = A(q, \hat{\eta}_N) u_t.$$

Then, the pre-filtered data satisfies

$$y_t^{\mathrm{pf}} = \frac{L^\circ(q)}{F^\circ(q)} u_t^{\mathrm{pf}} + A(q, \hat{\eta}_N) H^\circ(q) e_t, \tag{30}$$

which asymptotically is according to, due to Lemma 3.1,

$$y_t^{\mathrm{pf}} \approx \frac{L^\circ(q)}{F^\circ(q)} u_t^{\mathrm{pf}} + e_t. \tag{31}$$

Since (31) is of OE structure, the Steiglitz-McBride algorithm can be applied to the data set $\{y_t^{\mathrm{pf}}, u_t^{\mathrm{pf}}\}$.

Notice that, if we were to apply PEM to the pre-filtered data set, we would minimize, motivated by (31),

$$V_N(\theta) = \frac{1}{N} \sum_{t=1}^{N} \left(y_t^{\mathrm{pf}} - \frac{L(q,\theta)}{F(q,\theta)} u_t^{\mathrm{pf}} \right)^2. \tag{32}$$

To avoid an explicit non-convex minimization problem, we use the Steiglitz-McBride method instead. Although the Steiglitz-McBride is not asymptotically efficient, the BJSM method is when used with open loop data [18].

However, not all the information in $\hat{\eta}_N$ is being used, as the filtering (30) only uses $A(q, \hat{\eta}_N)$. In other words, the ARX model is not used as a sufficient statistic for this problem. For the method to still be asymptotically efficient, the output data are used when constructing the pre-filtering. This leads to two limitations.

The first is a counter-intuitive result. Suppose that $H^\circ(q) = 1$ (i.e., the true system is already of OE structure). Then, we have that $A^\circ(q) = 1$, and estimating a finite impulse response (FIR) model would suffice to asymptotically model the true system. However, this would maintain the data set unchanged when applying the filtering (30), and BJSM would simply be reduced to the Steiglitz-McBride method, which is not asymptotically efficient. If, on the other hand, it is not assumed that $A^\circ(q) = 1$ and an estimate $A(q, \hat{\eta}_N)$ is still computed, BJSM will be asymptotically efficient. Thus, although an FIR model is asymptotically a sufficient statistic for a system of OE structure (like the ARX model is for BJ structures) it is not possible to make use of this information when applying the BJSM method, since it does not exploit the full statistical properties of the high order model.

As for the second limitation, we observe that although BJSM avoids solving a non-convex optimization problem by applying the Steiglitz-McBride algorithm, it has the disadvantage of requiring the number of iterations of the Steiglitz-McBride to tend to infinity in order to provide consistent and asymptotically efficient estimates [18]. To bypass this problem but still avoid a non-convex minimization procedure, we use the Steiglitz-McBride with the ASYM method. This will allow us to obtain an asymptotically efficient estimate in one iteration.

5 Model Order Reduction Steiglitz-McBride

The objective of our approach is to minimize (26) without using a non-convex optimization method. To do so, we use an approach that combines ideas from ASYM and BJSM.

First, we write (26) as

$$V_N(\theta) = \frac{1}{N} \sum_{t=1}^{N} \left[B(q, \hat{\eta}_N) u_t - \frac{L(q,\theta)}{F(q,\theta)} A(q, \hat{\eta}_N) u_t \right]^2. \tag{33}$$

Then, we notice that (33) has the same form as (32) if we define

$$y_t^{\mathrm{pf}} := B(q, \hat{\eta}_N) u_t, \qquad u_t^{\mathrm{pf}} := A(q, \hat{\eta}_N) u_t, \tag{34}$$

and thus the same idea (i.e., applying the Steiglitz-McBride to $\{y_t^{\mathrm{pf}}, u_t^{\mathrm{pf}}\}$) can be used.

The only difference between this approach and BJSM is in the pre-filtered output. Comparing (34) and (30), we observe that u_t^{pf} are defined similarly, but y_t^{pf} are different. The difference lies in the true output not being used to construct the new pre-filtered data set. Rather, it is simulated from the input and the ARX model estimate. Indeed, we can simulate the output with

$$y_t^s := \frac{B(q, \hat{\eta}_N)}{A(q, \hat{\eta}_N)} u_t, \tag{35}$$

and then apply the same filter as in (30), but on the simulated output, according to

$$y_t^{\rm pf} = A(q,\hat{\eta}_N)y_t^s = B(q,\hat{\eta}_N)u_t, \tag{36}$$

obtaining the proposed pre-filter (34).

In summary, the proposed method is as follows:

(1) estimate an ARX model using the input-output data $\{u_t, y_t\}$, $t = 1,\dots,N$, according to (10);
(2) construct the pre-filtered data $\{u_t^{\rm pf}, y_t^{\rm pf}\}$, according to (34);
(3) apply the Steiglitz-McBride method with $\{u_t^{\rm pf}, y_t^{\rm pf}\}$ to obtain estimates $L(q,\hat{\theta}_N)$ and $F(q,\hat{\theta}_N)$ of $L^\circ(q)$ and $F^\circ(q)$, respectively.

Note that the pre-filtered data set (34) only depends on the original output data $\{y_t\}$ through the least squares estimate $\hat{\eta}_N$. With this method, we use the high order ARX model as an asymptotic sufficient statistic for our problem, and disregard the data without loss of information. Indeed, as will be shown in the next section, this procedure is asymptotically efficient for open loop data.

Moreover, there are two advantages for disregarding the data after the high order ARX model has been estimated. Although these are formally shown in the next section, we observe them here, supported by intuitive explanations.

First, the pre-filter (34) uses the complete statistical information contained in the estimate $\hat{\eta}_N$. So, if the noise contribution affecting the true system (1) is white, a high-order FIR model can be estimated instead of an ARX. In this case, $A(q,\hat{\eta}_N) = 1$.

Second, this procedure asymptotically (in N) only requires one iteration. To intuitively understand why this is the case, we recall why the Steiglitz-McBride is an iterative method. Note that the initialization step of the Steiglitz-McBride minimizes (27), which, when evaluated at the true parameters, as in (28), does not correspond to a cost function of white sequence. Therefore, the initialization estimate $\hat{\theta}_N^0$ is biased. Then, we start iterating. At the first iteration, the cost function evaluated at $\theta_\circ$ is given by (29) with $F(q,\hat{\theta}_N^0)$. Because $F(q,\hat{\theta}_N^0)$ is biased, the true parameter will still not correspond to the cost function of a white sequence. Therefore, the new estimate $\hat{\theta}_N^1$ will not be consistent either. However, by continuing to iterate, it can be shown that, under the conditions observed in [13], $\hat{\theta}_N^k \to \theta_\circ$, as $k \to \infty$ and $N \to \infty$. Concerning the original BJSM method, since the pre-filtered data is according to (30), it is asymptotically approximately an OE model structure, and a similar procedure takes place.

On the other hand, the alternative pre-filtering, which disregards the original data, satisfies

$$y_t^{\rm pf} = \frac{L^\circ(q)}{F^\circ(q)}u_t^{\rm pf} + \left(\frac{B(q,\hat{\eta}_N)}{A(q,\hat{\eta}_N)} - \frac{L^\circ(q)}{F^\circ(q)}\right)u_t^{\rm pf}. \tag{37}$$

This is a noise-free equation, except for the noisy parameters in the ARX model. However, from Lemma 3.1, the second term in (37) tends to zero asymptotically. As consequence, the variance of the error sequence being minimized by the Steiglitz-McBride iterations disappears asymptotically, and only one iteration is required.

We observe that the proposed method essentially consists of applying the Steiglitz-McBride algorithm to perform model order reduction based on an asymptotic ML criterion. We will thus refer to the method as Model Order Reduction Steiglitz-McBride (MORSM). The idea of using the Steiglitz-McBride to, in some sense, perform model order reduction, is not new. Variants of the Steiglitz-McBride method have been applied to estimate rational filters from an impulse response estimate, instead of applying the method directly to data (see, e.g., [1, 7, 9]). However, although some of these procedures are in some sense optimal under specific conditions, we consider a quite general system identification problem and motivate the application of the method based on an ML criterion. This, as we proceed to show, not only provides asymptotically efficient estimates under a quite general class of systems and external signals, but also does so in one iteration.

6 Asymptotic Properties

In this section, we analyze both the convergence and asymptotic covariance of the proposed method. To derive these results, we will need a formal expression for the estimate of θ at iteration $k+1$ of the MORSM algorithm. Define

$$y_t(\eta,\theta) = \frac{B(q,\eta)}{F(q,\theta)}u_t, \quad y_t(\eta_\circ,\theta) = \frac{B^\circ(q)}{F(q,\theta)}u_t,$$
$$u_t(\eta,\theta) = \frac{A(q,\eta)}{F(q,\theta)}u_t, \quad y_t(\eta_\circ,\theta) = \frac{A^\circ(q)}{F(q,\theta)}u_t,$$

and

$$\xi_t(\eta,\theta) = \frac{L^\circ(q)}{F^\circ(q)}\frac{B(q,\eta)-B^\circ(q)}{B^\circ(q)}u_t(\eta,\theta) - \frac{A(q,\eta)-A^\circ(q)}{A^\circ(q)}y_t(\eta,\theta).$$

The same definition also applies to vector valued signals, such as (9).

Using that the pre-filtered data set consists of filtered versions of u_t and that $G(q)$ can be represented both

using $L^\circ(q)$ and $F^\circ(q)$ as well as using $B^\circ(q)$ and $A^\circ(q)$, we have that

$$u_t = \frac{1}{B(q,\hat{\eta}_N)} y_t^{\text{pf}} = \frac{L^\circ(q)A^\circ(q)}{F^\circ(q)B^\circ(q)} \frac{1}{A(q,\hat{\eta}_N)} u_t^{\text{pf}}. \quad (38)$$

Filtering (38) by

$$F^\circ(q)\frac{A(q,\hat{\eta}_N)B(q,\hat{\eta}_N)}{A^\circ(q)F(q,\hat{\eta}_N)},$$

we arrive at the noise-free equation

$$F^\circ(q)\frac{A(q,\hat{\eta}_N)}{A^\circ(q)} y_t(\hat{\eta}_N,\hat{\theta}_N^k) = L^\circ(q)\frac{B(q,\hat{\eta}_N)}{B^\circ(q)} u_t(\hat{\eta}_N,\hat{\theta}_N^k)$$

relating the pre-filtered data. Equivalently,

$$F^\circ(q)y_t(\hat{\eta}_N,\hat{\theta}_N^k) = L^\circ(q)u_t(\hat{\eta}_N,\hat{\theta}_N^k) + F^\circ(q)\xi_t(\hat{\eta}_N,\hat{\theta}_N^k),$$

which can be written in regression form as

$$y_t(\hat{\eta}_N,\hat{\theta}_N^k) = [\varphi^m(\hat{\eta}_N,\hat{\theta}_N^k)]^\top \theta_\circ + F^\circ(q)\xi_t(\hat{\eta}_N,\hat{\theta}_N^k). \quad (39)$$

Given $\hat{\theta}_N^k$, the next parameter estimate in the Steiglitz-McBride iterations $\hat{\theta}_N^{k+1}$, is defined as the least squares estimate of $\theta_\circ$ in the linear regression (39):

$$\hat{\theta}_N^{k+1} = [R^m(\hat{\eta}_N,\hat{\theta}_N^k)]^{-1} r^m(\hat{\eta}_N,\hat{\theta}_N^k), \quad (40)$$

where

$$R^m(\eta^n,\theta) = \frac{1}{N}\sum_{t=m+1}^{N} \varphi_t^m(\eta^n,\theta)(\varphi_t^m(\eta^n,\theta))^\top,$$
$$r^m(\eta^n,\theta) = \frac{1}{N}\sum_{t=m+1}^{N} \varphi_t^m(\eta^n,\theta) y_t(\eta^n,\theta).$$

Notice that (39) is a linear regression form of (37) with the notable difference that the error made in the ARX model enters linearly into $\xi_t(\hat{\eta}_N,\hat{\theta}_N^k)$. As before, the ARX model error tends to zero asymptotically. This is, in essence, what enables the following results.

Theorem 6.1 *Let Assumptions 2.1, 2.2, 2.3, and 3.1 hold. Then,*

$$\hat{\theta}_N^k \to \theta_\circ \quad \text{as } N\to\infty, \text{ w.p. 1}, \quad \text{for all } k \geq 0$$

PROOF. See Appendix B.

Theorem 6.1 implies that the proposed algorithm achieves consistency in the initialization estimate—that is, $\hat{\theta}_N^0$ is a consistent estimate of $\theta_\circ$. This was not the case for the BJSM algorithm.

For the asymptotic covariance, we have the following theorem.

Theorem 6.2 *Let Assumptions 2.1, 2.2, 2.3, and 3.1 hold. Then,*

$$\lim_{N\to\infty} N\text{E}\left[(\hat{\theta}_N^k - \theta_\circ)(\hat{\theta}_N^k - \theta_\circ)^\top\right] = \sigma_\circ^2 M_{CR}^{-1},$$

and $\sqrt{N}(\hat{\theta}_N^k - \theta_\circ) \sim \text{AsN}(0, \sigma_\circ^2 M_{CR}^{-1})$ *for* $k \geq 1$, *where* N *stands for the normal distribution.*

PROOF. See Appendix D.

From Theorem 6.2, we observe that the proposed method has the same asymptotic covariance as PEM with Gaussian noise (6). Therefore, it is asymptotically efficient with open loop data. Moreover, the asymptotic efficient estimate is obtained in one iteration, at $k = 1$.

7 Simulations

In this section, we perform two Monte Carlo simulations to study the performance of the method. First, we illustrate how it converges in one iteration of the Steiglitz-McBride, while BJSM does not. Then, we perform a study with random systems, and observe that the method often has better finite sample convergence properties than PEM.

7.1 One iteration scheme

In the first simulation, we compare MORSM and BJSM. The practical difference between these methods is in the pre-filtering only. In particular, MORSM does not use the noisy output to construct the pre-filtered data set. The consequence is that the method provides asymptotically efficient estimates in one iteration.

For the simulation, the data are generated by

$$y_t = \frac{q^{-1} + 0.1q^{-2}}{1 - 1.2q^{-1} + 0.6q^{-2}} u_t + \frac{1 + 0.7q^{-1}}{1 - 0.9q^{-1}} e_t. \quad (41)$$

One hundred Monte Carlo simulations are performed with eight sample sizes logarithmically spaced between $N = 200$ and $N = 20000$. The sequence $\{u_t\}$ is obtained by

$$u_t = \frac{1}{1 - q^{-1} + 0.89q^{-2}} w_t, \quad (42)$$

where $\{w_t\}$ and $\{e_t\}$ are independent Gaussian white sequences with unit variance.

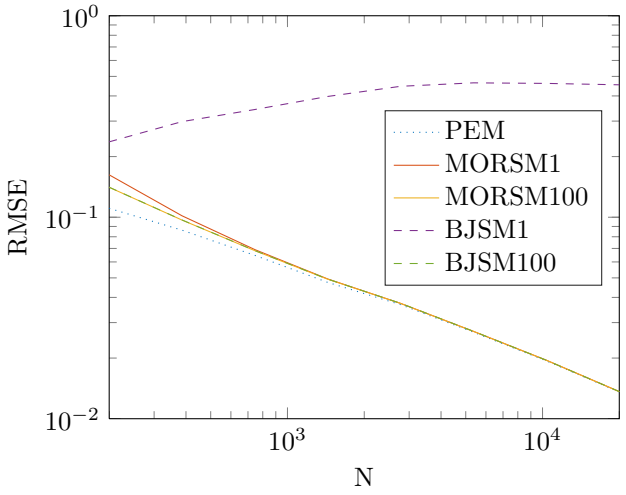

Fig. 1. Average RMSE as function of sample size for several methods, obtained from 100 Monte Carlo runs with a fixed system.

We compare PEM, BJSM (one and 100 iterations), and MORSM (one and 100 iterations). All methods estimate a plant parameterized with the correct orders, and PEM also estimates a correctly parameterized noise model. For BJSM and the proposed method, an ARX model of order 50 is estimated in the first step. As the objective of this simulation is to observe convergence and asymptotic variance properties, PEM is started at the true parameters, and all methods assume known initial conditions.

The results are presented in Fig. 1, where the average root mean square error (RMSE) of the impulse response from 1000 Monte Carlo runs is presented for each sample size. The RMSE is given by

$$\text{RMSE} = \|g^\circ - \hat{g}\|_2 \,, \tag{43}$$

where g° is the impulse response of $G^\circ(q)$ and $\hat{g}$ is the impulse response of the estimated plant model. In Fig. 1, we observe that MORSM and BJSM perform similarly with 100 iterations for all the sample size range used. MORSM performs slightly worse with one iteration than with 100 for small sample sizes, but they have the same performance for larger N. However, the same is not true for BJSM with one iteration, for which the RMSE does not even decrease with increasing sample size.

In conclusion, if a sufficiently amount of iterations are performed, both MORSM and BJSM attain the asymptotic covariance of PEM. However, BJSM theoretically needs the Steiglitz-McBride iterations to tend to infinity, while MORSM only needs one iteration.

7.2 Comparison with PEM

According to a prediction error criterion, the best model is the one that minimizes a cost function of the prediction errors. The estimate corresponding to the minimizer of this cost function is asymptotically efficient, meaning that is has asymptotically the minimum possible covariance for a consistent estimator. The limitation is that this cost function is, in general, non-convex. Seeking the global minimum requires local non-linear optimization techniques, and it is not guaranteed to be found. As the number of parameters to estimate increases, PEM has increasingly more difficulty in finding the global minimum.

In the following simulation, we will compare the performance of PEM and the proposed method with randomly generated systems, with structure

$$y_t = \frac{l_1^\circ q^{-1} + l_2^\circ q^{-2} + l_3^\circ q^{-3} + l_4^\circ q^{-4}}{1 + f_1^\circ q^{-1} + f_2^\circ q^{-2} + f_3^\circ q^{-3} + f_4^\circ q^{-4}} u_t + \frac{1 + c_1^\circ q^{-1} + c_2^\circ q^{-2} + c_3^\circ q^{-3} + c_4^\circ q^{-4}}{1 + d_1^\circ q^{-1} + d_2^\circ q^{-2} + d_3^\circ q^{-3} + d_4^\circ q^{-4}} e_t, \tag{44}$$

where $\{u_t\}$ is given as in the previous simulation, and $\{e_t\}$ is Gaussian white noise with variance chosen to obtain a signal-to-noise ratio

$$\text{SNR} = \frac{\sum_{t=1}^{N} (u_t)^2}{\sum_{t=1}^{N} (H^\circ(q) e_t)^2} = 10. \tag{45}$$

The coefficients of $L^\circ(q)$ are generated from a uniform distribution, with values between -1 and 1. The coefficients of the remaining polynomials are generated such that $F^\circ(q)$, $C^\circ(q)$, and $D^\circ(q)$ have all roots inside a half-ring in the unit disc with a radius between 0.7 and 0.9, with positive real part. We do this with the objective of studying a particular class of systems: namely, the systems are effectively of fourth order (i.e., no poles are considerably dominant over others), they can be approximated by ARX models roughly of orders between 30 and 100, and they resemble physical systems.

An important practical aspect in implementing the proposed method is how to choose the ARX model order, in case we do not previously have information of an appropriate order to choose. As we have seen, theoretically the ARX model order should tend to infinity as function of the sample size. However, for practical purposes it is sufficient to choose an order that can correctly capture the dynamics of the true system. We then propose the following procedure to choose the order of the ARX model. Since our objective is to minimize the loss function (5) using an indirect approach, we repeat the estimation for a grid of ARX model orders, and choose the low order model that minimizes (5). Since we do not compute a low order noise model, the highest order ARX polynomial $A(q, \hat{\eta}_N)$ is used instead of $1/H(q, \alpha)$ when computing this loss function. Although this is a very noisy estimate, the error induced will be the same for every computation, and should not have a considerable influence

in choosing the best model. For the class of systems we consider, we choose the ARX model order from a grid of values between 25 and 125, spaced with intervals of 25.

Moreover, when more than one iteration is used, the same criterion can be applied to optimize over the number of iterations—that is, we choose the model obtained at the iteration that minimizes the cost function (5).

We compare the following methods:

- the prediction error method, initialized at the true parameters (PEM true);
- the prediction error method, initialized with the standard MATLAB procedure (PEM);
- the Box-Jenkins Steiglitz-McBride method, with 20 iterations (BJSM20);
- the Model Order Reduction Steiglitz-McBride method, with 20 iterations (MORSM20);
- the Model Order Reduction Steiglitz-McBride, with one iteration (MORSM1).

PEM stops with a maximum of 1000 iterations and a function tolerance of 10^{-5}, and estimates initial conditions. MORSM and BJSM truncate initial conditions. Note that a procedure to estimate initial conditions for this type of methods has been proposed in [3], but it is only applicable if the plant and noise model share the same poles (e.g., ARMA, ARMAX) or if the noise model poles are known (e.g., OE), which is not the case of BJ models.

The performance of each method is evaluated by calculating the FIT of the impulse response of the plant, given by, in percent,

$$\text{FIT} = 100\left(1 - \frac{\text{RMSE}}{\|g^\circ - \bar{g}^o\|}\right), \tag{46}$$

where $\bar{g}^o$ is the average of g°.

The results are presented in Fig. 2, with the average FIT as function of sample size. We assume that PEM, when initialized at the true parameters, converges to the global optimum. Comparing PEM initialized at the true parameters and with the standard MATLAB procedure, we conclude that the latter must sometimes fail to reach the global optimum.

With 20 iterations, MORSM does not seem to reach the global minimum of the prediction error cost function for small sample sizes. However, for sample sizes around 4000 and larger, this minimum seems to be attained since MORSM performs similar to PEM initialized at the true parameters. This suggests that MORSM may be a viable alternative to PEM when PEM has difficulty in finding the global minimum.

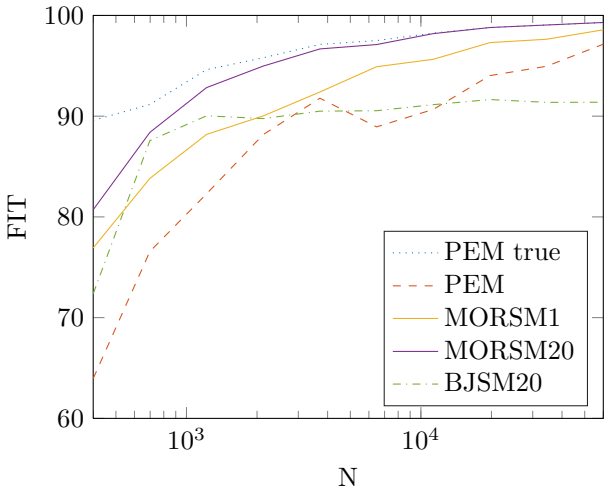

Fig. 2. Average FIT for several methods, obtained from 100 Monte Carlo runs with random systems.

With only one iteration, MORSM performs worse than with 20 iterations for the range of sample sizes used, but their performances becomes closer as larger sample sizes are used. Theoretically, we have shown that MORSM only requires one iteration to provide asymptotically efficient estimates. However, we observe that in practice (i.e., for finite sample size), MORSM performs better with more iterations, comparing MORSM1 and MORSM20 in Fig. 2. The fact that in practice MORSM requires more than one iteration to converge does not render it irrelevant in comparison to BJSM. As we observe in Fig. 2, BJSM with 20 iterations does not attain the same asymptotic performance of MORSM because 20 iterations do not seem to be sufficient for BJSM to converge in this simulation, while they are sufficient for MORSM.

8 Conclusion

In this paper, we propose a least squares method for estimation of models with a plant parameterized by a rational transfer function and a non-parametric noise model. We show that the method provides consistent and asymptotically efficient estimates of the plant if data are obtained in open loop.

Essentially, the method performs model order reduction based on an asymptotic ML criterion using the Steiglitz-McBride method. We thus name it Model Order Reduction Steiglitz-McBride (MORSM). The method uses ideas from the ASYM and BJSM methods. However, unlike ASYM, we avoid a non-convex optimization procedure by applying Steiglitz-McBride; unlike BJSM, we propose a procedure that only requires one iteration to provide asymptotically efficient estimates.

Finally, we perform two simulation studies to analyze the performance of the method, from which the following

are observed. First, MORSM is asymptotically efficient in one iteration, while BJSM is not. Second, even when extra iterations are required for convergence with finite sample sizes, MORSM still converges in less iterations than BJSM. Third, MORSM may be a viable alternative to PEM, specially when PEM has difficulty in finding the global minimum.

Future work will include application of MORSM for closed loop and for estimation of systems embedded in networks.

A Proof of Lemma 3.1

The result follows from Theorem 3.1 in [6]. Next, we verify the conditions of that theorem. Assumption 2.1 and the finite dimensionality of G° and H° implies that

$$\max(|a_k|, |b_k|) \leq C\rho^k \tag{A.1}$$

for some $C < \infty$ and $0 < \rho < 1$. This implies that Condition S1 holds. Furthermore, the bound (A.1) implies the inequality in (14) for some $\bar{C} < \infty$. Assumption 2.3 clearly implies Condition S2 (for any $p \leq 5$). Assumption 3.1 implies Conditions D1 and D3. Thus all conditions in Theorem 3.1 of [6] have been verified and the result in the lemma follows from this theorem.

B Proof of Theorem 6.1

Using Parseval's formula, we have

$$\bar{R}(\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\begin{bmatrix}-B^\circ\Gamma_m\\ A^\circ\Gamma_m\end{bmatrix}\begin{bmatrix}-B^\circ\Gamma_m\\ A^\circ\Gamma_m\end{bmatrix}^*\frac{\Phi_u}{|F(\theta)|^2}\,\mathrm{d}\omega \tag{B.1}$$

We notice that $\bar{R}(\theta) > 0$ whenever θ is in the stability region for the coefficients of polynomials of degree m

$$\bar{S} := \{\theta : F(z,\theta) = 0 \Rightarrow |z| < 1\} \subset \mathbb{R}^{2m} \tag{B.2}$$

We introduce the notation

$$f(N) = \mathcal{O}(g(N))$$

to mean that $f(N)$ decays to zero with the rate $g(N)$, i.e., that there exists some positive constants C and N_0 such that for all $N \geq N_0$,

$$\|f(N)\| \leq C|g(N)| \text{ as } N \to \infty.$$

From Lemma 3.1 it follows that

$$R^m(\hat{\eta}_N, \theta) - \bar{R}(\theta) = \mathcal{O}(m(N)). \tag{B.3}$$

By standard continuity arguments, with probability 1

$$R^m(\hat{\eta}_N, \theta) > 0$$

for large enough N. Hence, for N large enough, using (39) in (40)

$$\hat{\theta}_N^{k+1} = \theta_\circ + [R^m(\hat{\eta}_N, \theta_N^k)]^{-1} \cdot\frac{1}{N}\sum_{t=m+1}^{N}\varphi_t^m(\eta^n, \theta_N^k)F^\circ(q)\xi_t(\hat{\eta}_N, \hat{\theta}_N^k). \tag{B.4}$$

Now, since $\{u_t\}$ is uniformly bounded and $1/F(q,\theta)$ is uniformly stable, it follows that

$$\left\|\varphi_t^m(\hat{\eta}_N, \theta_N^k)\right\| \leq C_1,$$

for some $C_1 < \infty$, and furthermore, by Lemma 3.1, it follows that

$$F^\circ(q)\xi_t(\hat{\eta}_N, \hat{\theta}_N^k) = \mathcal{O}(m(N)).$$

It thus follows that

$$\hat{\theta}_N^{k+1} - \theta_\circ = \mathcal{O}(m(N)), \tag{B.5}$$

for any $k \geq 0$ and

$$\left\|\hat{\theta}_N^{k+1} - \theta_\circ\right\| \to 0, \quad \text{as N} \to \infty, \text{ w.p. 1.}$$

C Auxiliary lemmas

This section includes a few results needed for the proof of Theorem 6.2 in Section D.

Lemma C.1 *Assume that* $X(q) = \sum_{k=1}^{n} x_k q^{-k}$ *and* $Z(q) = \sum_{l=1}^{n} z_l q^{-l}$ *are stable filters and let* $v(t)$ *be quasi-stationary. Then,*

$$\left\|\frac{1}{N}\sum_{t=m+1}^{N}X(q)v(t)Z(q)v(t)\right\|_2 \leq \|X\|_2\,\|Z\|_2\,C$$

for some $C < \infty$.

PROOF.

$$\begin{aligned}
&\Big\|\frac{1}{N}\sum_{t=m+1}^{N} X(q)v(t)Z(q)v(t)\Big\|^2 \\
&\quad= \left\|\frac{1}{N}\sum_{t=m+1}^{N}\sum_{k=1}^{n} x_k v_{t-k}\sum_{l=1}^{n} z_l v_{t-l}\right\|^2 \\
&\quad= \left\|\sum_{k=1}^{n} x_k \sum_{l=1}^{n} z_l \frac{1}{N}\sum_{t=m+1}^{N} v_{t-k}v_{t-l}\right\|^2 \\
&\quad\le \sum_{k=1}^{n}|x_k|^2 \sum_{l=1}^{n}|z_l|^2 \left|\frac{1}{N}\sum_{t=m+1}^{N} v_{t-k}v_{t-l}\right|^2 \\
&\quad\le \sum_{k=1}^{n}|x_k|^2 \sum_{l=1}^{n}|z_l|^2 \left|R_{vv}^N(k-l)\right|^2 \\
&\quad\le \|X\|_2^2\,\|Z\|_2^2\,C^2,
\end{aligned}$$

where the last equality is due to the quasi stationarity of $v(t)$.

Lemma C.2 *Let Assumptions 2.1, 2.2, 2.3, and 3.1 be in force. Let Υ^n be an $m \times 2n$ deterministic matrix, with m fixed. Then, we have that*

$$\sqrt{N}\Upsilon^n(\hat{\eta}_N - \bar{\eta}^n) \sim \mathrm{AsN}(0, P), \tag{C.1}$$

where

$$P = \sigma_\circ^2 \lim_{n\to\infty} \Upsilon^n[\bar{R}^n]^{-1}(\Upsilon^n)^\top, \tag{C.2}$$

if the limit exists.

PROOF. See [6, Theorem 7.3].

Lemma C.3 *Let $\{x_n\}$ be a sequence of random variables that is asymptotically Gaussian distributed—$\{x_n\} \sim AsN(0, P)$. Let $\{M_n\}$ be a sequence of random square matrices that converge in probability to a non-singular matrix M, and $\{b_n\}$ be a sequence of random vectors that converges in probability to b. Also, let*

$$y_n = M_n x_n + b_n. \tag{C.3}$$

Then, y_n converges in distribution to $\mathcal{N}(b, MPM^\top)$.

PROOF. See [11, Lemma B.4].

Lemma C.4 *Let $\mathcal{S}_n$ be the subspace of $\mathcal{L}_2^2$ spanned by the rows of*

$$\begin{bmatrix} -F_1F_3\Gamma_n & F_2\Gamma_n \\ F_3\Gamma_n & 0 \end{bmatrix}, \tag{C.4}$$

where

$$\Gamma_n(q) = \begin{bmatrix} q^{-1} & \dots & q^{-n} \end{bmatrix}, \tag{C.5}$$

$$F_i(q) = \sum_{k=0}^{\infty} f_k^i q^{-k}. \tag{C.6}$$

Suppose that F_1, F_2 and F_3 are exponentially stable, i.e., for an exponentially stable F_i

$$|f_k^i| \le C\lambda^k, \quad \text{for some } C < \infty, \quad \lambda < 1, \tag{C.7}$$

and that there is a causal exponentially stable inverse

$$\tilde{F}_2(q) = \sum_{k=0}^{\infty} \tilde{f}_k^2 q^{-k}, \quad |\tilde{f}_k^2| < C\lambda^k. \tag{C.8}$$

Let $\gamma = [\sum_{k=1}^{\infty} d_k q^{-k} \quad 0\,]$ be exponentially stable. Then

$$\|\gamma - \mathbf{P}_{\mathcal{S}_n}[\gamma]\|_2 \le C\lambda^n, \text{ for some } C < \infty,\ \lambda < 1. \tag{C.9}$$

PROOF. We will construct an explicit approximation to γ that belongs to $\mathcal{S}_n$. Let

$$\tilde{F}_u\gamma = \begin{bmatrix} \sum_{l=1}^{\infty} \beta_l z^{-l} & 0 \end{bmatrix},$$

which is exponentially stable since both γ and $\tilde{F}_2$ are exponentially stable. Take as approximation for γ

$$\hat{\gamma}_n = \begin{bmatrix} \sum_{l=1}^{n} \beta_l F_2(z) z^{-l} & 0 \end{bmatrix},$$

which by construction belongs to $\mathcal{S}_\Psi$. Introduce the notation $\gamma = \begin{bmatrix} \gamma_1 & \gamma_2 \end{bmatrix}$. Hence

$$\begin{aligned}
\big\|\gamma_k - \mathbf{P}_{\mathcal{S}_{\tilde{\Psi}}}[\gamma]\big\|_2 &\le \|\gamma - \hat{\gamma}_n\|_2 \\
&= \left\|\gamma_1 - \sum_{l=1}^{n} \beta_l F_2(z) z^{-l}\right\|_2 \\
&= \left\|F_2(z)\left(\tilde{F}_2(z)\gamma_1 - \sum_{l=1}^{n} \beta_l z^{-l}\right)\right\|_2 \\
&\le \|F_2(z)\|_2 \left\|\sum_{l=n+1}^{\infty} \beta_l z^{-l}\right\|_2 \le C\lambda^n,
\end{aligned}$$

for some $C < \infty$ and $\lambda < 1$ since F_2 and $\tilde{F}_2\gamma$ are exponentially stable.

D Proof of Theorem 6.2

We start by using (B.4) to write

$$\sqrt{N}(\hat{\theta}_N^{k+1} - \theta_\circ) = M_N^{-1} x_N,$$

where

$$M_N = R^m(\hat{\eta}_N, \theta_N^k)$$
$$x_N = \frac{1}{\sqrt{N}} \sum_{t=m+1}^{N} \varphi_t^m(\hat{\eta}_N, \theta_N^k) F^\circ(q) \xi_t(\hat{\eta}_N, \hat{\theta}_N^k).$$

From (B.3) and Theorem 6.1, for $k \geq 1$, we have that

$$M_N \to M_{CR}, \quad \text{as N} \to \infty, \quad \text{w.p. 1.}$$

Assume for now (we will prove it later) that

$$x_N \sim \text{AsN}(0, P).$$

Then, using Lemma C.3, we have that

$$\sqrt{N}(\hat{\theta}_N^{k+1} - \theta_\circ) \sim \text{AsN}(0, M_{CR}^{-1} P M_{CR}^{-1}). \quad \text{(D.1)}$$

D.1 x_N

We will now establish the asymptotic distribution and covariance of x_N. To this end, we first define

$$\Phi^m(\eta^n, \theta) := \frac{1}{F(q,\theta)} \begin{bmatrix} -B(q,\eta^n)\Gamma_m \\ A(q,\eta^n)\Gamma_m \end{bmatrix},$$
$$\Xi^m(\eta^n, \theta) := \frac{F^\circ(q)}{A^\circ(q)F(q,\theta)} \cdot \begin{bmatrix} -B^\circ(q) & A^\circ(q) \end{bmatrix} \begin{bmatrix} A(q,\eta^n) - A^\circ(q) \\ B(q,\eta^n) - B^\circ(q) \end{bmatrix}.$$

Then we rewrite $\xi_t(\hat{\eta}_N, \theta_N^k)$ as

$$\begin{aligned}
\xi_t(\hat{\eta}_N, \theta_N^k) &= -\frac{B(q,\hat{\eta}_N)}{A^\circ(q)F(q,\theta_N^k)}(A(q,\hat{\eta}_N) - A^\circ(q))u_t \\
&\quad + \frac{A(q,\hat{\eta}_N)}{A^\circ(q)F(q,\theta_N^k)}(B(q,\hat{\eta}_N) - B^\circ(q))u_t \\
&= -\frac{B^\circ(q)}{A^\circ(q)F(q,\theta_N^k)}(A(q,\hat{\eta}_N) - A^\circ(q))u_t \\
&\quad + \frac{A^\circ(q)}{A^\circ(q)F(q,\theta_N^k)}(B(q,\hat{\eta}_N) - B^\circ(q))u_t \\
&= \frac{1}{F^\circ(q)} \Xi^m(\hat{\eta}_N, \theta_N^k) u_t.
\end{aligned}$$

We can thus express x_N as

$$x_N = \frac{1}{\sqrt{N}} \sum_{t=m+1}^{N} \Phi^m(\hat{\eta}_N, \theta_N^k) u_t \Xi^m(\hat{\eta}_N, \theta_N^k) u_t.$$

We will in the remainder of the proof need some properties regarding the filters Φ^m and Ξ^m that are easily established using Lemma 3.1:

$$\left\|\Xi^m(\hat{\eta}_N, \theta_N^k)\right\| = \mathcal{O}(m(N)) \quad \text{(D.2)}$$
$$\left\|\Phi^m(\hat{\eta}_N, \theta_N^k) - \Phi^m(\hat{\eta}_N, \theta^\circ)\right\| = \mathcal{O}(m(N)) \quad \text{(D.3)}$$
$$\left\|\Phi^m(\hat{\eta}_N, \theta^\circ) - \Phi^m(\eta^\circ, \theta^\circ)\right\| = \mathcal{O}(m(N)) \quad \text{(D.4)}$$
$$\left\|\Xi^m(\hat{\eta}_N, \theta_N^k) - \Xi^m(\hat{\eta}_N, \theta^\circ)\right\| = \mathcal{O}(m^2(N)) \quad \text{(D.5)}$$
$$\left\|\Phi^m(\eta^\circ, \theta^\circ)\right\| = \mathcal{O}(1) \quad \text{(D.6)}$$

For future reference, we will establish the limit of $\sqrt{N}m^2(N)$. The dominating term in $m(N)$ is $n(N)\sqrt{\log N/N}$ and terms with $d(N)$ will be neglected. For N large enough, we have

$$\begin{aligned}
\lim_{N\to\infty} \sqrt{N} m^2(N) &= \lim_{N\to\infty} \sqrt{N} n(N)^2 \frac{\log N}{N} \\
&= \lim_{N\to\infty} \left(\frac{n(N)^{4+\delta}}{N}\right)^{\frac{2}{4+\delta}} \frac{\log N}{N^{\frac{\delta}{4+\delta}}} = 0,
\end{aligned}$$

where the first term goes to zero by Assumption 3.1.

Using Lemma C.1 and Lemma C.3 with (D.2) and (D.3), it follows that difference between x_N and

$$\frac{1}{\sqrt{N}} \sum_{t=m+1}^{N} \Phi^m(\hat{\eta}_N, \theta_\circ) u_t \Xi^m(\hat{\eta}_N, \theta_N^k) u_t \quad \text{(D.7)}$$

tend to zero as $N \to \infty$ w.p.1, and therefore they have the same asymptotic distribution and the same asymptotic covariance. We will analyze (D.7) instead. Similarly, using Lemma C.1 and Lemma C.3 with (D.2) and (D.4), it follows that difference between (D.7) and

$$\frac{1}{\sqrt{N}} \sum_{t=m+1}^{N} \Phi^m(\eta^\circ, \theta_\circ) u_t \Xi^m(\hat{\eta}_N, \theta_N^k) u_t \quad \text{(D.8)}$$

tend to zero as $N \to \infty$ w.p.1, and we will analyze (D.8) instead. Similarly, using Lemma C.1 and Lemma C.3 with (D.5) and (D.6), the difference between (D.8) and

$$\frac{1}{\sqrt{N}} \sum_{t=m+1}^{N} \Phi^m(\eta^\circ, \theta_\circ) u_t \Xi^m(\hat{\eta}_N, \theta^\circ) u_t \quad \text{(D.9)}$$

tend to zero as $N \to \infty$ w.p.1, and we will analyze (D.9) instead.

We rewrite $\Xi^m(\hat{\eta}_N, \theta^\circ) u_t$ as

$$\begin{aligned}
\Xi^m(\hat{\eta}_N, \theta^\circ) u_t &= \frac{1}{A^\circ(q)} \begin{bmatrix} -B^\circ(q) u_t \Gamma_n \\ A^\circ(q) u_t \Gamma_n \end{bmatrix}^\top (\hat{\eta}_N - \bar{\eta}^n) \\
&= \frac{1}{A^\circ(q)} \varphi_t^n(\eta_\circ, \theta^\circ)^\top (\hat{\eta}_N - \bar{\eta}^n). \quad \text{(D.10)}
\end{aligned}$$

Thus, we have shown that x_N has the same distribution and covariance as

$$T_N := Z^n \sqrt{N}(\hat{\eta}_N - \bar{\eta}^n), \tag{D.11}$$

where

$$Z^n = \sum_{t=m+1}^{N} \varphi_t^m(\eta_\circ, \theta_\circ) \frac{F^\circ(q)}{A^\circ(q)} \varphi_t^n(\eta_\circ, \theta_\circ)^\top, \tag{D.12}$$

and we will analyze T_N instead.

D.2 Asymptotic covariance of T_N

Using Lemma C.2, we have that

$$T_N \sim \mathrm{AsN}(0, Q),$$

where

$$Q = \sigma_\circ^2 \lim_{n\to\infty} Z^n [\bar{R}^n]^{-1} (Z^n)^\top, \tag{D.13}$$

provided the right hand side limit exists. This will be shown next. We start by analyzing $\bar{R}^n$.

$$\begin{aligned} \bar{R}^n &= \mathrm{E}\left[\varphi_t^n (\varphi_t^n)^\top\right] \\ &= \langle \Psi, \Psi \rangle, \end{aligned} \tag{D.14}$$

where

$$\langle f, g \rangle \coloneqq \int_{-\pi}^{\pi} f(e^{j\omega}) g(e^{j\omega})^* \, \mathrm{d}\omega,$$

and with Ψ given by

$$\Psi = \begin{bmatrix} -G^\circ \Gamma_n & H^\circ \Gamma_n \\ \Gamma_n & 0_{n\times 1} \end{bmatrix} U_\circ$$

and $U_\circ$ is a spectral factor of the the covariance matrix of the input u_t and the noise e_t, given by

$$U_\circ = \begin{bmatrix} F_u & 0 \\ 0 & \sigma_\circ \end{bmatrix}.$$

For (D.12), we have that

$$\begin{aligned} Z^n &= \mathrm{E}\left[\varphi_t^m(\eta_\circ, \theta_\circ) \frac{F^\circ(q)}{A^\circ(q)} \varphi_t^n(\eta_\circ, \theta_\circ)^\top\right] \\ &= \mathrm{E}\left[\begin{bmatrix} -\frac{B^\circ}{F^\circ} \Gamma_m u_t \\ \frac{A^\circ}{F^\circ} \Gamma_m u_t \end{bmatrix} \begin{bmatrix} -G^\circ \Gamma_n u_t \\ \Gamma_n u_t \end{bmatrix}^\top\right] \\ &= \left\langle \begin{bmatrix} -\frac{G^\circ}{F^\circ H^\circ} \Gamma_m & 0_{n\times 1} \\ \frac{1}{F^\circ H^\circ} \Gamma_m & 0_{n\times 1} \end{bmatrix} F_u, \begin{bmatrix} -G^\circ \Gamma_n & 0_{n\times 1} \\ \Gamma_n & 0_{n\times 1} \end{bmatrix} F_u \right\rangle \\ &= \langle \gamma, \Psi \rangle, \end{aligned} \tag{D.15}$$

with

$$\gamma = \begin{bmatrix} -\frac{G^\circ}{F^\circ H^\circ} \Gamma_m & 0_{m\times 1} \\ \frac{1}{F^\circ H^\circ} \Gamma_m & 0_{m\times 1} \end{bmatrix} F_u,$$

where the last equality is due to the fact that the added column in the right argument of the inner product is multiplied by the zero column in γ when the inner product is taken. Hence, we can write the asymptotic covariance matrix of T_N as

$$\begin{aligned} \lim_{N\to\infty} \mathrm{E}\left[T_N T_N^\top\right] &= \sigma_\circ^2 \langle \gamma, \Psi \rangle \langle \Psi, \Psi \rangle^{-1} \langle \Psi, \gamma \rangle \\ &= \sigma_\circ^2 \langle \mathbf{P}_{\mathcal{S}_\Psi}[\gamma], \mathbf{P}_{\mathcal{S}_\Psi}[\gamma] \rangle, \end{aligned} \tag{D.16}$$

where $\mathcal{S}_\Psi$ is the subspace in $\mathcal{L}_2^{1\times 2}$ spanned by the rows of Ψ. Lemma C.4 gives that, as $n \to \infty$, $S_\gamma \subseteq \mathcal{S}_\Psi$ and

$$\lim_{N\to\infty} \mathrm{E}\left[T_N T_N^\top\right] = \sigma_\circ^2 \langle \gamma, \gamma \rangle = \sigma_\circ^2 M_{CR}.$$

D.3 Summing up

Consider T_N defined in (D.11). As observed in Section D.2, it follows from Lemma C.2 that

$$T_N \sim \mathrm{AsN}(0, \sigma_\circ^2 M_{CR}). \tag{D.17}$$

The asymptotic normality of $\sqrt{N}(\hat{\theta}_N - \hat{\theta}_\circ)$ follows from (D.1) and (D.17), together with that $\sqrt{N}(\hat{\theta}_N - \hat{\theta}_\circ)$ has the same asymptotic distribution as T_N. From (D.1) and (D.17), it now follows that

$$\sqrt{N}(\hat{\theta}_N^k - \theta_\circ) \sim \mathrm{AsN}(0, \sigma_\circ^2 M_{CR}^{-1}). \tag{D.18}$$

References

[1] A. G. Evans and R. Fischl. Optimal least squares time-domain synthesis of recursive digital filters. *IEEE Transactions on Audio and Electroacoustics*, 21(1):61–65, 1973.

[2] M. Galrinho, C. R. Rojas, and H. Hjalmarsson. A weighted least-squares method for parameter estimation of structured models. In *53rd IEEE Conference on Decision and Control*, pages 3322–3327, 2014.

[3] M. Galrinho, C. R. Rojas, and H. Hjalmarsson. On estimating initial conditions in unstructured models. In *54th IEEE Conference on Decision and Control*, pages 2725–2730, 2015.

[4] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer texts in statistics. Springer, New York, 1998.

[5] L. Ljung. *System Identification. Theory for the User, 2nd ed.* Prentice-Hall, 1999.

[6] L. Ljung and B. Wahlberg. Asymptotic properties of the least-squares method for estimating transfer functions and disturbance spectra. *Adv. Appl. Prob.*, 24:412–440, 1992.

[7] J. H. McClellan and D. Lee. Exact equivalence of the Steiglitz-McBride iteration and IQML. *IEEE Transactions on Signal Processing*, 39(2):509–5012, 1991.

[8] J. Schoukens, Y. Rolain, G. Vandersteen, and R. Pintelon. User friendly Box-Jenkins identification using nonparametric noise models. In *50th IEEE Conference on Decision and Control*, Orlando, Florida, USA, 2011.

[9] A. K. Shaw. Optimal identification of discrete-time systems from impulse response data. *IEEE Transactions on Signal Processing*, 42(1):113–120, 1994.

[10] T. Söderström and P. Stoica. *Instrumental Variable Methods for System Identificaiton*. Springer Verlag, New York, 1983.

[11] T. Söderström and P. Stoica. *System identification*. Prentice-Hall, Inc., 2001.

[12] K. Steiglitz and L. E. McBride. A technique for the identification of linear systems. *IEEE Transactions on Automatic Control*, 10:461–464, 1965.

[13] P. Stoica and M. Jansson. MIMO system identification: State-space and subspace approximations versus transfer function and instrumental variables. *IEEE Transactions on Signal Processing*, 48(11):3087–3099, 2000.

[14] P. Stoica and T. Söderström. The Steiglitz-McBride identification algorithm revisited—convergence analysis and accuracy aspects. *IEEE Transactions on Automatic Control*, 26(3):712–717, 1981.

[15] P. van Overschee and B. de Moor. *Subspace identification for linear systems: theory, implementation, applications*. Kluwer Academic Publishers, Boston, 1996.

[16] B. Wahlberg. Model reduction of high-order estimated models: the asymptotic ML approach. *International Journal of Control*, 49(1):169–192, 1989.

[17] Y. Zhu. *Multivariable System Identification for Process Control*. Pergamon, 2001.

[18] Y. Zhu and H. Hjalmarsson. The Box-Jenkins Steiglitz-McBride algorithm. *Automatica*, 65:170–182, 2016.