Open-loop asymptotically efficient model reduction with the Steiglitz-McBride method [⋆]

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 use a physical intuition when choosing 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 can 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 risk of getting trapped in local minima. Here, we consider the following alternative approach. First, estimate a high-order ARX model with least squares, providing non-parametric estimates of the plant and noise model. Second, reduce the high-order model to obtain a parametric model of the plant only. We review existing methods to do 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 suggesting that the proposed method is competitive with state-of-the-art 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 [13]. 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 [27] or instrumental variable methods [20], are appealing for their low computational complexity, and are hence useful to initialize PEM. However, they are in general not as accurate as PEM, although multistep or iterative versions of IV methods can be asymptotically efficient [21, 31].

It is also possible to apply PEM to a more flexible model than the one of interest, and then perform model order reduction. With indirect PEM [23], the model-reduction step is based on a maximum likelihood cost function. In some settings, this procedure is advantageous with respect to a direct PEM estimation (see [23] for examples).

However, for settings with output-error or Box-Jenkins models, the more flexible model must be taken as non-parametric (i.e., arbitrarily large order). In general, this can be taken 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 [28], but solving a non-convex optimization problem is still needed in general.

This approach differs from the setting in [23] because, for a given order, the non-parametric model does not contain the true system. To analyze this type of approach theoretically, it is therefore instrumental to let the order depend on the sample size [14]; in particular, the order has to tend to infinity with some maximum rate to achieve consistency and asymptotic efficiency.

The model order reduction need not necessarily be done on the high-order model itself, but the residuals of this

[⋆] This work was supported by the Swedish Research Council under contracts 015-05285 and 2016-06079. 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).

model can be used in a second stage to estimate the low order model. This idea dates back to [2]. For the class of ARMAX models, the method was complemented with the proper filtering for efficiency in [15] and letting the high-order model order depend on the data [9].

Model-order selection and estimation based on ML has a long history (e.g., [1, 8, 29]). One classical approach is to estimate the model orders from data. For ARMAX models, one iterative procedure is the Hannan-Rissanen-Kavalieris type methods [8, 10, 11]. These methods do not use an intermediate high-order model; instead, at each iteration, they estimate the innovations and select new model orders according to an information criterion.

Another possibility to perform model order reduction from a high-order non-parametric model is with the weighted null-space fitting (WNSF) method [6]. Although it can be motivated by an exact ML criterion [28], this criterion is not minimized explicitly. Rather, it is interpreted as a weighted least squares problem by fixing the parameters in the weighting.

While the plant model order can sometimes be based on physical intuition, the noise model order is usually a more abstract concept. In [18], 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, the authors call it "user-friendly".

If the data are obtained in open loop, the asymptotic properties of the plant and noise-model estimates obtained with PEM are uncorrelated if the two transfer functions are independently parametrized [13, 17]. 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.

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 [28]. 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 [37] is based on this approach.

Another approach that does not require a parametric noise model is the BJSM method [38]. This method uses a non-parametric ARX model to extend the applicability of the Steiglitz-McBride method [24] to colored noise settings. BJSM uses the ARX model to create a pre-filtered data set for which the output noise is approximately white, and the Steiglitz-McBride method is applied to the pre-filtered data set. In [38], it is shown that this procedure is asymptotically efficient in open loop. However, consistency has only been established when the number of Steiglitz-McBride iterations tends to infinity.

In this paper, we start from an asymptotic ML criterion to propose a method that uses the Steiglitz-McBride method instead of non-convex optimization algorithms, but with improved convergence properties compared with BJSM. Our contributions are the following. First, we propose the new method and contextualize it with other related methods. Second, we perform a theoretical analysis, showing that the proposed method is consistent and asymptotically efficient in open loop with one Steiglitz-McBride iteration. This analysis is rather elaborate due to the necessity, as mentioned earlier, to let the ARX-model order depend on the sample size. Third, we perform a simulation study, where we observe that the proposed method has better finite sample convergence properties than BJSM, and that it may be a viable alternative to other competitive methods.

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. These signals are related by*

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

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

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

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, with $\{w_t\}$ 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 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*

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

for some positive finite constant C.

The assumption that the expected value of the tenth moment is bounded is stronger than what is required for the analysis of PEM [13]. This assumption can be made weaker for some of our theoretical results (see [14] for details), but for simplicity we take a sufficiently strong requirement that will apply for all theoretical results.

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 as

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

where

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

with parameter vectors $\theta = [f_1 \ \dots \ f_{m_f} \ l_1 \ \dots \ l_{m_l}]^\top$ and $\alpha = [c_1 \ \dots \ c_{m_c} \ d_1 \ \dots \ d_{m_d}]^\top$. 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 known (i.e., $m_f = m_f^\circ$ and $m_l = m_l^\circ$). For notational simplicity, we let $m := m_f = m_l$.

The one step ahead prediction errors of the BJ model (2) are given by [13]

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

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

$$V_N(\theta,\alpha) = \frac{1}{N}\sum_{t=1}^{N}\varepsilon_t^2(\theta,\alpha), \tag{3}$$

where N is the number of data samples. We denote by $\hat{\theta}_N^{\mathrm{PEM}}$ the estimate of θ obtained by minimizing (3). Moreover, $\theta_\circ$ corresponds to the vector θ evaluated at the coefficients of $F^\circ(q)$ and $L^\circ(q)$. The cost function (3) is non-convex, and may require good intialization points to converge to the global minimum.

Because the system operates in open loop (Assumption 2.2), it is well known that, when PEM is applied to the model (2), under weak conditions, the (normalized) error of the estimated parameters $\hat{\theta}_N^{\mathrm{PEM}}$ converges to the Gaussian distribution [13]

$$\sqrt{N}\left(\hat{\theta}_N^{\mathrm{PEM}} - \theta_\circ\right) \in \mathcal{N}\left(0, \sigma_\circ^2 M_{\mathrm{CR}}^{-1}\right), \quad \text{as } N \to \infty, \tag{4}$$

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 $\Gamma_m(q) = \left[\, q^{-1} \ \dots \ q^{-m} \,\right]^\top$ 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 [13]. 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 in principle be chosen arbitrarily large, guaranteeing that $H^\circ(q)$ is in the model set. However, this makes the problem numerically more challenging, as we increase the number of parameters in the non-convex cost function (3).

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

$$A^\circ(q)y_t = B^\circ(q)u_t + e_t, \tag{5}$$

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

$$A(q,\eta^n) = 1 + \sum_{k=1}^{n} a_k q^{-k}, \quad B(q,\eta^n) = \sum_{k=1}^{n} b_k q^{-k}, \quad (6)$$

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. Because $\{a_k^\circ\}$ and $\{b_k^\circ\}$ are exponentially decaying, (6) can model (5) with good accuracy if n is chosen large enough.

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

$$V_N(\eta^n) = \frac{1}{N}\sum_{t=1}^{N} \left[A(q,\eta^n)y_t - B(q,\eta^n)u_t\right]^2, \quad (7)$$

which can be done by linear least squares. Thus, the estimate of η^n is given by

$$\hat{\eta}_N^{n,ls} := [R_N^n]^{-1} r_N^n, \quad (8)$$

where

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

with

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

In the analysis, we will use the slightly modified estimate

$$\hat{\eta}_N^n := [R_{N,\text{reg}}^n]^{-1} r_N^n, \quad (10)$$

where

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

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 [14]. It follows from Assumption 2.2 and Assumption 2.3 (see [14] for details) that

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

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

To guarantee that the true system (5) is asymptotically in the model set defined by the ARX model (6), 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 (ARX-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 $\delta > 0$.

We define $\hat{\eta}_N := \hat{\eta}_N^{n(N)}$ and, for future reference,

$$\eta_\circ^n := \begin{bmatrix} a_1^\circ \dots a_n^\circ \; b_1^\circ \dots b_n^\circ \end{bmatrix}^\top, \quad (11)$$
$$\eta_\circ := \begin{bmatrix} a_1^\circ \; a_2^\circ \dots b_1^\circ \; b_2^\circ \dots \end{bmatrix}^\top. \quad (12)$$

The asymptotic properties of $\hat{\eta}_N$ have been established in [14]. 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

$$m(N) = n(N)\sqrt{\log N/N}(1+d(N)) + d(N),$$
$$d(N) := \sum_{k=n(N)+1}^{\infty} |a_k^\circ| + |b_k^\circ| \le \bar{C}\rho^{n(N)}, \quad (13)$$

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

$$G(q,\hat{\eta}_N) := \frac{B(q,\hat{\eta}_N)}{A(q,\hat{\eta}_N)}, \quad H(q,\hat{\eta}_N) := \frac{1}{A(q,\hat{\eta}_N)}. \qquad (14)$$

Despite the simplicity of ARX models, they are not appropriate to model (2.1) for most practical uses: because the order n may have to be chosen large, the estimated model will have 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, as long as the order n tends to infinity according to Assumption 3.1: in this way, the estimate $\hat{\eta}_N$ and its covariance become a sufficient statistic [12] for our problem as the number of data samples increases to infinity. Thus, the data could in principle be disregarded, and $\hat{\eta}_N$ and the respective covariance 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 Maximum Likelihood

Because, as the ARX-model order increases, $\hat{\eta}_N$ and its covariance approach a sufficient statistic, they can be used to obtain an estimate of θ that is asymptotically efficient. This can be done using an exact ML criterion [28]. Let $\eta^n(\theta,\alpha)$ be the truncated parameter vector η^n obtained from θ and α, satisfying the relations

$$A(q,\eta) = \frac{1}{H(q,\alpha)}, \quad B(q,\eta) = \frac{G(q,\theta)}{H(q,\alpha)}. \qquad (15)$$

This procedure consists in minimizing

$$[\hat{\eta}_N - \eta^n(\theta,\alpha)]^\top [\operatorname{cov}(\hat{\eta}_N)]^{-1} [\hat{\eta}_N - \eta^n(\theta,\alpha)], \qquad (16)$$

where $\operatorname{cov}(\hat{\eta}_N)$ denotes the covariance of the estimated vector $\hat{\eta}_N$. This cost function can be approximated by the asymptotic ML criterion

$$\int_0^{2\pi} \left|G(e^{i\omega},\hat{\eta}_N) - G(e^{i\omega},\theta)\right|^2 \frac{\Phi_u(e^{i\omega})}{|H(e^{i\omega},\hat{\eta}_N)|^2} d\omega + \frac{\hat{\sigma}^2}{2\pi}\int_0^{2\pi} \frac{\left|H(e^{i\omega},\hat{\eta}_N) - H(e^{i\omega},\alpha)\right|^2}{|H(e^{i\omega},\hat{\eta}_N)|^2} d\omega, \qquad (17)$$

where $\hat{\sigma}^2$ is a consistent estimate of $\sigma_\circ^2$, without changing the asymptotic statistical properties of the estimate [28]. Moreover, the first term in (17) is only dependent on $G(q,\theta)$ and the second term on $H(q,\alpha)$. Therefore, $G(q,\theta)$ can be estimated independently of $H(q,\alpha)$ by minimizing

$$V_N(\theta) = \int_0^{2\pi} \left|G(e^{i\omega},\hat{\eta}_N^n) - G(e^{i\omega},\theta)\right|^2 \frac{\Phi_u(e^{i\omega})}{|H(e^{i\omega},\hat{\eta}_N)|^2} d\omega. \qquad (18)$$

The idea of the ASYM method [37] is to minimize the time domain equivalent to (18) for finite sample size:

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} \left[\left(\frac{B(q,\hat{\eta}_N)}{A(q,\hat{\eta}_N)} - G(q,\theta)\right) A(q,\hat{\eta}_N) u_t\right]^2. \qquad (19)$$

Minimizing (19) is still a non-convex optimization problem. However, it is pointed out in [37] 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 (19), and the noise contribution is only present indirectly through the high-order estimates, the influence of the disturbance is reduced.

4.2 BJSM method

The BJSM method can be seen as an extension of the Steiglitz-McBride algorithm [24]. The latter consists in using iterative least squares to estimate $L(q,\theta)$ and $F(q,\theta)$ when the transfer function $H^\circ(q) = 1$. Then, the idea of BJSM is to extend the applicability of the Steiglitz-McBride method to other noise spectra, using a pre-filtering step by a high-order ARX model.

We start by reviewing the Steiglitz-McBride method. Consider the following three steps. First, an ARX model

$$F(q,\theta)y_t = L(q,\theta)u_t + e_t \qquad (20)$$

is estimated with least squares, providing an initialization estimate $\hat{\theta}_N^0$. Second, the data are filtered by

$$y_t^f = \frac{1}{F(q,\hat{\theta}_N^0)} y_t, \qquad u_t^f = \frac{1}{F(q,\hat{\theta}_N^0)} u_t.$$

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.

The motivation for the Steiglitz-McBride algorithm is the following. Let the estimate obtained at iteration k

by $\hat{\theta}_N^k$. At iteration $k+1$, we minimize the cost function

$$V_N(\theta_N^{k+1}) = \frac{1}{N}\sum_{t=1}^{N}\left[\frac{F(q,\theta^{k+1})}{F(q,\hat{\theta}_N^k)}y_t - \frac{L(q,\theta^{k+1})}{F(q,\hat{\theta}_N^k)}u_t\right]^2 . \quad (21)$$

If the estimates converge to a consistent estimate of $\theta_\circ$, this cost function corresponds to minimizing a quadratic cost function of prediction errors, as with (3).

Convergence of the Steiglitz-McBride has been studied in [26], 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 the Cramér-Rao bound M_{CR}^{-1}.

The Box-Jenkins Steiglitz-McBride (BJSM) algorithm [38] consists of an extension of Steiglitz-McBride that is consistent for colored noise and is asymptotically efficient for open loop data. The method uses the following procedure. First, a high-order ARX model (6) is estimated with least squares. Second, the original data set is pre-filtered by $A(q,\hat{\eta}_N)$, according to

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

Third, the Steiglitz-McBride algorithm is applied to the pre-filtered data set.

The motivation for this procedure is that the pre-filtered data satisfy

$$y_t^{\text{pf}} = \frac{L^\circ(q)}{F^\circ(q)}u_t^{\text{pf}} + A(q,\hat{\eta}_N)H^\circ(q)e_t, \quad (22)$$

which asymptotically is according to (Lemma 3.1)

$$y_t^{\text{pf}} \approx \frac{L^\circ(q)}{F^\circ(q)}u_t^{\text{pf}} + e_t. \quad (23)$$

Because the noise in (23) is white, the Steiglitz-McBride method is convergent with the data set $\{y_t^{\text{pf}}, u_t^{\text{pf}}\}$.

If we were to apply PEM to the pre-filtered data set, we would minimize, motivated by (23),

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N}\left(y_t^{\text{pf}} - \frac{L(q,\theta)}{F(q,\theta)}u_t^{\text{pf}}\right)^2 . \quad (24)$$

To avoid an explicit non-convex minimization problem, BJSM uses the Steiglitz-McBride method instead.

Although the BJSM method is asymptotically efficient in open loop [38], not all the information in $\hat{\eta}_N$ is being used, as the filtering (22) 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 $H^\circ(q) = 1$. Then, we would have that $A^\circ(q) = 1$, and only $B(q,\eta^n)$ would need to be estimated in order to model the true system. However, this would maintain the data set unchanged when applying the filtering (22), 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 $B(q,\eta^n)$ should be sufficient to model system with additive white noise (as n increases), it is not possible to make use of this information when applying the BJSM method, because it does not use the full statistical properties of the high-order model.

The second limitation is that, although BJSM avoids solving a non-convex optimization problem, it requires the number of Steiglitz-McBride iterations to tend to infinity in order to provide consistent and asymptotically efficient estimates [38].

5 Model Order Reduction Steiglitz-McBride

Similarly to a direct minimization of (18) or BJSM, the objective of our approach is to obtain an estimate of $G(q,\theta)$ from the high-order ARX model estimate without having to estimate $H(q,\alpha)$. However, we would also like to do so without using a non-convex optimization method and in a finite number of steps.

We start by re-writing (19), the time-domain approximation of (18), as

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N}\left[B(q,\hat{\eta}_N)u_t - \frac{L(q,\theta)}{F(q,\theta)}A(q,\hat{\eta}_N)u_t\right]^2 . \quad (25)$$

Then, we define

$$\hat{y}_t^{\text{pf}} := B(q,\hat{\eta}_N)u_t, \qquad u_t^{\text{pf}} := A(q,\hat{\eta}_N)u_t. \quad (26)$$

With this definition, (25) has the same form as (24) if we replace y_t^{pf} by $\hat{y}_t^{\text{pf}}$. Thus, similarly to BJSM, we may apply the Steiglitz-McBride method to an alternative data set instead of minimizing (25) with a non-convex optimization algorithm; the difference is that we use the data set $\{\hat{y}_t^{\text{pf}}, u_t^{\text{pf}}\}$ instead of $\{y_t^{\text{pf}}, u_t^{\text{pf}}\}$

The proposed method is as follows:

(1) estimate an ARX model using the input-output data $\{u_t, y_t\}$, $t = 1, \ldots, N$, according to (8);
(2) construct the pre-filtered data $\{\hat{y}_t^{\text{pf}}, u_t^{\text{pf}}\}$, according to (26);
(3) apply the Steiglitz-McBride method with $\{\hat{y}_t^{\text{pf}}, u_t^{\text{pf}}\}$ to obtain estimates $L(q, \hat{\theta}_N)$ and $F(q, \hat{\theta}_N)$ of $L^\circ(q)$ and $F^\circ(q)$, respectively.

Because this method can be seen as a way of applying the Steiglitz-McBride algorithm to reduce a high-order model to a parametric one, we will refer to it as Model Order Reduction Steiglitz-McBride (MORSM).

In terms of the algorithm, comparing (26) and (22) shows that the difference between this approach and BJSM is in the pre-filtered output only: here, the pre-filtered output is simulated from the input and the ARX-model estimate, depending on the original output data $\{y_t\}$ only through the least squares estimate $\hat{\eta}_N$.

Although the difference between the algorithms is minimal, the methods have different motivations. BJSM follows from extending the Steiglitz-McBride method to be consistent with colored noise, using only part of the high-order ARX model for that purpose. MORSM follows from observing that the Steiglitz-McBride can be applied to an asymptotic ML cost function, where all the information in the high-order ARX model is used. This provides two advantages with respect to BJSM, which will be formally proven in the next section, but we introduce in the following paragraphs.

First, the pre-filter (26) uses the complete statistical information contained in the estimate $\hat{\eta}_N$. So, if the noise contribution affecting the true system is white, a high-order FIR model can be estimated instead of an ARX. In this case, $A(q, \hat{\eta}_N) = 1$. This was not the case with BJSM, for which $A(q, \hat{\eta}_N^n)$ must always be estimated.

Second, this procedure only requires one iteration to provide an efficient estimate. To intuitively understand why this is the case, we recall that the Steiglitz-McBride is an iterative method that only after successive iterations minimizes (3). By continuing to iterate, it can be shown that, under the conditions observed in [25], $\hat{\theta}_N^k \to \theta_\circ$, as $k \to \infty$ and $N \to \infty$. Concerning the BJSM method, since the pre-filtered data is according to (22), it is asymptotically approximately an OE model structure, and a similar procedure takes place. On the other hand, the proposed pre-filtered data set, which does not use the original data, satisfies

$$\hat{y}_t^{\text{pf}} = \frac{L^\circ(q)}{F^\circ(q)} u_t^{\text{pf}} + \left(\frac{B(q, \hat{\eta}_N)}{A(q, \hat{\eta}_N)} - \frac{L^\circ(q)}{F^\circ(q)} \right) u_t^{\text{pf}}. \quad (27)$$

This is a noise-free equation, except for the noisy parameters in the ARX model. However, from Lemma 3.1, the second term in (27) tends to zero asymptotically (in N). As consequence, the variance of the error sequence being minimized by the Steiglitz-McBride iterations disappears asymptotically, and only one iteration is required.

5.1 Noise Model

One of the advantages of MORSM is that it does not require a noise-model order selection. However, if a low-order noise model is required, it can be estimated independently of the plant model using a similar procedure, starting from the second term in (17). Minimizing this term is the same as minimizing, in the time domain,

$$V_N^H(\alpha) = \frac{1}{N} \sum_{t=1}^{N} \left[e_t - \frac{C(q, \alpha)}{D(q, \alpha)} A(q, \hat{\eta}_N) e_t \right]^2. \quad (28)$$

We recognize that (28) has the same form as (24) if we replace y_t^{pf} and u_t^{pf} by

$$\tilde{y}_t^{\text{pf}} := e_t, \qquad \tilde{u}_t^{\text{pf}} := A(q, \hat{\eta}_N) e_t, \quad (29)$$

respectively, and let $C(q, \alpha)$ and $D(q, \alpha)$ play the role of $L(q, \theta)$ and $F(q, \theta)$. The difference here is that $\{e_t\}$ is not known. However, it may be replaced by an estimate based on the high-order model,

$$\hat{e}_t := A(q, \hat{\eta}_N) y_t - B(q, \hat{\eta}_N) u_t. \quad (30)$$

Alternatively, the products $e_i e_j$ in the least-squares equations may be replaced by a scaling of the expected value (i.e., $\mathbb{E}[e_i e_j] = 0$ for $i \neq j$ and $\mathbb{E}[e_i e_i] = 1$).

5.2 Relation to other methods

As discussed in previous sections, the proposed method builds on a family of methods that use high-order ARX models and an auxiliary step to obtain the model of interest; mainly, [28, 37]. Instead of performing model reduction on the high-order ARX model, this model can be used to estimate the residuals from which a low-order model is estimated [2, 9, 15]. The method has also close similarities with BJSM, which also uses the Steiglitz-McBride algorithm. In contrast with the proposed method, these approaches keep the original data, while MORSM does not. Thus, MORSM can be seen as a model-reduction method using Steiglitz-McBride.

The idea of using 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 (e.g., [3, 16, 19]). Although some of these procedures are optimal under specific conditions, here we consider a quite general system identification problem for which application of the Steiglitz-McBride algorithm provides asymptotically efficient estimates in one iteration.

MORSM shares also similarities with the Refined Instrumental Variables (RIV) method [30–34], as both use an iterative procedure with filtering according to the estimate update from the previous iteration. This observation has already been made in [38] for BJSM, and most of the similarities and differences observed apply for MORSM. However, MORSM is a model-reduction method, while BJSM and RIV are not, although RIV also computes a simulated output and it has been applied to reduce high-order dynamic simulation models with a technique called model emulation [35, 36].

The main differences are the following. First, MORSM uses least squares while RIV uses instrumental variables. Second, RIV iterates between the plant model estimate update and the noise model estimate update, while in MORSM the plant model estimate only depends on the noise model through the high-order estimate, which is the same for all iterations. Third, RIV, as mentioned earlier, computes a simulated output, but only to construct the instrument vector, as the measured output is still used; MORSM, on the other hand, discards the measured output and only uses simulated data after the high-order ARX model has been obtained. Fourth, with RIV the order of the instrument vector is kept fixed, while for MORSM the order of the ARX model needs to tend to infinity for efficiency to be obtained; on the other hand, RIV (unlike other IV algorithms) exploits a multiple-iteration procedure to guarantee optimal statistical properties (see [30] for details on this discussion), while MORSM is asymptotically efficient in one iteration, as we proceed to show in the following section.

6 Asymptotic Properties

In this section, we analyze 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.

With this purpose, we start by defining

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

and

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

which also applies to vector valued signals such as (9). Then, using (26), we have that

$$u_t = \frac{1}{B(q,\hat{\eta}_N)}y_t^{\text{pf}} = \frac{L^\circ(q)A^\circ(q)}{F^\circ(q)B^\circ(q)}\frac{1}{A(q,\hat{\eta}_N)}u_t^{\text{pf}}, \quad (31)$$

which we filter by

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

where $\hat{\theta}_N^k$ is the estimate obtained at iteration k of the Steiglitz-McBride algorithm (the step (20) to obtain the initialization estimate $\hat{\theta}_N^0$ corresponds to setting $F(q,\theta)\equiv 1$). From here, we write 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

$$y_t(\hat{\eta}_N,\hat{\theta}_N^k) = [\varphi^m(\hat{\eta}_N,\hat{\theta}_N^k)]^\top\theta_\circ + F^\circ(q)\xi_t(\hat{\eta}_N,\hat{\theta}_N^k). \quad (32)$$

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 (32):

$$\hat{\theta}_N^{k+1} = [R^m(\hat{\eta}_N,\hat{\theta}_N^k)]^{-1}r^m(\hat{\eta}_N,\hat{\theta}_N^k), \quad (33)$$

where

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

Having written (32) in the regression form (27) will be instrumental for our analysis, because the error made in the ARX model appears explicitly and linearly in $\xi_t(\hat{\eta}_N,\hat{\theta}_N^k)$, which tends to zero according to Lemma 3.1.

Regarding consistency, we have the following theorem.

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.

Regarding the asymptotic distribution and covariance, we have the following theorem.

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

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

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

PROOF. See Appendix D.

Theorem 6.1 implies that the initial estimate $\hat{\theta}_N^0$ is a consistent estimate of $\theta_\circ$. Moreover, Theorem 6.2 implies that MORSM has the same asymptotic covariance as PEM with Gaussian noise (4). Therefore, it is asymptotically efficient with open loop data, and asymptotic efficiency is obtained in one iteration, with $\hat{\theta}_N^1$. This is a main difference to the BJSM algorithm, for which consistency and asymptotically efficiency have been established only for $k \to \infty$ [38].

7 Practical Considerations

In the previous section, we showed that MORSM provides an asymptotically efficient estimate in one iteration if the ARX-model order tends to infinity according to Assumption 3.1. However, in practice (i.e., for finite sample size), we need to choose some value for the ARX-model order, and the estimate may improve by iterating. We now consider these practical choices.

7.1 ARX-model order

There is a trade-off in the choice of the ARX-model order: the larger the order is, the more accurately the system dynamics are captured, but the higher the estimation variance becomes. A too high variance in the high-order model may have an impact on the estimation variance of the low-order model.

For practical purposes, the user usually has a rough idea of how large this order should be chosen for the dynamics of the system to be captured sufficiently accurately, depending on whether the system has fast or slow dynamics. However, the ideal order depends also on the sample size, as the order should be made larger as more samples are available.

The following approach can be used to choose the high-order model order. First, choose a set of orders that may be appropriate to model the system, following initial knowledge about the speed of the system. Second, run MORSM for all these orders separately, obtaining low-order models corresponding to each of the high-order models. Third, choose the low-order model that minimizes the prediction-error criterion (3).

Recall that the purpose of MORSM and other similar methods is to attempt to attain the global minimizer of (3) without using non-convex methods. Hence, the choice of using (3) to distinguish between several model estimates is appropriate. However, it should be taken into account that MORSM does not necessarily estimate a low-order noise model to plug in (3). In this case, the highest-order of the non-parametric estimates can be used: although this is a very noisy estimate, the noise introduced will be the same for all the model estimates.

7.2 Iterations

Although MORSM provides an asymptotically efficient estimate in one iteration, for finite sample size the estimate may improve by iterating. Analogously to the choice of high-order, criterion (3) can also be used to choose the best estimate among all the iterations.

8 Simulations

In this section, we perform Monte Carlo simulations to study the performance of the method. First, we illustrate the advantages with respect to the BJSM algorithm. Second, we illustrate how the method can be appropriate to initialize PEM. Third, we perform comparisons with other methods using systems where PEM can have difficulties with local minima.

8.1 Advantages with respect to BJSM

Although using different motivations, the algorithms for MORSM and BJSM have close similarities. In [38], simulation studies have been performed with examples where BJSM can be an alternative to PEM when PEM has convergence problems. Here, we illustrate how BJSM and MORSM typically perform similarly given that the algorithms converge, but that MORSM converges faster. First, we illustrate how MORSM only requires one iteration for an asymptotically efficient estimate, while this is not sufficient for BJSM. Second, we illustrate how even when (for finite sample size) MORSM requires more than one iteration for convergence, it is still a faster converging method than BJSM.

8.1.1 Example 1: one iteration scheme

In the first simulation, we illustrate that MORSM, unlike BJSM, gives an asymptotically efficient estimate in one iteration. For the simulation, the data are generated by

$$y_t = \frac{q^{-1} + 0.1q^{-2}}{1 - 1.2q^{-1} + 0.6q^{-2}} u_t + \frac{1 + 0.7q^{-1}}{1 - 0.9q^{-1}} e_t. \tag{34}$$

Two hundred Monte Carlo simulations are performed with eight sample sizes equally spaced between $N = 200$ and $N = 20000$ on a logarithmic scale. The input $\{u_t\}$ is obtained by

$$u_t = \frac{1}{1 - q^{-1} + 0.89q^{-2}} w_t, \tag{35}$$

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

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 MORSM, an ARX model of order 50 is estimated in the first step. In the iterative versions, the estimate obtained in the last iteration is the one used. 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 Figure 1, where the average root mean square error (RMSE) of the impulse response is presented for each sample size. The RMSE is given by

$$\text{RMSE} = \sqrt{\text{E}\left[\|g^\circ - \hat{g}\|_2^2\right]}, \tag{36}$$

where g° is a vector with the first 50 impulse response coefficients of $G^\circ(q)$ and similarly for $\hat{g}$ for the estimated plant model. In Figure 1, we observe that MORSM and BJSM perform similarly with 100 iterations for all the sample sizes 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. Table 1 compares the parameter estimates of MORSM with 100 iterations associated with the results in Figure 1 for two chosen sample sizes. The sample standard deviation σ agrees with what is predicted by the asymptotic theory.

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

8.1.2 *Example 2: convergence speed*

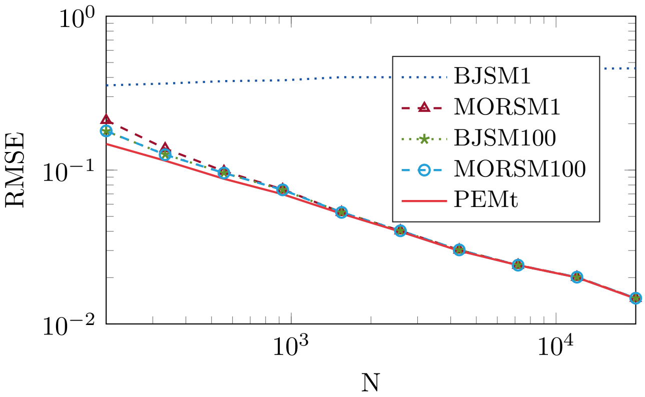

Fig. 1. Example 1: average RMSE as function of sample size for several methods, obtained from 200 Monte Carlo runs with a fixed system.

Table 1
Example 1: sample mean, sample standard deviation and theoretical standard deviation for MORSM with 100 iterations.

N	True values	-1.200	0.600	1.000	0.100
200	Sample mean	-1.195	0.594	1.000	0.109
	Sample σ	0.043	0.039	0.063	0.105
	Theoretical σ	0.033	0.031	0.050	0.080
20000	Sample mean	-1.200	0.600	1.000	0.100
	Sample σ	0.003	0.003	0.005	0.008
	Theoretical σ	0.003	0.003	0.005	0.008

In the following simulation, we will compare the performance of BJSM and MORSM with randomly generated systems, with structure

$$\begin{aligned} y_t = &\frac{l_1^\circ q^{-1} + l_2^\circ q^{-2} + l_3^\circ q^{-3} + l_4^\circ q^{-4}}{1 + f_1^\circ q^{-1} + f_2^\circ q^{-2} + f_3^\circ q^{-3} + f_4^\circ q^{-4}} u_t \\ &+ \frac{1 + c_1^\circ q^{-1} + c_2^\circ q^{-2} + c_3^\circ q^{-3} + c_4^\circ q^{-4}}{1 + d_1^\circ q^{-1} + d_2^\circ q^{-2} + d_3^\circ q^{-3} + d_4^\circ q^{-4}} e_t, \end{aligned} \tag{37}$$

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

$$\text{SNR} = \frac{\sum_{t=1}^N [G^\circ(q)u_t]^2}{\sum_{t=1}^N [H^\circ(q)e_t]^2} = 5. \tag{38}$$

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 an annulus 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 extremely dominant over others), they can be approximated by ARX models roughly of orders between

30 and 100, and they resemble physical systems.

We consider the following methods:

- the prediction error method, initialized at the true parameters (PEMt);
- the Box-Jenkins Steiglitz-McBride method (BJSM).
- the iterative Model Order Reduction Steiglitz-McBride method (MORSM);
- the iterative Model Order Reduction Steiglitz-McBride method, estimating also a noise model (MORSMh);
- the one-iteration Model Order Reduction Steiglitz-McBride, estimating also a low-order noise model (MORSM1h).

All the iterative methods perform a maximum of 1000 iterations. MORSM and BJSM have a stopping criterion of 10^{-4} as tolerance for the normalized norm of the last iteration (for PEM, the stopping criterion depends on the optimization algorithm used by MATLAB, which is set as automatic). For MORSM, we choose the ARX-model order from a grid of values between 25 and 125, spaced with intervals of 25. With PEM, we estimate initial conditions, and with MORSM and BJSM we truncate them. Although a procedure to estimate initial conditions for this type of methods has been proposed in [7], 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 for BJ models.

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

$$\text{FIT} = 100\left(1 - \frac{\text{RMSE}}{\|g^\circ - \bar{g}^o\|}\right), \tag{39}$$

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

The results are presented in Figure 2, with the average FIT as function of sample size. Unlike in Figure 1, this more challenging scenario does not allow to observe that (for this range of sample sizes) one iteration of MORSM provides an asymptotically efficient estimate. If we continue to iterate, there is an improvement in the obtained model estimate, and both MORSM (without low-order noise-model estimate) and BJSM perform similarly, attaining the performance of PEM (initialized at the true parameters) for this range of sample sizes.

The performance of MORSM also improves for small sample size if a noise model is estimated in the iterative version. The estimation of $H(q, \alpha)$, as pointed out in Section 5, is independent of the estimation of $G(q, \theta)$. The improvement observed in the model estimate is because having an estimate of $H(q, \alpha)$ allows for a more accurate computation of (3) when choosing the best model over all the iterations.

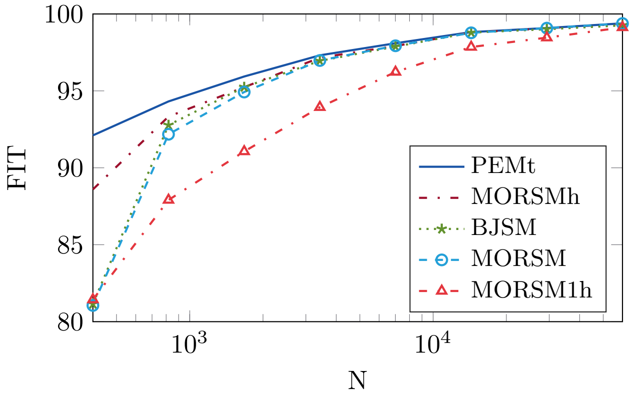

Fig. 2. Example 2: average FIT for several methods, obtained from 100 Monte Carlo runs with random systems.

The fact that, for finite sample size, both MORSM and BJSM require more than one iteration for convergence does not render MORSM useless with respect to BJSM. In Table 2, we indicate the average number of iterations required for MORSM and BJSM to converge, for the difference sample sizes used. From here, we conclude that, even when MORSM needs more than one iteration to converge, it still converges faster than BJSM. Moreover, BJSM needs approximately the same amount of iterations independently of sample size, while the number of iterations required for MORSM decreases with sample size. This is in accordance with our theoretical result that, asymptotically, MORSM provides an efficient estimate in one iteration.

8.2 Example 3: initialization for PEM

Here, we illustrate how MORSM can be an appropriate method to initialize PEM. For that, we repeat the simulation in Section 8.1.2 with the following methods:

- the prediction error method, with default MATLAB initialization (PEMd);
- the prediction error method, with MORSM1h as initialization (PEMm1);
- the prediction error method, with MORSMh as initialization (PEMm).

The results are presented in Figure 3, where we can compare how PEM performs with the default MATLAB initialization and with the MORSM initialization, as well as how much PEM can improve the MORSM estimate. For easier comparison, we also include PEMt and MORSMh from the previous simulation.

In Figure 3, we see that the standard MATLAB initialization for PEM is not always accurate enough to find the global minimum: for all the sample sizes used, there

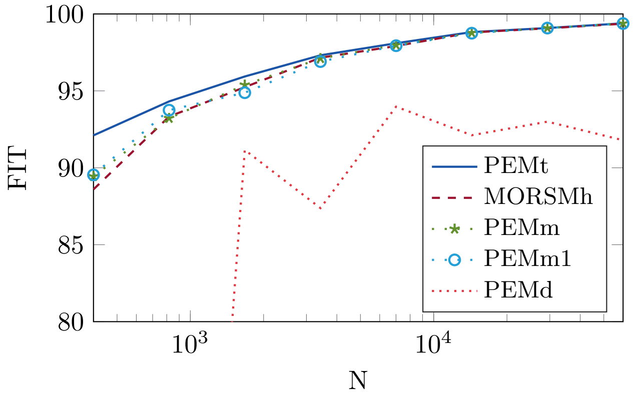

Fig. 3. Example 3: average FIT for several methods, obtained from 100 Monte Carlo runs with random systems.

Table 2
Examples 2 and 3: average number of iterations until convergence for several methods and different sample sizes. For PEM with different initializations, only the PEM iterations are counted.

Method\\N	400	818	3425	7007	29328	60000
BJSM	46	54	116	111	119	117
MORSM	23	13	8	6	4	3
PEM	18	18	15	16	12	18
PEMt	7	5	3	2	2	2
PEMm	7	4	2	1	1	1
PEMm1	10	6	4	3	2	1

are cases when PEM does not converge to the same parameters as PEMt, which decreases the average FIT. On the other hand, there is a considerable improvement for PEM if it is initialized with MORSM, performing close to PEM initialized at the true parameters for the smallest sample sizes and identically otherwise. However, initializing with one iteration of MORSM or with the iterative version gives identical performance. Therefore, if MORSM is used to initialize PEM, it may not be needed to wait for MORSM to converge. Alternatively, the iterative version of MORSM with low-order noise model estimate also performs similar to these: in this case, using it to initialize PEM showed no improvement. As initialization for PEM, a few iterations of MORSM might be a good compromise as the number of PEM iterations decreases with iterative MORSM compared to only one MORSM iteration (*cf.* Table 2).

8.3 Comparison with other methods

For parametric models, when PEM converges to the global optimum, it provides the best possible estimate in a maximum-likelihood sense. Hence, upon convergence to the global optimum, it is not expectable that other methods that achieve the same statistical properties using alternative numerical algorithms (e.g., MORSM or RIV) do better than PEM. The question is how robust a method is against failures—that is, cases where it converges to low-accuracy estimates that do not correspond to the global optimum.

In the following simulations, we will use two systems, each in two different settings, where PEM often converges to a non-global minimum. We will compare with different methods, with MORSM showing robustness against failures of the algorithm, while still having a median performance competitive to other methods with the same theoretical statistical properties.

The following methods will be compared:

- the prediction error method, with default MATLAB initialization (PEMd);
- subspace method with CVA weighting (SS);
- refined instrumental variable method (RIV);
- the iterative Model Order Reduction Steiglitz-McBride method (MORSM);
- the prediction error method, initialized at the true parameters (PEMt).

PEMd, SS and PEMt are according to the implementation in MATLAB2016b with default settings. RIV is according to the implementation in the CAPTAIN toolbox v7.5:11 with default settings. With PEM and RIV, the plant and noise models always have the correct order. With SS, the state-space model order is chosen as the maximum order of the plant and noise model. With MORSM, the plant is estimated with the correct order, and the noise model is non-parametric. Here, we do not initialize PEM with MORSM, as we want to make use of the feature that MORSM does not require estimating a low-order noise model. However, we include PEM initialized at the true parameters as benchmark.

8.3.1 System 1: widely separated eigenvalues

The first system we consider is given by

$$G_\circ(q) = \frac{0.016 + 0.026q^{-1} - 0.0375q^{-2}}{1 - 1.6252q^{-1} + 0.642q^{-2}}. \tag{40}$$

This system, with its widely separated eigenvalues, is problematic for PEM in some conditions if the initial conditions are not very close to the true parameter values [30]. Here, we begin by repeating the simulation in [30], for which RIV does not have the same convergence problems as PEM. In the considered scenario,

$$H_\circ(q) = \frac{1 + 0.5q^{-1}}{1 - 0.85q^{-1}}, \tag{41}$$

$\{u_t\}$ and $\{e_t\}$ are zero-mean Gaussian white-noise sequences with variances 8.8 and 0.0009, respectively, and the sample size is $N = 1700$. We perform 100 Monte Carlo simulations.

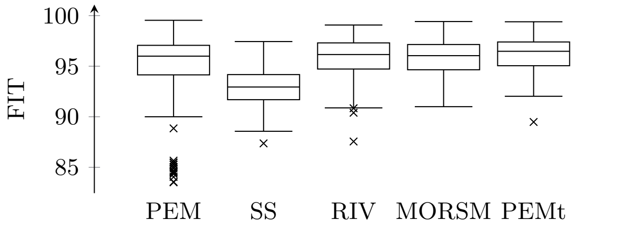

Fig. 4. System 1 with true noise model (41): boxplot of FIT for several methods, obtained from 100 Monte Carlo runs.

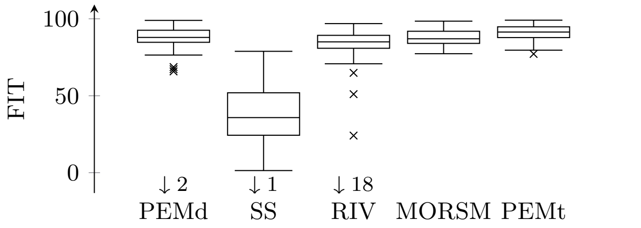

Fig. 5. System 1 with true noise model (42): boxplot of FIT for several methods, obtained from 100 Monte Carlo runs.

The obtained FITs are shown in Figure 4. Confirming the results in [30], there is probably a local minimum for the PEM cost function giving a FIT around 85, where the optimization procedure often converges to with the default initialization in MATLAB. In this simulation, the subspace method CVA is an appropriate approach to avoid the local-minimum issue with PEM; however, the median performance is inferior to PEM. Also RIV avoids the problematic local minimum of PEM, and has a median performance superior to subspace. Finally, MORSM performs similarly to RIV and to PEM initialized at the true parameters.

We now consider the same simulation settings except for the noise model, now given by

$$H_\circ(q) = \frac{1 + 0.23q^{-1} + 0.07q^{-2} + 0.05q^{-3} + 0.014q^{-4}}{1 - 3.04q^{-1} + 3.85q^{-2} - 2.36q^{-3} + 0.616q^{-4}}. \tag{42}$$

In this case, the results are given in Figure 5. PEM with the default MATLAB initialization has less poor-performance estimates than in the previous scenario, but two cases with negative FIT are still encountered. Subspace CVA is not competitive, having poor median performance. Although RIV has a median performance similar to PEM, it has a considerable amount of low-performance outliers. Finally, MORSM has a median performance similar to RIV and PEM, but with no outliers. Despite the robustness of MORSM against failures of the algorithm, the estimates obtained not always correspond to the global minimum of the PEM cost function, as PEM when initialized at the true parameters has slightly better performance.

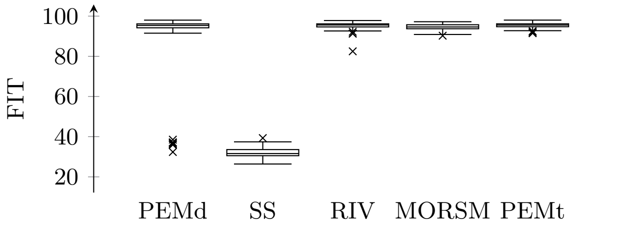

Fig. 6. System 2 with white input: boxplot of FIT for several methods, obtained from 100 Monte Carlo runs.

8.3.2 System 2: resonance peaks

The second system we consider is a 6th order system with three resonance peaks, given by

$$\begin{aligned} L_\circ(q) &= 0.08q^{-1} + 0.53q^{-2} - 0.29q^{-3} \\ &\quad - 0.51q^{-4} + 0.23q^{-5} + 0.04q^{-6} \\ F_\circ(q) &= 1 - 1.89q^{-1} + 2.26q^{-2} - 1.78q^{-3} \\ &\quad + 1.63q^{-4} - 1.09q^{-5} + 0.56q^{-6} \end{aligned} \tag{43}$$

The noise model is given by

$$H_\circ(q) = \frac{1 + 0.8q^{-1}}{1 - 0.9q^{-1}}, \tag{44}$$

and $\{e_t\}$ is a Gaussian white-noise sequence with unit variance, and the sample size is $N = 2600$. We consider two scenarios, with different inputs. In the first, $\{u_t\}$ is a Gaussian white-noise sequence with unit variance.

The FITs obtained from 100 Monte Carlo simulations are shown in Figure 6. Like the case in Figure 4, there is a local minimum in the PEM cost function corresponding to FIT between 30 and 40%, to where the optimization procedure often converges with the standard MATLAB initialization. The subspace method also provides a model that gives a FIT around 30–40%. Also like in Figure 4, RIV and MORSM always avoid the non-global minimum that PEM sometimes converges to, and both perform close to PEM initialized at the true parameters.

In the following, we show a scenario where MORSM is advantageous with respect to the other methods in this study: compared to subspace, it has better median performance; compared to PEM (default MATLAB initialization) and RIV, it has similar median performance but, unlike these, shows no algorithm failures. The setting is the same as before, except for the input. In this case, the input is given by

$$u_t = \frac{\sqrt{0.05}}{1 - 1.85q^{-1} + 0.87q^{-2}} u_t^w, \tag{45}$$

where $\{u_t^w\}$ is the input from the previous simulation. This is a low-pass filter with cut-off frequency 0.1rad/s.

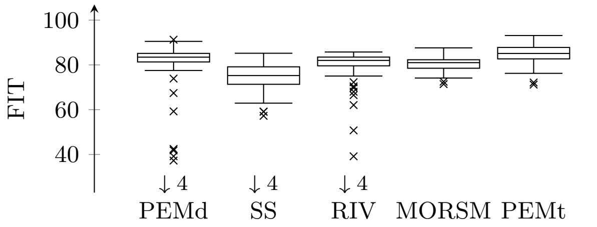

Fig. 7. System 2 with colored input (45): boxplot of FIT for several methods, obtained from 100 Monte Carlo runs.

The FITs obtained from 100 Monte Carlo simulations are shown in Figure 7. Here, PEM with the default MATLAB initialization has a considerable amount of low-accuracy outliers where the algorithm fails to find the global optimum. The subspace methods is favored with this setting compared to the previous one, but the median performance is still worse than PEM. Like PEM, the RIV algorithm also fails a considerable number of times, and the median performance is slightly inferior to PEM. On the other hand, MORSM has no algorithm failures, while the median performance is only slightly inferior to that of PEM and RIV. Despite the robust performance of MORSM, it does not attain the performance of PEM initialized with the true parameters, meaning that MORSM is not converging to the global minimum of PEM, but always finds a model with good performance.

9 Discussion and Conclusion

In this paper, we proposed a least-squares method for estimation of models with a plant parameterized by a rational transfer function and a non-parametric noise model. The method reduces a non-parametric model to a parametric one based on an asymptotic ML criterion using the Steiglitz-McBride method. We thus name it Model Order Reduction Steiglitz-McBride (MORSM). We show that the method provides consistent and asymptotically efficient estimates of the plant in one iteration if data are obtained in open loop.

We also perform simulations to study the performance of the method. The following results were 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 competitive with PEM and other methods; in particular, robustness against outliers in challenging scenarios was a main advantage compared with competitive methods.

Several extensions of MORSM to other system structures are possible. The multi-input multi-output (MIMO) case with a diagonal noise model is straightforward. In this case, each output can be considered individually, and a set of multi-input single-output (MISO) systems remain to be estimated. Then, each of these systems can also be considered individually, as all but the one of interest can be replaced by their corresponding high-order estimates. The MIMO case with fully-parametrized noise model is not as straightforward, and will be considered for future work.

Also the closed-loop case can be addressed with a direct estimation (i.e., using $\{y_t, u_t\}$ as data) by starting from the asymptotic maximum likelihood cost function for closed loop. In this case, the estimation of the plant model independently of the noise model makes it impossible to attain the CR bound. The asymptotic covariance attained will correspond to the best possible with a non-parametric noise model [4]. However, such a theoretical analysis does not follow straightforwardly from the open-loop case, and is also considered for future work.

Finally, MORSM has already been applied to estimate systems embedded in networks, showing promising performance [5]. A theoretical analysis and a more in-depth simulation study are already under development.

The theoretical analysis and simulation study of MORSM performed in this paper illustrate the potential of the method applied in open loop to single-input single-output (SISO) systems. However, the potential for extensions also makes the results in this paper essential for the development and analysis of future extensions.

A Proof of Lemma 3.1

The result follows from Theorem 3.1 in [14]. Next, we verify the conditions S1, S2, U1, D1 and D3 of that theorem. Assumption 2.1 and the finite dimensionality of G° and H° imply that

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

for some $C < \infty$ and $0 < \rho < 1$. In turn, this implies that Condition S1 holds. Furthermore, the bound (A.1) implies the inequality in (13) for some $\bar{C} < \infty$. Assumption 2.3 clearly implies Condition S2 (for any $p \leq 5$). Assumption 2.2 clearly implies Condition U1. Assumption 3.1 implies Conditions D1 and D3. Thus all conditions in Theorem 3.1 of [14] 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

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

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

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

We introduce the notation $f(N) = \mathcal{O}(g(N))$ to mean that $f(N)$ decays to zero with the rate $g(N)$: that is, that there are 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

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

By standard continuity arguments, with probability 1, $R^m(\hat{\eta}_N, \theta) > 0$ for large enough N. Hence, for large enough N, replacing (32) in (33), we have

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

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 $C_1 < \infty$. Furthermore, by Lemma 3.1, it follows that $F^\circ(q)\xi_t(\hat{\eta}_N, \hat{\theta}_N^k) = \mathcal{O}(m(N))$. Finally, we have that

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

and

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

This is valid for any previous estimate $\hat{\theta}_N^k$ within the above mentioned stability region. In particular, it also holds for the initialization estimate (i.e., the estimate obtained in Step 1 of the Steiglitz-McBride algorithm, when the previous estimate is set to $F(q,\theta) \equiv 1$).

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.

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

as v_t is quasi-stationarity.

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

$$\sqrt{N}\Upsilon^n(\hat{\eta}_N - \bar{\eta}^n) \sim As\mathcal{N}(0, P), \qquad \textit{(C.1)}$$

where

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

if the limit exists.

PROOF. See [14, Theorem 7.3].

Lemma C.3 *Let* $\{x_n\}$ *be a sequence of random variables that is asymptotically Gaussian distributed—* $\{x_n\} \sim As\mathcal{N}(0,P)$. *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*

$$y_n = M_n x_n + b_n. \qquad \text{(C.3)}$$

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

PROOF. See [22, Lemma B.4].

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

$$\begin{bmatrix} -F_1F_2\Gamma_n & F_3\Gamma_n \\ F_2\Gamma_n & 0 \end{bmatrix}, \qquad \textit{(C.4)}$$

where

$$\varGamma_n(q) = \begin{bmatrix} q^{-1} \; \dots \; q^{-n} \end{bmatrix}^\top, \qquad (C.5)$$

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

Suppose that F_1, F_2 *and* F_3 *are exponentially stable: for an exponentially stable* F_i*,*

$$|f_k^i| \leq C\lambda^k, \quad \text{for some } C < \infty, \quad \lambda < 1, \qquad (C.7)$$

and that there is a causal exponentially stable inverse

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

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

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

PROOF. We will construct an explicit approximation of γ that belongs to $\mathcal{S}_n$. Let $\tilde{F}_2\gamma := \left[\sum_{l=1}^{\infty} \beta_l z^{-l} \;\; 0\right]$, which is exponentially stable since both γ and $\tilde{F}_2$ are exponentially stable. Take as approximation for γ, $\hat{\gamma}_n := \left[\sum_{l=1}^{n} \beta_l F_2(z) z^{-l} \;\; 0\right]$, which by construction belongs to $\mathcal{S}_\Psi$. Introduce the notation $\gamma = [\,\gamma_1 \; \gamma_2\,]$. Then,

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

for some $C < \infty$ and $\lambda < 1$, as 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 $M_N = R^m(\hat{\eta}_N, \theta_N^k)$ and

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

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

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

Assume for now (we will prove it later) that

$$x_N \sim \text{As}\mathcal{N}(0, P).$$

Then, using Lemma C.3, we have that

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

D.1 x_N

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

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

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

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

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

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

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 tends to zero by Assumption 3.1.

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

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

tends 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 the difference between (D.7) and

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

tends 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

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

tends 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 $T_N \sim \text{As}\mathcal{N}(0,Q)$, 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

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

where

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

and with Ψ given by

$$\Psi = \begin{bmatrix}-G^\circ\Gamma_n & H^\circ\Gamma_n \\ \Gamma_n & 0_{n\times1}\end{bmatrix}U_\circ, \qquad U_\circ = \begin{bmatrix}F_u & 0 \\ 0 & \sigma_\circ\end{bmatrix}.$$

For (D.12), we have that

$$\begin{aligned}Z^n &= \text{E}\left[\varphi_t^m(\eta_\circ,\theta_\circ)\frac{F^\circ(q)}{A^\circ(q)}\varphi_t^n(\eta_\circ,\theta_\circ)^\top\right] \\ &= \text{E}\left[\begin{bmatrix}-\frac{B^\circ}{F^\circ}\Gamma_m u_t \\ \frac{A^\circ}{F^\circ}\Gamma_m u_t\end{bmatrix}\begin{bmatrix}-G^\circ\Gamma_n u_t \\ \Gamma_n u_t\end{bmatrix}^\top\right] \\ &= \left\langle\begin{bmatrix}-\frac{G^\circ}{F^\circ H^\circ}\Gamma_m & 0_{n\times1} \\ \frac{1}{F^\circ H^\circ}\Gamma_m & 0_{n\times1}\end{bmatrix}F_u, \begin{bmatrix}-G^\circ\Gamma_n & 0_{n\times1} \\ \Gamma_n & 0_{n\times1}\end{bmatrix}F_u\right\rangle \\ &= \langle\gamma,\Psi\rangle, \end{aligned} \tag{D.15}$$

with

$$\gamma = \begin{bmatrix}-\frac{G^\circ}{F^\circ H^\circ}\Gamma_m & 0_{m\times1} \\ \frac{1}{F^\circ H^\circ}\Gamma_m & 0_{m\times1}\end{bmatrix}F_u,$$

Hence, we can write the asymptotic covariance matrix of T_N as

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

where $\mathcal{S}_\Psi$ is the subspace in $\mathcal{L}_2^{1\times2}$ spanned by the rows of Ψ. Lemma C.4 gives that, as $n\to\infty$, $S_\gamma\subseteq\mathcal{S}_\Psi$ and

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

D.3 Summing up

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

$$T_N \sim \text{As}\mathcal{N}(0,\sigma_\circ^2 M_{CR}). \tag{D.17}$$

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

$$\sqrt{N}(\hat{\theta}_N^k - \theta_\circ) \sim \text{As}\mathcal{N}(0, \sigma_\circ^2 M_{CR}^{-1}). \qquad \text{(D.18)}$$

References

[1] G.E.P. Box and G. M. Jenkins. *Time series analysis: forecasting and control.* Holden-Day series in time series analysis and digital processing. Holden-Day.

[2] J. Durbin. Efficient estimation of parameters in moving-average models. *Biometrika*, 46(3/4):306, dec 1959.

[3] A. 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.

[4] U. Forssell and L. Ljung. Closed-loop identification revisited. *Automatica*, 35:1215–1241, 1999.

[5] M. Galrinho, N. Everitt, and H. Hjalmarsson. In *20th World Congress of the International Federation of Automatic Control*, Toulouse, France, 2017.

[6] 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.

[7] 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.

[8] E. J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems*, volume 70. SIAM, 1988.

[9] E. J. Hannan and L. Kavalieris. Linear estimation of ARMA processes. *Automatica*, 19(4):447–448, jul 1983.

[10] E. J. Hannan and L. Kavalieris. Multivariate linear time series models. *Advances in Applied Probability*, 16(03):492–561, 1984.

[11] E. J. Hannan and J. Rissanen. Recursive estimation of mixed autoregressive-moving average order. *Biometrika*, pages 81–94, 1982.

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

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

[14] 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.

[15] D. Q. Mayne and F. Firoozan. Linear identification of ARMA processes. *Automatica*, 18(4):461–466, jul 1982.

[16] 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.

[17] D. A. Pierce. Least squares estimation in dynamic disturbance time-series models. *Biometrika*, 5:73–78, 1972.

[18] 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.

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

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

[21] T. Söderström and P. Stoica. Optimal instrumental variable estimation and approximate implementation. *IEEE Transactions on Automatic Control*, 28:757–772, 1983.

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

[23] T. Söderström, P. Stoica, and B. Friedlander. An indirect prediction error method for system identification. *Automatica*, 27(1):183–188, 1991.

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

[25] 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.

[26] 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.

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

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

[29] P. Whittle. *Hypothesis testing in time series analysis*, volume 4. Almqvist & Wiksells, 1951.

[30] P. Young. The refined instrumental variable method: Unified estimation of discrete and continuous-time transfer function models. *Journal Europeen des Systemes Automatises*, 42:149–179, 2008.

[31] P. Young. Refined instrumental variable estimation. *Automatica*, 52(C):35–46, 2015.

[32] P. Young and A. Jakeman. Refined instrumental variable methods of recursive time-series analysis part I. single input, single output systems. *International Journal of Control*, 29(1):1–30, 1979.

[33] P. Young and A. Jakeman. Refined instrumental variable methods of recursive time-series analysis part II. multivariable systems. *International Journal of Control*, 29(4):621–644, 1979.

[34] P. Young and A. Jakeman. Refined instrumental variable methods of recursive time-series analysis part III. extensions. *International Journal of Control*, 31(4):741–764, 1980.

[35] P. Young, S. Parkinson, and M. Lees. Simplicity out of complexity in environmental modelling: Occam's razor revisited. *Journal of applied statistics*, 23(2-3):165–210, 1996.

[36] P. Young and M. Ratto. Statistical emulation of large linear dynamic models. *Technometrics*, 53(1):29–43, 2011.

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

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