**Abstract**—In identification of dynamical systems, the prediction error method (PEM) with a quadratic cost function provides asymptotically efficient estimates under Gaussian noise, but in general it requires solving a non-convex optimization problem, which may imply convergence to non-global minima. An alternative class of methods uses a non-parametric model as intermediate step to obtain the model of interest. Weighted null-space fitting (WNSF) belongs to this class, starting with the estimate of a non-parametric ARX model with least squares. Then, the reduction to a parametric model is a multi-step procedure where each step consists of the solution of a quadratic optimization problem, which can be obtained with weighted least squares. The method is suitable for both open- and closed-loop data, and can be applied to many common parametric model structures, including output-error, ARMAX, and Box-Jenkins. The price to pay is the increase of dimensionality in the non-parametric model, which needs to tend to infinity as function of the sample size for certain asymptotic statistical properties to hold. In this paper, we conduct a rigorous analysis of these properties: namely, consistency and asymptotic efficiency. Also, we perform a simulation study illustrating the performance of WNSF and identify scenarios where it can be particularly advantageous compared with state-of-the-art methods.

**Index Terms**—System identification, least squares.

**I. INTRODUCTION**

In model estimation from observed data, maximum likelihood (ML) is a classic framework to obtain parameter estimates. The basic idea is to maximize the likelihood function, using the underlying noise distribution to describe the likelihood that the process represented by the model produced the observed data. In the field of time-series inference and analysis, significant contributions were made to establish the appropriate assumptions and further derive consistency and asymptotic distribution results for the ML estimator. In particular, its asymptotic covariance corresponds to the inverse of the Fisher information matrix, which makes ML an asymptotically efficient estimator, having the best possible accuracy for consistent estimators: the Cramér-Rao bound. Several contributions can be referred to regarding these results: [1] established a quite complete theory for univariate linear stationary processes; multivariate models were addressed in [2]; and exogenous inputs included in [3]. For a general overview of the field containing many of the aforementioned results, we refer to [4].

For identification of dynamical systems for control design, the numerical application of maximum likelihood dates back to [5]. This led to the development of the prediction error method (PEM), consisting of minimizing a cost function of prediction errors—the difference between the observed output and its prediction based on the model and past data. When the noise is Gaussian, PEM with a quadratic cost function and ML are equivalent. These developments are treated in [6], where more practical conditions on the input signal from a control and identification perspective are assumed for proving consistency; in [7], asymptotic normality and accuracy are considered. A general treatment of PEM from numerical and theoretical perspectives is provided in [8]. Being widely available in software packages such as MATLAB, the method has become a benchmark in system identification.

There are two issues that may hinder successful application of PEM. The first—and most critical—is the risk of converging to a non-global minimum of the cost function, which is in general not convex. Thus, PEM requires local non-linear optimization algorithms and good initialization points. The second issue concerns closed-loop data. In this case, PEM is biased unless the noise model is flexible enough. For asymptotic efficiency, the noise model must be of correct order and estimated simultaneously with the dynamic model.

During the half century since the publication of [5], alternatives to PEM/ML have appeared, addressing one or both of the aforementioned issues. We will not attempt to fully review this vast field, but below we highlight some of the milestones.

Instrumental variable (IV) methods [9] allow consistency to be obtained in a large variety of settings without the issue of non-convexity. Asymptotic efficiency can be obtained for some problems using iterative algorithms [10,11]. However, IV methods cannot achieve the CR bound in closed loop [12].

Realization-based methods [13], which later evolved into subspace methods [14], are based on numerically stable procedures, having attractive computational properties. The bias issue for closed-loop data has been overcome by more recent algorithms [15]–[18]. However, structural information is difficult to incorporate, and—even if a complete analysis is still unavailable (significant contributions have been provided [18]–[20])—subspace methods are in general not believed to be as accurate as PEM.

Some methods are based on fixing some parameters in some
places of the cost function to obtain a quadratic optimization problem, so that the estimate can be obtained with (weighted) least squares. Then, the fixed coefficients are replaced by an estimate from the previous iteration in the weighting or in a filtering step. This leads to iterative methods, which date back to [21]. Some of these methods have been denoted iterative quadratic maximum likelihood (IQML), originally developed for filter design [22,23] and later applied to dynamical systems [24]–[26]. Another classical example is the Steiglitz-McBride method [27] for estimating output-error models, which is equivalent to IQML for an impulse-input case [28]. In the identification field, weightings or filterings have not been determined by statistical considerations. In this perspective, the result in [29], showing that the Steiglitz-McBride method is not asymptotically efficient, is not surprising.

Another approach is to estimate, in an intermediate step, a more flexible model, followed by a model reduction step to recover a model with the desired structure. The motivation for this procedure is that, in some cases, each step corresponds to a convex optimization problem or a numerically reliable procedure. To guarantee asymptotic efficiency, it is important that the intermediate model is a sufficient statistic and the model reduction step is performed in a statistically sound way. Indirect PEM [30] formalizes the requirements starting with an over-parametrized model of fixed order and uses ML in the model reduction step. The latter step corresponds in general to a weighted non-linear least-squares problem.

It has also been recognized that the intermediate model does not need to capture the true system perfectly, but only with sufficient accuracy. Subspace algorithms can be interpreted in this way: for example, SSARX [17] estimates an ARX model followed by a singular-value-decomposition (SVD) step and least-squares estimation. For spectral estimation, the THREE-like approach is also a two-step procedure that first obtains a non-parametric spectral estimate and then reduces it to a parametric estimate that in some sense is closest to the non-parametric one, and whose optimization function is convex [31].

Parameter estimation based on an intermediate high-order model has a long history in the field of time series, where such methods have arisen because of the difficulties associated with finding the global optimum of the likelihood function. Durbin’s first method [32] for auto-regressive moving-average (ARMA) time series uses an intermediate high-order AR model to simulate the innovations, which allows to obtain the ARMA parameters with least squares. This method does not have the same asymptotic properties as ML, unlike Durbin’s second method [32]: an extension of [33], where the parameters of an MA time series are estimated from the high-order AR estimates using least squares, with an accuracy that can be made arbitrarily close to the Cramér-Rao bound by increasing the AR-model order. When applied to ARMA time series in [32], the idea to achieve efficiency is to iterate between estimating the AR and MA polynomials using this procedure, initialized with Durbin’s first method. Another way to achieve efficiency from Durbin’s first method as starting point was proposed in [34] by using an additional filtering step with the MA estimates from Durbin’s first method, and then re-estimating the ARMA parameters.

The asymptotic properties of these methods have been analyzed by considering the high order tending to infinity, but “small” compared to the sample size. A preferable analysis should handle the relation between the high order and the sample size formally, as done in [35] to prove consistency of the method in [34], where the high order is assumed to tend to infinity as function of the sample size at a particular rate. This class of methods has become popular for vector ARMA time series, with several available algorithms using different procedures for obtaining the asymptotic efficient model parameter estimates from the estimated innovations (e.g., [36]–[39]). Despite sharing the same asymptotic properties, which have been analyzed with the high order as a function of the sample size, these algorithms may have different computational requirements and finite sample properties.

For identification of dynamical systems, instead of using the high-order model to estimate the innovations, it has been suggested that identification of the model of interest can be done by applying asymptotic ML directly to the high-order model [40]. The ASYM method [41] is an instantiation of this approach. Because an ARX-model estimate and its covariance constitute a sufficient statistic as the model order grows, this approach can produce asymptotically efficient estimates. However, the plant and noise models are estimated separately, preventing asymptotic efficiency for closed-loop data. Also, although such model reduction procedures may have numerical advantages over direct application of PEM [41], this approach still requires local non-linear optimization techniques. The Box-Jenkins Steiglitz-McBride (BJSM) method [42] instead uses the Steiglitz-McBride method in the model reduction step, resulting in asymptotically efficient estimates of the plant in open loop. Two drawbacks of BJSM are that the number of iterations is required to tend to infinity (as for the Steiglitz-McBride method) and that, similarly to [40] and [41], the CR bound cannot be attained in closed loop. The Model Order Reduction Steiglitz-McBride (MORSM) method solves the first drawback of BJSM, but not the second [43].

In this contribution, we focus on weighted null-space fitting (WNSF), introduced in [44]. This method uses two of the features of the methods above: i) an intermediate high-order ARX model; ii) the high-order model is directly used for estimating the low-order model using ML-based model reduction. However, instead of an explicit minimization of the model-reduction cost function—as in indirect PEM (directly via the model parameters), ASYM (in the time domain), and [40] (in the frequency domain)—the model reduction step consists of a weighted least-squares problem. Asymptotic efficiency requires that the weighting depend on the (to be estimated) model parameters. To handle this, an additional least-squares step is introduced. Consisting of three (weighted) least-squares steps, WNSF has attractive computational properties in comparison with, for example, PEM, ASYM, and BJSM.

More steps may be added to this standard procedure, using an iterative weighted least-squares algorithm, which may improve the estimate for finite sample size. Although also one Gauss-Newton iteration based of a PEM/ML cost function initialized at consistent estimate is enough to obtain an asymptotically efficient estimate, convergence for finite sample size may take
considerably longer, and simulations suggest that the multi-step WNSF approach may be beneficial for this purpose.

Another interesting feature of WNSF is that, unlike many of the methods above (including MORSM), the dynamic model and the noise model are estimated jointly. If this is not done, an algorithm cannot be asymptotically efficient for closed-loop data [45]. Nevertheless, in some applications, the noise model may be of no concern. WNSF can then be simplified and a noise model not estimated, still maintaining asymptotic efficiency for open-loop data. In closed loop, consistency is still maintained because the high-order model captures the noise spectrum consistently, while the resulting accuracy corresponds to the covariance of PEM with an infinite-order noise model [45]. Thus, besides the attractive numerical properties, WNSF has theoretical properties matched only by PEM. However, WNSF has the additional benefit that an explicit noise model is not required to obtain consistency with closed-loop data.

In [44], some theoretical properties of WNSF are claimed and supported by simulations, but with no formal proof. The robust performance that the method has shown has provided the motivation to extend the simulation study and deepen the theoretical analysis. Take Fig. 1 as an example, showing the FITs from 100 Monte Carlo runs with a highly resonant system.

![Fig. 1. FITs from 100 Monte Carlo runs with a highly resonant system.](image)

### II. PRELIMINARIES

#### A. Notation

- $||x||_p = (\sum_{k=1}^n |x_k|^p)^{1/p}$, with $x_k$ the $k^{th}$ entry of the $n \times 1$ vector $x$, and $p \in \mathbb{N}$ (for simplicity $||x|| := ||x||_2$).
- $||A||_p = \sup_{x \neq 0} ||Ax||_p/||x||_p$, with $A$ a matrix, $x$ a vector of appropriate dimensions, and $p \in \mathbb{N}$ (for simplicity $||A|| := ||A||_2$; also, $||A||_\infty = ||A^T||_1$).
- $\Gamma_n(q) = [q^{-1} \cdots q^{-n}]^T$, where $q^{-1}$ is the backward time-shift operator.
- $||G(q)||_{\mathcal{H}_\infty} := \sup_{\omega \in \mathbb{R}^+} ||G(e^{j\omega})||$, with $G(q)$ a transfer function.
- $C$ and $N$ denote constants, which need not be the same in different expressions, and may be random variables.
- $A^*$ is the complex conjugate transpose of the matrix $A$.
- $T_{n,m}(X(q))$ is the Toeplitz matrix of size $n \times m$ ($m \leq n$) with first column $[x_0 \cdots x_{n-1}]^T$ and first row $[x_0 \cdots 0_{1\times m-1}]$, where $X(q) = \sum_{k=0}^{\infty} x_k q^{-k}$. The dimension $n$ may be infinity, denoted $T_{\infty,m}(X(q))$.
- $\mathbb{E}$ denotes expectation of the random vector $x$.
- $\mathbb{E}_x := \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_x$.
- $x_N = \mathcal{O}(f_N)$: the function $x_N$ tends to zero at a rate not slower than $f_N$ as $N \to \infty$ w.p.1.
- $x_N \sim \mathcal{N}(\alpha, P)$: the random variable $x_N$ is normally distributed with mean $\alpha$ and covariance $P$ as $N \to \infty$.

#### B. Definitions and Assumptions

**Assumption 1** (Model and true system). The model has input $\{u_t\}$, output $\{y_t\}$ and is subject to the noise $\{e_t\}$, all real-valued, related by

$$y_t = G(q, \theta)u_t + H(q, \theta)e_t.$$  \hspace{1cm} (1)

The transfer functions $G(q, \theta)$ and $H(q, \theta)$ are rational functions in $q^{-1}$, according to

$$G(q, \theta) := \frac{L(q, \theta)}{P(q, \theta)} := \frac{l_1 q^{-1} + \cdots + l_m q^{-m_l}}{1 + f_1 q^{-1} + \cdots + f_m q^{-m_f}},$$

$$H(q, \theta) := \frac{C(q, \theta)}{D(q, \theta)} := \frac{1 + c_1 q^{-1} + \cdots + c_m q^{-m_c}}{1 + d_1 q^{-1} + \cdots + d_m q^{-m_d}},$$

where $\theta$ is the parameter vector to be estimated, given by

$$\theta = [f^T \ l^T \ c^T \ d^T]^T \in \mathbb{R}^{m_f + m_l + m_c + m_d},$$ \hspace{1cm} (2)

with

$$f = \begin{bmatrix} f_1 \\ \vdots \\ f_{m_f} \end{bmatrix}, \quad l = \begin{bmatrix} l_1 \\ \vdots \\ l_m \end{bmatrix}, \quad c = \begin{bmatrix} c_1 \\ \vdots \\ c_m \end{bmatrix}, \quad d = \begin{bmatrix} d_1 \\ \vdots \\ d_m \end{bmatrix}.$$
If the noise model is not of interest, we consider that we want to obtain an estimate \( G(q, \theta) \), where \( \theta = [f^\top \ I] ^\top \).

The true system is described by (1) when \( \theta = \theta_0 \). The transfer functions \( G_0 := G(q, \theta_0) \) and \( H_0 := H(q, \theta_0) \) are assumed to be stable, and \( H_0 \) inversely stable. The polynomials \( L_0 := L(q, \theta_0) \) and \( F_0 := F(q, \theta_0) \), as well as \( C_0 := C(q, \theta_0) \) and \( D_0 := D(q, \theta_0) \), do not share common factors.

Because we allow for data to be collected in closed loop, the input \( \{u_t\} \) is a stochastic process. Then, let \( \mathcal{F}_{t-1} \) be the \( \sigma \)-algebra generated by \( \{e_s, u_s, s \leq t - 1\} \). For the noise, the following assumption applies.

**Assumption 2 (Noise).** The noise sequence \( \{e_t\} \) is a stochastic process that satisfies
\[
E[e_t | \mathcal{F}_{t-1}] = 0, \quad E[e_t^2 | \mathcal{F}_{t-1}] = \sigma_a^2, \quad E[|e_t|^{10}] \leq C, \forall t.
\]

Before stating the assumption on the input sequence, we introduce the following definitions, used in [46].

**Definition 1 \((f_N\text{-}quasi-stationarity).** Let \( f_N \) be a decreasing sequence of positive scalars, with \( f_N \to 0 \) as \( N \to \infty \), and
\[
R_{\text{uv}}(\tau) = \left\{ \begin{array}{ll} \frac{1}{N} \sum_{n=-N}^{N} v_n u_{n+\tau}^\top, & 0 \leq \tau < N, \\ \frac{1}{N} \sum_{n=1}^{N} v_n u_{n-\tau}^\top, & -N < \tau \leq 0, \\ 0, & \text{otherwise.} \end{array} \right.
\]
The vector sequence \( \{v_t\} \) is \( f_N\text{-}quasi-stationary if

i) There exists \( R_{\text{uv}}(\tau) \) such that
\[
\sup_{|\tau| \leq N} \| R_{\text{uv}}(\tau) - R_{\text{uv}}(0) \| \leq C_1 f_N,
\]

ii) \( \frac{1}{N} \sum_{n=-N}^{N} v_n^2 \leq C_2 \)

for all \( N \) large enough, where \( C_1 \) and \( C_2 \) are finite constants.

The sample covariances in Definition 1 are according to [46] (although the normalization with \( N \) gives a biased estimate, this is irrelevant asymptotically). Also, condition ii) is used to guarantee that \( u_t \) does not explode in the infinite past. Definition 1 allows us to use quasi-stationary signals: signals containing stationary stochastic components and deterministic components, according to the definition in [8, Chapter 2]. Here, in addition to the standard definition of quasi-stationarity, a rate of convergence for the sample covariances is defined.

**Definition 2 \((f_N\text{-}stability).** A filter \( G(q) = \sum_{k=0}^{\infty} g_k q^{-k} \) is \( f_N\text{-}stable if \( \sum_{k=0}^{\infty} |g_k| f_k < \infty \).

**Definition 3** (Power spectral density). The power spectral density of an \( f_N\text{-}quasi-stationary sequence \( \{v_t\} \) is given by \( \Phi_{\text{uv}}(z) = \sum_{\tau=-\infty}^{\infty} R_{\text{uv}}(\tau) z^{-\tau} \), if the sum exists for \( |z| = 1 \).

For the input, the following assumption applies.

**Assumption 3 (Input).** The input sequence \( \{u_t\} \) is defined by
\[
u_t = -(K(q) y_t + r_t)
\]
under the following conditions.

i) The sequence \( \{r_t\} \) is independent of \( \{e_t\}, \quad \text{\( f_N\text{-}quasi-stationary with } f_N = \sqrt{\log N/N}, \text{ and uniformly bounded.}

ii) With \( \Phi_v(z) = F_v(z) \Phi_r(z^{-1}) \) the spectral factorization of \( \{r_t\} \) and \( F_v(q) \) causal, \( F_v(q) \) is BIBO stable.

iii) The closed loop system is \( f_N\text{-}stable with } f_N = 1/\sqrt{N}.

iv) The transfer function \( K(z) \) is bounded on the unit circle.

v) The spectral density of \( \{|r_t, e_t|^\top\} \) is coercive (i.e., bounded from below by the matrix \( \delta I \), for some \( \delta > 0 \).

Operation in open loop is obtained by taking \( K(q) = 0 \). The choice of \( f_N \) in iii) guarantees that the impulse responses of the closed-loop system have a minimum rate of decay, necessary to derive the results in [46]. This minimum decay rate is trivially satisfied here, as the system is stable and finite dimensional, and hence has exponentially decaying impulse responses.

### C. The Prediction Error Method

We now proceed with an overview of PEM for estimating \( \theta \). The idea is to minimize a cost function of prediction errors, which, for the model structure (1), are
\[
e_t(\theta) = \frac{D(q, \theta)}{C(q, \theta)} \left( y_t - \frac{L(q, \theta)}{F(q, \theta)} u_t \right).
\]

Using a quadratic cost function, the PEM estimate of \( \theta \) is obtained by minimizing
\[
J(\theta) = \frac{1}{N} \sum_{t=1}^{N} e_t^2(\theta), \tag{3}
\]
where \( N \) is the sample size. Assuming that \( \theta \) belongs to an appropriate domain [8, Def. 4.3], when the data set is informative [8, Def. 8.1] and under appropriate technical conditions [8, Chap. 8], the global minimizer \( \hat{\theta}_{\text{NPEM}} \) of (3) is asymptotically distributed as [8, Theorem 9.1]
\[
\sqrt{N}(\hat{\theta}_{\text{NPEM}} - \theta_0) \sim A_N(0, \sigma_{\text{NPEM}}^2 M_{\text{CL}}^{-1}), \tag{4}
\]
with (for notational simplicity, we omit the argument \( e^{i\omega} \))
\[
\Omega = \begin{bmatrix} \frac{C_2}{\pi M_{\text{CR}}} \Gamma_{m_f} & 0 \\
\frac{C_2}{\pi M_{\text{CR}}} \Gamma_{m_i} & \frac{C_2}{\pi M_{\text{CR}}} \Gamma_{m_i}
\end{bmatrix}, \tag{6}
\]
and \( \Phi_v \) the spectrum of \( \{u_t, e_t\}^\top \). When the error sequence is Gaussian, PEM with a quadratic cost function is asymptotically efficient, with (5) corresponding to the CR bound [8, Chap. 9].

In open loop, the asymptotic covariance of the dynamic-model parameters is the top-left square block (with dimension \( m_f + m_i \)) of the matrix \( M_{\text{CL}}^{-1} \) in (4) even if the noise-model orders \( m_e \) and \( m_d \) are larger than the true ones; if smaller, the dynamic-model estimates are consistent but do not have the same covariance. In closed loop, the covariance of the dynamic-model estimates only corresponds to the top-left block of (4) if the noise-model orders are the true ones; if smaller, the dynamic-model estimates are biased; if larger, they are consistent and the asymptotic covariance can be bounded by \( \sigma^2 \) \( M_{\text{CL}}^{-1} \), where [45]
\[
M_{\text{CL}} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \tilde{\Omega}(e^{i\omega}) \Phi_v^\top(e^{i\omega}) \tilde{\Omega}^\top(e^{i\omega}) d\omega \tag{7}
\]
with \( \Phi_v^\top \) the spectrum of the input due to the reference. This corresponds to the case with infinite noise-model order.

The main drawback with PEM is that minimizing (3) is in general a non-convex optimization problem. Therefore, the global minimizer \( \hat{\theta}_{\text{NPEM}} \) is not guaranteed to be found. An exception is the ARX model.
D. High-Order ARX Modeling

The true system can alternatively be written as

$$A_o(q)y_t = B_o(q)u_t + e_t, \quad (8)$$

where the transfer functions

$$A_o(q) := \frac{1}{H_o(q)} := 1 + \sum_{k=1}^{\infty} a_k q^{-k},$$

$$B_o(q) := \frac{G_o(q)}{H_o(q)} =: \sum_{k=1}^{\infty} b_k q^{-k} \quad (9)$$

are stable (Assumption 1). Therefore, the ARX model

$$A(q, \eta^n)y_t = B(q, \eta^n)u_t + e_t, \quad (10)$$

where

$$\eta^n = [a_1 \cdots a_n b_1 \cdots b_n]^\top,$$

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

can approximate (8) arbitrarily well if the model order n is chosen arbitrarily large.

Because the prediction errors for the ARX model (10),

$$\epsilon_t(\eta^n) = A(q, \eta^n)y_t - B(q, \eta^n)u_t,$$

are linear in the model parameters \(\eta^n\), the corresponding PEM cost function (3) can be minimized with least squares. This is done as follows. First, re-write (10) in regression form as

$$y_t = (\varphi_t^n)^\top \eta^n + e_t, \quad (11)$$

where

$$\varphi_t^n = \begin{bmatrix} y_{t-1} \cdots y_{t-n} u_{t-1} \cdots u_{t-n} \end{bmatrix}^\top. \quad (12)$$

Then, the least-squares estimate of \(\eta^n\) is obtained by

$$\hat{\eta}^{n,ls}_N = [R_N^n]^{-1}r_N^n, \quad (13)$$

where

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

$$r_N^n = \frac{1}{N} \sum_{t=n+1}^{n+N} \varphi_t^n y_t. \quad (14)$$

As the sample size increases, we have [46]

$$R_N^n \rightarrow R^n := \mathbb{E} \left[ \varphi_t^n (\varphi_t^n)^\top \right], \quad \text{as } N \rightarrow \infty \text{ w.p.1},$$

$$r_N^n \rightarrow r^n := \mathbb{E} \left[ \varphi_t^n y_t \right], \quad \text{as } N \rightarrow \infty \text{ w.p.1}. \quad (15)$$

Consequently,

$$\hat{\eta}^{n,ls}_N \rightarrow \hat{\eta}^n := [R^n]^{-1}r^n, \quad \text{as } N \rightarrow \infty \text{ w.p.1}. \quad (16)$$

For future reference, we define

$$\eta_0^n := \begin{bmatrix} a_1^n \cdots a_n^n b_1^n \cdots b_n^n \end{bmatrix}^\top,$$

$$\eta_0 := \begin{bmatrix} a_1^0 \cdots a_n^0 b_1^0 \cdots b_n^0 \end{bmatrix}^\top. \quad (17)$$

The attractiveness of ARX modeling is the simplicity of estimation while approximating more general classes of systems with arbitrary accuracy. As the order n typically has to be taken large, the estimated ARX model will have high variance. Nevertheless, this estimate can be used as a means to estimate a parametric model of interest, such as (1). If ML is used for the reduction from the non-parametric to a parametric model estimate, it is possible to asymptotically do so without loss of information and obtain an asymptotically efficient estimate [40].

III. WEIGHTED NULL-SPACE FITTING METHOD

The idea of weighted null-space fitting [44] is to use the non-parametric ARX model estimate and its covariance to estimate the parametric model of interest. Unlike with ML, we do this with a multi-step least squares procedure. The method then consists of three steps. In the first step, a high-order ARX model is estimated with least squares, whose order increases with the number of observations. In the second step, the parametric model is estimated from the high-order ARX model with least squares, providing a consistent estimate. In the third step, the parametric model is re-estimated with weighted least squares. Because the optimal weighting depends on the true parameters, we replace these by the consistent estimate obtained in the previous step, which is sufficient to obtain an asymptotically efficient estimate. This corresponds to a common procedure in signal processing (see, e.g. [47, Lemma 5]), having also conceptual similarities with initializing ML with a strongly consistent estimate, where one Gauss-Newton iteration provides an asymptotically efficient estimate (e.g., [48, Chap. 23]). We now proceed to detail each of these steps, and refer the reader to [44] for a more descriptive presentation.

The first step consists in estimating \(\hat{\eta}^{n,ls}_N\) from (13). As discussed before, \(\hat{\eta}^{n,ls}_N\) and \(R_N^n\) are almost a sufficient statistic for our problem, if the ARX-model truncation error is small enough (later, this will be treated formally). Then, we will use \(\hat{\eta}^{n,ls}_N\) and \(R_N^n\) instead of data to estimate the model of interest.

The second step implements this as follows. Re-write (9) as

$$C_o(q)A_o(q) - D_o(q) = 0,$$

$$F_o(q)B_o(q) - L_o(q)A_o(q) = 0. \quad (18)$$

Then, (18) can be expanded as

$$\left(1 + c_1^0 q^{-1} + \cdots + c_{m_1}^0 q^{-m_1}\right) \left(1 + \sum_{k=1}^{\infty} a_k q^{-k}\right) - \left(1 + d_1^0 q^{-1} + \cdots + d_{m_2}^0 q^{-m_2}\right) = 0,$$

$$\left(1 + f_1^0 q^{-1} + \cdots + f_{m_3}^0 q^{-m_3}\right) \sum_{k=1}^{\infty} b_k q^{-k} - \left(1 + r_1^0 q^{-1} + \cdots + r_{m_4}^0 q^{-m_4}\right) = 0. \quad (19a)$$

To express \(\theta_0\) in terms of \(\eta_0\), we can use a vector form. Take the example with \(m_f = 1 = m_l\); then, the first n coefficients of (19b) can be expressed as

$$[b_1^0 \ b_2^0 \ \ldots \ b_{n-1}^0] - [a_1^0 \ a_2^0 \ \ldots \ a_{n-1}^0] \left[ f_1^0 \ f_2^0 \ \ldots \ f_n^0 \right] = 0. \quad (19b)$$

In general, a power-series product \(\sum_{k=0}^{\infty} \alpha_k q^{-k} \sum_{k=0}^{\infty} \beta_k q^{-k}\) can be written as the (Toeplitz-)matrix-vector product

$$\begin{bmatrix} \alpha_0 & \alpha_0 & 0 & \beta_0 & \beta_0 & 0 & \cdots \\ \alpha_1 & \alpha_0 & 0 & \beta_1 & \beta_1 & \beta_0 & 0 & \cdots \\ \alpha_2 & \alpha_1 & \alpha_0 & \beta_2 & \beta_2 & \beta_1 & \beta_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \cdots \\ \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \end{bmatrix}. \quad (20)$$
So, we may write (19a) and (19b) as (keeping the first \( n \) equations of each)

\[ 0 = Q_n(\eta^n)\theta_0, \]

with \( \theta_0 \) defined by (2) evaluated at the true parameters and

\[ Q_n(\eta^n) = \begin{bmatrix} 0 & 0 & -Q_2'(\eta^n) & Q_2^T \end{bmatrix}, \]

where, when evaluated at the true parameters \( \eta_n^0 \),

\[ Q_n(\eta_n^0) = T_{n,m_0}(A(q, \eta_0)), \]

which are Toeplitz matrices. Motivated by (21), we replace \( \eta_n^0 \) by its estimate \( \hat{\eta}_n^{\text{LS}} \), obtaining an over-determined system of equations, which may be solved for \( \theta \) using least squares:

\[ \hat{\theta}_N^{\text{LS}} = \left( Q_n^T(\hat{\eta}_N^{\text{LS}})Q_n(\hat{\eta}_N^{\text{LS}}) \right)^{-1} Q_n^T(\hat{\eta}_N^{\text{LS}}) \hat{\eta}_N^{\text{LS}}. \]

In (23), invertibility follows from convergence of \( \eta_n^{\text{LS}} \) to a vector arbitrarily close to \( \eta_n^0 \) as \( n \) increases, which is of larger dimension than \( \theta \), and the block-Toeplitz structure of \( Q_n(\eta^n) \) (this is treated formally in Lemma 1).

With (23), we have not accounted for the residuals in (21) when \( \eta_n^{\text{LS}} \) replaces \( \eta_n^0 \). The third step remedies this by re-estimating \( \theta \) in a statistically sound way. Consider (18) with \( A_o(q) \) and \( B_o(q) \) replaced by \( A(q, \eta^n) \) and \( B(q, \eta^n) \), respectively, for any \( \eta^n \). Then, the right-hand side does not equal zero; however, it can be re-written as

\[ C_o(q)A(q, \eta^n) - D_o(q) = C_o(q)[A(q, \eta^n) - A_o(q)] \]

\[ = F_o(q)[B(q, \eta^n) - B_o(q)] - L_o(q)[A(q, \eta^n) - A_o(q)], \]

where the right side of the equalities were obtained by summing \( A_o(q) - A_o(q) \) to \( A(q, \eta^n) \) and using (18), and analogously for \( B(q, \eta^n) \). Then, as we used (20) to write (18) in the vector form (21), we may also use it to write (24) in the vector form

\[ \eta^n - Q_n(\eta^n)\theta_0 = T_n(\theta_0)(\eta^n - \eta_n^0) =: \delta_n(\eta^n, \theta_0), \]

where

\[ T_n(\theta) = \begin{bmatrix} T^n_\theta(\theta) & 0 \\ -T^n_\theta(\theta) & T^n_\theta(\theta) \end{bmatrix}, \]

with \( T^n_\theta(\theta_0) = T_{n,m_0}(C(q, \eta_0), T^n_\theta(\theta_0) = T_{n,m_0}(L(q, \theta_0)), \) and \( T^n_\theta(\theta_0) = T_{n,m_0}(F(q, \theta_0)). \) The objective is then to estimate \( \theta \) that minimizes the residuals \( \delta_n(\hat{\eta}_N^{\text{LS}}, \theta) \). If we neglect the bias error from truncation of the ARX model, which should be close to zero for sufficiently large \( n \), we have that, approximately,

\[ \sqrt{N} (\hat{\eta}_N^{\text{LS}} - \eta_n^0) \sim AsN(0, \sigma_\theta^2 [R_n]^{-1}). \]

Then, using (27) and (25), we may write that, approximately,

\[ \delta_n(\hat{\eta}_N^{\text{LS}}) \sim AsN\left(0, T_n(\theta_0)\sigma_\theta^2 [R_n]^{-1} T_n^T(\theta_0)\right). \]

Because the residuals we want to minimize, given by \( \delta_n(\hat{\eta}_N^{\text{LS}}, \theta) = \hat{\eta}_N^{\text{LS}} - Q_n(\hat{\eta}_N^{\text{LS}})\theta \), are asymptotically distributed by (28), the estimate of \( \theta \) with minimum variance is the weighted least-squares estimate

\[ \hat{\theta}_N^{\text{WLS}} = \left( Q_n^T(\hat{\eta}_N^{\text{LS}})W_n(\theta_0)Q_n(\hat{\eta}_N^{\text{LS}}) \right)^{-1} Q_n^T(\hat{\eta}_N^{\text{LS}})W_n(\theta_0)\hat{\eta}_N^{\text{LS}}, \]

where the weighting matrix

\[ W_n(\theta_0) = \left( T_n(\theta_0)\sigma_\theta^2 [R_n]^{-1} T_n^T(\theta_0) \right)^{-1} \]

is the inverse of the covariance of the residuals [49]. Because \( \theta_0 \) and \( R_n^o \) are not available, we replace them by \( \hat{\theta}_N^{\text{LS}} \) and \( R_n^{\text{LS}} \), respectively (\( \sigma_\theta^2 \) can be disregarded, because the weighting can be scaled arbitrarily without influencing the solution). Thus, the third step consists in re-estimating \( \theta \) according to

\[ \hat{\theta}_N^{\text{WLS}} = \left( Q_n^T(\hat{\eta}_N^{\text{LS}})W_n(\hat{\theta}_N^{\text{LS}})Q_n(\hat{\eta}_N^{\text{LS}}) \right)^{-1} Q_n^T(\hat{\eta}_N^{\text{LS}})W_n(\hat{\theta}_N^{\text{LS}}), \]

where (we take the inverses of the matrices individually)

\[ W_n(\hat{\theta}_N^{\text{LS}}) = T_n^\top(\hat{\theta}_N^{\text{LS}})R_n^{\text{LS}}T_n^{-\top}(\hat{\theta}_N^{\text{LS}}), \]

with \( T_n(\hat{\theta}_N^{\text{LS}}) \) obtained using (26). Invertibility in (30) follows from (besides what was mentioned for Step 2) invertibility of \( W_n(\hat{\theta}_N^{\text{LS}}) \), which in turn follows from the lower-Toeplitz structure of \( T_n(\theta) \), convergence of \( \hat{\theta}_N^{\text{LS}} \) to \( \theta_0 \) and \( R_n^{\text{LS}} \) to \( R_n^o \) (this is treated formally in Lemmas 2 and 3). Because \( \hat{\theta}_N^{\text{LS}} \) is a consistent estimate of \( \theta_0 \) with an error decaying sufficiently fast, using \( \hat{\theta}_N^{\text{LS}} \) in the weighting should not change the asymptotic properties of \( \hat{\theta}_N^{\text{WLS}} \), which is an asymptotically efficient estimate, as will be proven in the next section.

In summary, WNSF consists of the following three steps:

1) estimate a high-order ARX model with least squares (13);
2) reduce the high-order ARX model to the model of interest with least squares (23);
3) re-estimate the model of interest by weighted least squares (30) using the weighting (31).

Two notes can be made about this procedure. First, the objective of the second step is to obtain a consistent estimate to construct the weighting; hence, the choice of least squares is arbitrary, and weighted least squares with any invertible weighting (e.g., \( W_n = R_n^o \)) can be used. Second, although \( \hat{\theta}_N^{\text{WLS}} \) is asymptotically efficient, it is possible to continue iterating, which may improve the estimate for finite sample size.

Other Settings

Despite having been presented for a fully parametrized SISO BJ model, we point out that the method is flexible in parametrization. For example, it is possible to fix some parameters in \( \theta \) if they are known, or to impose linear relations between parameters. Hence, other common model structures (e.g., OE, ARMA, ARMAX) may also be used, as well as multi-input multi-output (MIMO) versions of such structures. The requirement is that a relation between the high- and low-order parameters can be written in the form (21).

Moreover, a parametric noise model does not need to be estimated. In this case, disregard (19a) and consider only (19b). The subsequent steps can then be derived similarly. This approach is presented in detail and analyzed in [50]. In open
loop, it provides asymptotically efficient estimates of the
dynamic model; in closed loop, the estimates are consistent
and with asymptotic covariance corresponding to (7).

IV. ASYMPTOTIC PROPERTIES

We now turn to the asymptotic analysis of WNSF. Here, we
make a distinction between the main algorithm presented here
and the case without a low-order noise model estimate. Al-
though apparently simpler because of the smaller dimension of
the problem, the case without a noise model requires additional
care in the analysis. The reason is that the corresponding \( T_n(\theta) \)
in that case will not be square. Then, inverting the weighting as
in (31) (a relation that will be used for the analysis in this paper)
will not be valid, requiring another approach. Including in this
paper under-parametrized noise models is then not possible
for space concerns. Thus, the asymptotic analysis in this paper
considers the dynamic and noise models correctly parametrized,
in which case the algorithm is consistent and asymptotically
efficient. The case with an under-parametrized noise model
(in particular, the limit case where (19a) is neglected and no
noise-model is estimated) is considered in [50].

Because the ARX model (11) is a truncation of the true
system (8), some information is lost when using this estimate
to obtain a parametric model. Then, we need to ensure that,
as \( N \) grows, the truncation error will be sufficiently small
so that, asymptotically, no information is lost. To keep track
of the truncation error in the analysis (see appendices), we let
the model order \( n \) depend on the sample size \( N \)—denoting
\( n = n(N) \)—according to the following assumption.

**Assumption 4** (ARX-model order). It holds that

1. \( n(N) \to \infty \), as \( N \to \infty \);
2. \( n^{1+\delta}(N)/N \to 0 \), for some \( \delta > 0 \), as \( N \to \infty \);
3. \( \sqrt{N}d(N) \to 0 \) as \( N \to \infty \), where \( d(N) := \sum_{k=n(N)+1}^{\infty} |a_k^2| + |b_k^2| \).

Condition D1 implies that, as the sample size \( N \) tends to
infinity, so does the model order \( n \). Condition D2 establishes a
maximum rate at which the model order \( n \) is allowed to grow,
as we cannot use too high order compared with the number of
observations. A consequence of Condition D2 is that [46]
\[
n^2(N) \log(N)/N \to 0, \quad N \to \infty,
\]
which appears often in the theoretical analysis. Condition D3
establishes a minimum rate at which the model order \( n \) is
allowed to grow, which depends on the true system. For stable
systems with rational description, D3 can be re-written as
\[
\sqrt{N}n(N) \to 0 \quad \text{as} \quad N \to \infty
\]
for some \( \lambda < 1 \). Taking the logarithm of the expression in (33)
\[
\frac{1}{2} \log N + n(N) \log \lambda \to -\infty \quad \text{as} \quad N \to \infty,
\]
which implies that \( n(N) \) must tend to infinity faster than \( \log N \).

To facilitate the statistical analysis, the results in this section
consider, instead of (13), a regularized estimate
\[
\hat{\eta}_N^{n,\text{reg}} := \left[ R^{n}_{\text{reg}}(N) \right]^{-1} r_N^n,
\]
where
\[
R^{n}_{\text{reg}}(N) = \begin{cases} R^N_N & \text{if } ||[R^N_N]^{-1}|| < 2/\delta, \\
R^N_N + \frac{1}{2} I_{2n} & \text{otherwise},
\end{cases}
\]
for some small \( \delta > 0 \). Asymptotically, the first and second
order properties of \( \hat{\eta}_N^{n,\text{LS}} \) and \( \hat{\eta}_N^{n,\text{WLS}} \) are identical [46].

When we let \( n = n(N) \) according to Assumption 4, we use
\( \hat{\eta}_N := \hat{\eta}_N^{n(N)} \). We will also denote \( \hat{\eta}_N^{n(N)} \) and \( \eta_0^{n(N)} \), defined
in (16) and (17), respectively. Concerning the matrices (15),
(22), (26), (29), and (31), for notational simplicity we maintain
the subscript \( n \) even if \( n = n(N) \).

Some of the technical assumptions used in this paper differ
from those used for the asymptotic analysis of PEM [8]. For
example, the bound in Assumption 2 is stronger than what is
required for PEM. On the other hand, for PEM the parameter
vector \( \theta \) is required to belong to a compact set, which is
not imposed here. However, such differences in technical
assumptions have little relevance in practice.

We have the following result for consistency \( \hat{\theta}_N^{\text{LS}} \).

**Theorem 1.** Let Assumptions 1, 2, 3, and 4 hold, and \( \hat{\theta}_N^{\text{LS}} \) be defined by (23). Then,
\[
\hat{\theta}_N^{\text{LS}} \to \theta_o, \quad \text{as} \quad N \to \infty \; \text{w.p.1}.
\]

Moreover, we have that
\[
||\hat{\theta}_N^{\text{LS}} - \theta_o|| = O\left( \sqrt{n(N) \log(N) / N (1 + d(N))} \right).
\]

**Proof.** See Appendix I. \( \square \)

We have the following result for consistency \( \hat{\theta}_N^{\text{WLS}} \).

**Theorem 2.** Let Assumptions 1, 2, 3, and 4 hold, and \( \hat{\theta}_N^{\text{WLS}} \) be defined by (30). Then,
\[
\hat{\theta}_N^{\text{WLS}} \to \theta_o, \quad \text{as} \quad N \to \infty \; \text{w.p.1}.
\]

**Proof.** See Appendix II. \( \square \)

We have the following result for asymptotic distribution and
covariance of \( \hat{\theta}_N^{\text{WLS}} \).

**Theorem 3.** Let Assumptions 1, 2, 3, and 4 hold, and \( \hat{\theta}_N^{\text{WLS}} \) be defined by (30). Then,
\[
\sqrt{N}(\hat{\theta}_N^{\text{WLS}} - \theta_o) \sim \text{AsN}(0, \sigma^2(\hat{\theta}_N^{\text{WLS}})) \text{Cov}(\theta),
\]
where \( \text{MC subt.} \) is given by (5).

**Proof.** See Appendix III. \( \square \)

Theorem 3 implies, comparing with (4), that WNSF has the
same asymptotic properties as PEM. For Gaussian noise, this
corresponds to an asymptotically efficient estimate.

V. SIMULATION STUDIES

In this section, we perform simulation studies and discuss
practical issues. First, we illustrate the asymptotic properties
of the method. Second, we consider how to choose the order
of the non-parametric model. Third, we exemplify with two
difficult scenarios for PEM how WNSF can be advantageous in
terms of robustness against convergence to non-global minima
and convergence speed. Fourth, we perform a simulation with
random systems to test the robustness of the method compared with other state-of-the-art methods.

Although WNSF and the approach in [43] are different algorithms, they share similarities in using high-order models and iterative least squares. However, [43] is only applicable in open loop. Here, to differentiate WNSF as a more general approach that is applicable in open or closed loop without changing the algorithm, we focus on the typically more challenging closed-loop setting, for which many standard methods are not consistent.

A. Illustration of Asymptotic Properties

The first simulation has the purpose of illustrating that the method is asymptotically efficient. Here, we consider only the case where we estimate a correct noise model (the case where a low-order noise model is not estimated is illustrated in [50]). We perform open- and closed-loop simulations, where the closed-loop data are generated by

\[
\begin{align*}
  u_t &= \frac{1}{1 + K(q)G_o(q)} r_t - \frac{K(q)H_o(q)}{1 + K(q)G_o(q)} e_t, \\
  y_t &= \frac{G_o(q)}{1 + K(q)G_o(q)} r_t + \frac{H_o(q)}{1 + K(q)G_o(q)} e_t,
\end{align*}
\]

and the open-loop data by

\[
\begin{align*}
  u_t &= \frac{1}{1 + K(q)G_o(q)} r_t, \\
  y_t &= G_o(q)u_t + H_o(q)e_t,
\end{align*}
\]

where \( \{r_t\} \) and \( \{e_t\} \) are independent Gaussian white sequences with unit variance, \( K(q) = 1 \), and

\[
G_o(q) = \frac{q^{-1} + 0.1q^{-2}}{1 - 0.5q^{-1} + 0.75q^{-2}}, \quad H_o(q) = \frac{1 + 0.7q^{-1}}{1 - 0.9q^{-1}}.
\]

We perform 1000 Monte Carlo runs, with sample sizes \( N \in \{300, 600, 1000, 3000, 6000, 10000\} \). We apply WNSF with an ARX model of order 50 with open- and closed-loop data. Performance is evaluated by the mean-squared error of the estimated parameter vector of the dynamic model, \( \text{MSE} = ||\hat{\theta}_{\text{WNSF}} - \theta||^2 \), where \( \theta \) contains only the elements of \( \theta \) contributing to \( G(q, \theta) \). As this simulation has the purpose of illustrating asymptotic properties, initial conditions are zero and assumed known—that is, the sums in (14) start at \( t = 1 \) instead of \( t = n + 1 \).

The results are presented in Fig. 2, with the average MSE over 1000 Monte Carlo runs plotted as function of the sample size (closed loop in solid line, open loop in dash-dotted line), where we also plot the corresponding CR bounds (closed loop in dashed line, open loop in dotted line). The respective CR bounds are attained as the sample size increases.

B. Practical Issues

In the previous simulation, an ARX model of order 50 was estimated in the first step. Although the order of this model should, in theory, tend to infinity at some maximum rate to attain efficiency (Assumption 4), a fixed order was sufficient to illustrate the asymptotic properties of WNSF in this particular scenario. This suggests that, in practice but for sufficiently large sample size, a non-parametric model of fixed order with low enough bias error may be sufficient. However, for fixed sample size, the question remains on how to choose the most appropriate non-parametric model order. In this case, if \( n \) is chosen too large, the “tail” of the non-parametric model polynomials estimate will be too noisy, although the true values are close to zero; then, it may be better to neglect these coefficients by reducing the estimated non-parametric model order. Some previous knowledge about the speed of the system may help in choosing this order, but the most appropriate value may also depend on sample size and signal-to-noise ratio. In this paper, we use the PEM cost function (3) as criterion to choose \( n \) we compute \( \theta_{\text{WNSF}} \) for several \( n \), and choose the estimate that minimizes (3).

Also, \( \theta_{\text{WNSF}} \) need not be used as final estimate, or for finite sample size, performance may improve by iterating. However, because WNSF does not minimize the cost function (3) explicitly, it is not guaranteed that subsequent iterations correspond to a lower cost-function value than previous ones. Here, we will also use the cost function (3) as criterion to choose the best model among the iterations performed.

C. Comparison with PEM

One of the main limitations of PEM is the non-convex cost function, which may make the method sensitive to the initialization point. Here, we provide examples illustrating how WNSF may be a more robust method than PEM regarding initialization: in cases where the PEM cost function is highly non-convex, WNSF may require less iterations and be more robust against convergence to non-global minima of PEM.

We consider a system where \( H_o(q) = 1 \), \( K(q) = 0.3 \), and

\[
G_o(q) = \frac{1.0q^{-1} - 1.2q^{-2}}{1 - 2.5q^{-1} + 2.4q^{-2} - 0.88q^{-3}},
\]

with data generated according to (37), where

\[
r_t = \frac{1 + 0.7q^{-1}}{1 - 0.9q^{-1}} r^w_t,
\]

with \( \{e_t\} \) and \( \{r^w_t\} \) Gaussian white noise sequences with variances 4 and 0.25, respectively. The sample size is \( N = 2000 \). We estimate an OE model with the following algorithms:

- WNSF with a non-parametric model of order \( n = 250 \);
- PEM with default MATLAB initialization and Gauss-Newton (GN) algorithm;
All methods use a maximum of 100 iterations, but stop early upon convergence (default settings for PEM, 10^{-4} as tolerance for the normalized relative change in the parameter estimates) and initial conditions are zero.

Performance is evaluated by the FIT of the impulse response of the estimated OE model \( G(q, \hat{\theta}_N) \), given in percent by

\[
\text{FIT} = 100 \left(1 - \frac{\|g_0 - \hat{g}\|}{\|g_0 - \text{mean}(g_0)\|}\right), \tag{38}
\]

where \( g_0 \) is a vector with the impulse response parameters of \( G_0(q) \), and similarly for \( \hat{g} \) but for the estimated model. In (38), sufficiently long impulse responses are taken to make sure that the truncation of their tails does not affect the FIT.

The average FITs for 100 Monte Carlo runs are shown in Table I. For PEM, the results depend on the optimization method and the initialization point: as consequence of the non-convexity of PEM, the algorithms do not always converge to the global optimum. For PEM implementations, the average FIT is the same as for PEM started at the true parameters only with default MATLAB initialization and LM algorithm. For WNSF, the average FIT is the same as for PEM started at the true parameters independently of the initialization point used in the weighting matrix, suggesting robustness to different initial weighting matrices.

In this simulation, PEM was most robust with the LM algorithm and the default MATLAB initialization, having on average the same accuracy as WNSF. Then, it is appropriate to compare the performance of these methods by iteration when WNSF also is initialized with the same parameter values. In Fig. 3, we plot the average FITs for these methods as function of the maximum number of iterations. Here, WNSF reaches an average FIT of 98 after two iterations, while PEM with LM takes 20 iterations to reach the same value.

The robustness of WNSF against convergence to non-global minima compared with different instances of PEM can be even more evident than in Table I, as WNSF seems to be appropriate for modeling systems with many resonance peaks, for which the PEM cost function can be highly non-linear. Take the example in Fig. 1, based on 100 Monte Carlo runs for a system with

\[
L_0(q) = q^{-1} - 3.4q^{-2} + 4.8q^{-3} - 3.3q^{-4} + 0.96q^{-5},
\]

\[
F_0(q) = 1 - 5.4q^{-1} + 13.5q^{-2} - 20.1q^{-3} + 19.5q^{-4} - 12.1q^{-5} + 4.5q^{-6},
\]

and data generated according to (37) with \( K(q) = -0.05 \) and \( r_t = 0.05/(1-0.99q^{-1}) r_t^n \), where \( \{r_t^n\} \) and \( \{\epsilon_t\} \) are Gaussian white sequences with unit variance. Here, initial conditions are not assumed zero: PEM estimates initial conditions by backcasting and WNSF uses the approach in [51]. In this scenario, PEM with the LM algorithm and default initialization fails in most runs to find the global optimum. Subspace methods, often used to avoid the non-convexity of PEM, may not help in this scenario: SSARX [17], a subspace method that is consistent in closed loop, provides an average FIT around 20% (default MATLAB implementation). Here, WNSF with \( n \) between 100 and 600 spaced with intervals of 50 performs similarly to PEM initialized at the true parameters, accurately capturing the resonance peaks of the system.

### Table I

**Comparison with PEM: average FITs with different methods (Meth) and initializations (Init).**

| Meth         | Init | MtL | LS | true |
|--------------|------|-----|----|------|
| WNSF         | 98   | 98  | -  |      |
| PEM GN       | 74   | 87  | 98 |      |
| PEM LM       | 98   | 85  | 98 |      |

D. Random Systems

In order to test the robustness of the method, we now perform a simulation with random systems, comparing the performance of WNSF with other methods available in the MATLAB System Identification Toolbox. For a fair comparison, we only use methods that are consistent in closed loop using input and output data. From the subspace class, we use SSARX, as this method is competitive with other subspace algorithms such as CVA [52,53] and N4SID [54], while it is consistent in closed loop [17]. IV methods are not considered, as the instruments need to contain the reference signal in closed loop.

For the simulation, we use 100 systems with structure

\[
G_0(q) = \frac{p_0 q^{-1} + \cdots + p_m q^{-4}}{1 + f_0 q^{-6} + \cdots + f_m q^{-5}}.
\]

As we observed, PEM may have difficulties with slow resonant systems: therefore, it is for this class of systems that WNSF may be most advantageous. With this purpose, we generate the polynomial coefficients in the following way. The poles are located in an annulus with the radius uniformly distributed between 0.88 and 0.98, and the phase uniformly distributed between 0 and 90° (and respective complex conjugates). One pair of zeros is generated in the same way, and a third real zero
is uniformly distributed between $-1.2$ and $1.2$ (this allows for non-minimum-phase systems). The noise models have structure

$$H_n(q) = \frac{1 + c_1 q^{-1} + c_2 q^{-2}}{1 + d_1 q^{-1} + d_2 q^{-2}},$$

with the poles and zeros having uniformly distributed radius between $0$ and $0.95$, and uniformly distributed phase between $0$ and $180^\circ$ (and respective complex conjugates).

The data are generated in closed loop according to

$$u_t = \frac{K(q)}{1 + K(q)G_o(q)} r_t - \frac{K(q)H_n(q)}{1 + K(q)G_o(q)} e_t,$$
$$y_t = \frac{K(q)G_o(q)}{1 + K(q)G_o(q)} r_t + \frac{H_n(q)}{1 + K(q)G_o(q)} e_t,$$

where

$$r_t = \frac{1 - 1.273q^{-1} + 0.81q^{-2}}{1 - 1.559q^{-1} + 0.81q^{-2}} w_t$$

with $\{w_t\}$ a Gaussian white-noise sequence with unit variance, $\{e_t\}$ a Gaussian white-noise sequence with the variance chosen such that the signal-to-noise ratio (SNR) is

$$\text{SNR} = \frac{\sum_{t=1}^{N} \left[ \frac{K(q)G_o(q)}{1 + K(q)G_o(q)} r_t \right]^2}{\sum_{t=1}^{N} \left[ H_n(q)e_t \right]^2} = 2,$$

and the controller $K(q)$ is obtained using a Youla-parametrization to have an integrator and a closed-loop transfer function that has the same poles as the open loop except that the radius of the slowest open-loop pole pair is reduced by $80\%$. The sample size is $N = 2000$ and we perform 100 Monte Carlo runs (one for each system; different noise realizations).

We compare the following methods:

- PEM initialized at the true parameters (PEMt);
- PEM with default MATLAB initialization (PEMd);
- SSARX with the default MATLAB options;
- WNSF using the approach in Section V-B to choose $n$ from the grid $\{50, 100, 150, 200, 250, 300\}$;
- PEM initialized with WNSF (PEMw).

All methods estimate a fully parametrized noise model. We use the MATLAB2016b implementation of SSARX and PEM. For PEM, the optimization algorithm is LM. For SSARX, the horizons are chosen automatically by MATLAB, based on the Akaike Information Criterion. WNSF and PEM use a maximum of 100 iterations, but stop earlier upon convergence (default settings for PEM, $10^{-4}$ as tolerance for the normalized relative change in the parameter estimates. PEM estimates initial conditions by backcasting and WNSF truncates them ([51] does not apply to BJ models).

The FITs obtained in this simulation are presented in Fig. 4. In this scenario, PEM with default MATLAB initialization (PEMd) often fails to find a point close to the global optimum, which can be concluded by comparison with PEM initialized at the true parameters (PEMt). Also, SSARX is not an alternative for achieving better performance. WNSF can be an appropriate alternative, failing only once to provide an acceptable estimate, and having otherwise a performance close to the practically infeasible PEMt. The estimate obtained with WNSF may be used to initialize PEM. This provides a small improvement only, suggesting that the estimates obtained with WNSF are already close to a minimum of the PEM cost function.

VI. CONCLUSION

Methods for parameter estimation based on an intermediate unstructured model have a long history in system identification (e.g., [17,40]–[42]). Here, we believe to have taken a significant step further in this class of methods, with a method that is flexible in parametrization and provides consistent and asymptotically efficient estimates in open and closed loop without using a non-convex optimization or iterations.

In this paper, we provided a theoretical and experimental analysis of this method, named weighted null-space fitting (WNSF). Theoretically, we showed that the method is consistent and asymptotically efficient for stable Box-Jenkins systems. Experimentally, we performed Monte Carlo simulations, comparing PEM, subspace, and WNSF under settings where PEM typically performs poorly. The simulations suggest that WNSF is competitive with these methods, being a viable alternative to PEM or to provide initialization points for PEM.

Although WNSF was here presented for SISO BJ models, it was also pointed out that the flexibility in parametrization allows for a wider range of structures to be used, as well as for incorporating structural information (e.g., fixing specified parameters). Moreover, based on the analysis in [50], WNSF does not require a parametric noise model to achieve asymptotic efficiency in open loop and consistency in closed loop.

An extension that was not covered in this paper is the MIMO case, where subspace or IV methods are typically used [55], as PEM often has difficulty with estimation of such systems. Based on the theoretical foundation provided in this contribution, this important extension is already in preparation. Future work includes also extensions to dynamic networks and non-linear model structures.

APPENDIX I

CONSISTENCY OF STEP 2

The main purpose of this appendix is to prove Theorem 1. Before we do so, we introduce some auxiliary results.

- $\| \hat{\eta}_N - \eta_0^{n(N)} \| \text{ tends to zero as } N \text{ tends to infinity } w.p.1$

Consider the estimated parameter vector $\hat{\eta}_N := \eta_0^{n(N)}$ (34), and the truncated true parameter vector $\eta_0^{n(N)}$ (17). Using the triangular inequality, we have

$$\| \hat{\eta}_N - \eta_0^{n(N)} \| \leq \| \hat{\eta}_N - \tilde{\eta}_0^{n(N)} \| + \| \tilde{\eta}_0^{n(N)} - \eta_0^{n(N)} \|,$$ (39)
where \( \hat{\eta}^n \) is defined by (16). Then, from [46, Lemma 5.1], the second term on the right side of (39) tends to zero as \( n(N) \to \infty \). From [46, Theorem 5.1], the first term on the right side of (39) tends to zero as \( N \to \infty \) w.p.1. Thus,

\[
\| \hat{\eta}_N - \eta_0^{(N)} \| \to 0 \text{ as } N \to \infty \text{ w.p.1.} \tag{40}
\]

- \( \| Q_n(\hat{\eta}_N) - Q_n(\eta_0^{(N)}) \| \) tends to zero, as \( N \) to infinity, w.p.1

Consider \( Q_n(\eta_0^{(N)}) \), given by (22) evaluated at the truncated true parameter vector \( \eta_0^{(N)} \), and the matrix \( Q_n(\hat{\eta}_N) \), given by (22) evaluated at the estimated parameters \( \hat{\eta}_N \). We have

\[
\| Q_n(\hat{\eta}_N) - Q_n(\eta_0^{(N)}) \| \leq \| Q_n(\hat{\eta}_N) - Q_n(\hat{\eta}_N^{(N)}) \| + \| Q_n(\hat{\eta}_N^{(N)}) - Q_n(\eta_0^{(N)}) \| \leq C \| \hat{\eta}_N - \eta_0^{(N)} \|.
\]

(41)

Then, using (40), we conclude that

\[
\| Q_n(\hat{\eta}_N) - Q_n(\eta_0^{(N)}) \| \to 0 \text{ as } N \to \infty \text{ w.p.1.} \tag{42}
\]

- \( \| Q_n(\eta_0^{n}) \| \) is bounded for all \( n \).

We have that

\[
\| Q_n(\eta_0^{n}) \| \leq \| Q_n(\eta_0^{n}) \| + \| Q_n(\eta_0^{n}) \| + \| Q_n(\eta_0^{n}) \| \leq C \| \eta_0^{n} \| + 1 \quad \forall n,
\]

which is bounded because the coefficients of \( A_0(q) \) and \( B_0(q) \) in (9) are exponentially decaying (the true system is asymptotically stable and has a rational description).

- \( \| Q_n(\hat{\eta}_N) \| \) is bounded for large \( N \) w.p.1

Using the triangular inequality, we have

\[
\| Q_n(\hat{\eta}_N) \| \leq \| Q_n(\eta_0^{n}) \| + \| Q_n(\eta_0^{n}) \| + \| Q_n(\eta_0^{n}) \| \leq C \| \hat{\eta}_N \| + 1 \quad \forall n.
\]

(43)

Using now (43) and (41), the first term on the right side of (44) can be made arbitrarily small as \( N \) increases, while the second term is bounded for all \( n \). Then, there exists \( N \) such that

\[
\| Q_n(\hat{\eta}_N) \| \leq C \quad \forall N > N.
\]

(45)

- \( \| T_n(\theta_0) \| \) is bounded for all \( n \)

Consider the matrix \( T_n(\theta_0) \), given by (26). Bounding the norm of \( T_n(\theta_0) \) by its constituting blocks, we have that

\[
\| T_n(\theta_0) \| \leq \| T_n(\theta_0) \| + \| T_n(\theta_0) \| + \| T_n(\theta_0) \| \leq \| F_0(q) \|_H + \| C_0(q) \|_H + \| L_0(q) \|_H \leq C \quad \forall n,
\]

(46)

where the last two inequalities follow from [56, Theorem 3] and from asymptotic stability of \( F_0(q) \), \( C_0(q) \), and \( L_0(q) \).

**Proposition 1.** Consider the product \( \prod_{i=1}^p \hat{X}^{(i)}_N \), where \( p \) is finite and \( \hat{X}^{(i)}_N \) are stochastic matrices of appropriate dimensions (possibly a function of \( N \)) such that

\[
\| \hat{X}^{(i)}_N - X^{(i)}_N \| \to 0 \text{ as } N \to \infty \text{ w.p.1; } \| \hat{X}^{(i)}_N \| \leq C_i,
\]

(47)

where \( \hat{X}^{(i)}_N \) is a deterministic matrix for each \( N \), which may influence its dimensions according to the dimensions of \( X^{(i)}_N \). Then, we have that

\[
\| \prod_{i=1}^p \hat{X}^{(i)}_N - \prod_{i=1}^p X^{(i)}_N \| \to 0 \text{ as } N \to \infty \text{ w.p.1.} \tag{48}
\]

**Proof.** We show this by induction. First, let \( p = 2 \) and define \( \Delta^{(i)}_N := X^{(i)}_N - \hat{X}^{(i)}_N \). Then, we write

\[
\Delta^{(i)}_N X^{(2)}_N - \hat{X}^{(i)}_N \hat{X}^{(i)}_N = \Delta^{(1)}_N \hat{X}^{(i)}_N + \hat{X}^{(1)}_N \Delta^{(2)}_N + \Delta^{(1)}_N \Delta^{(2)}_N,
\]

which yields, using (47),

\[
\| \Delta^{(1)}_N \| \| \Delta^{(2)}_N \| \leq \| \Delta^{(1)}_N \| \| \Delta^{(2)}_N \|. \tag{49}
\]

Second, we consider an arbitrary \( p \), and assume that

\[
\| \prod_{i=1}^{p-1} \hat{X}^{(i)}_N - \prod_{i=1}^{p-1} X^{(i)}_N \| \to 0 \text{ as } N \to \infty \text{ w.p.1.} \tag{50}
\]

Then, using a similar procedure as (49), we have

\[
\| \prod_{i=1}^{p-1} \hat{X}^{(i)}_N - \prod_{i=1}^{p-1} X^{(i)}_N \| \leq \| \Delta^{(p)}_N \| \prod_{i=1}^{p-1} \| \hat{X}^{(i)}_N \| \leq \| \Delta^{(p)}_N \| \prod_{i=1}^{p-1} \| \Delta^{(i)}_N \| \tag{51}
\]

which, in turn, is bounded by

\[
\| \Delta^{(p)}_N \| \prod_{i=1}^{p-1} \| \hat{X}^{(i)}_N \| \leq \| \Delta^{(p)}_N \| \prod_{i=1}^{p-1} \| \Delta^{(i)}_N \| \to 0 \text{ as } N \to \infty \text{ w.p.1.}
\]

Thus, we can then write \( M(\eta_0) := Q^{\top}(\eta_0)Q(\eta_0) \). From this factorization, we observe that \( M(\eta_0) \) is non-singular if and only if \( Q^{\top}(\eta_0)Q(\eta_0) \) has a non-trivial null-space. Moreover, the block anti-diagonal structure of \( Q^{\top}(\eta_0)Q(\eta_0) \) implies that \( Q(\eta_0) \) has full column rank if and only if both matrices \( [-Q^{\top}(\eta_0)Q(\eta_0)] \) and \( [-Q^{\top}(\eta_0)Q(\eta_0)] \) have full column rank. We proceed by contradiction. Suppose that

\[
[-Q^{\top}(\eta_0)Q(\eta_0)] \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = -Q^{\top}(\eta_0)\alpha + Q(\eta_0)\beta = 0,
\]

(53)
where $\alpha$ and $\beta$ are vectors $\alpha = [\alpha_0 \ldots \alpha_{m-1}]^T$ and $\beta = [\beta_0 \ldots \beta_{m-1}]^T$. Then, (53) implies

$$B(q, \eta_0)\alpha(q) = A(q, \eta_0)\beta(q) \iff L(q, \theta_0)\alpha(q) = F(q, \theta_0)\beta(q).$$

where $\alpha(q) = \sum_{k=0}^{m-1} \alpha_k q^{-k}$ and $\beta(q) = \sum_{k=0}^{m-1} \beta_k q^{-k}$. Because $L(q, \theta_0)$ and $F(q, \theta_0)$ are coprime by Assumption 1 and polynomials of order $m_1 - 1$ and $m_T$, $\alpha(q)$ and $\beta(q)$ are polynomials of orders at most $m_T - 1$ and $m_1 - 1$, (54) can only be satisfied if $\alpha(q) \equiv 0 \equiv \beta(q)$. Hence, $[-Q^L_n(\eta_0) Q^S_n(\eta_0)]$ has full column rank.

Analogously, $[-Q^L_{\infty}(\eta_0) Q^S_{\infty}(\eta_0)]$ is full column rank if and only if $C(q, \theta_0)\alpha(q) = D(q, \theta_0)\beta(q)$ is satisfied only for $\alpha(q) \equiv 0 \equiv \beta(q)$, where $\alpha(q) = \sum_{k=0}^{m} \alpha_k q^{-k}$ and $\beta(q) = \sum_{k=0}^{m} \beta_k q^{-k}$. This is the case, as $C(q, \theta_0)$ and $D(q, \theta_0)$ are co-prime and polynomials of higher order than $\alpha(q)$ and $\beta(q)$. Hence, $[-Q^L_n(\eta_0) Q^S_n(\eta_0)]$ and $[-Q^L_{\infty}(\eta_0) Q^S_{\infty}(\eta_0)]$ are full column rank, implying that $Q_{\infty}(\eta_0)$ has a trivial right null-space and $M(\eta_0)$ is invertible.

Finally, we have the necessary results to prove Theorem 1.

**Proof of Theorem 1:** We start by using (23) to write

$$\hat{\theta}^S_N - \theta_0 = [Q^\top_n(\hat{\eta}_N)Q_n(\hat{\eta}_N)]^{-1}Q^\top_n(\hat{\eta}_N)[\hat{\eta}_N - N(\eta_0)\theta_0]$$

$$= [Q^\top_n(\hat{\eta}_N)Q_n(\hat{\eta}_N)]^{-1}Q^\top_n(\hat{\eta}_N)\hat{\eta}_N - Q_n(\eta_0)\theta_0$$

$$= [Q^\top_n(\hat{\eta}_N)Q_n(\hat{\eta}_N)]^{-1}Q^\top_n(\hat{\eta}_N)T_n(\theta_0)[\hat{\eta}_N - \eta_0^{(N)}].$$

(55)

where the last equality follows from (25). If $n$ were fixed, consistency would follow if $\hat{\eta}_N - \eta_0^{(N)}$ would approach zero as $N \to \infty$, provided the inverse of $Q^\top_n(\hat{\eta}_N)Q_n(\hat{\eta}_N)$ existed for sufficiently large $N$. However, $n = n(N)$ increases according to Assumption 4. This implies that the dimensions of the vectors $\hat{\eta}_N$ and $\eta_0^{(N)}$, and of the matrices $Q_n(\hat{\eta}_N)$ (number of rows) and $T_n(\theta_0)$ (number of rows and columns), become arbitrarily large. Therefore, extra requirements are necessary.

In particular, we use (55) to write

$$||\hat{\theta}^S_N - \theta_0|| = ||M^{-1}(\hat{\eta}_N)Q^\top_n(\hat{\eta}_N)T_n(\theta_0)[\hat{\eta}_N - \eta_0^{(N)}]||$$

$$\leq ||M^{-1}(\hat{\eta}_N)|| ||Q^\top_n(\hat{\eta}_N)|| ||T_n(\theta_0)|| ||\hat{\eta}_N - \eta_0^{(N)}||$$

(56)

where $M(\hat{\eta}_N) := [Q^\top_n(\hat{\eta}_N)Q_n(\hat{\eta}_N)]$. Consistency is achieved if the last factor on the right side of the inequality in (56) approaches zero as $N \to \infty$ w.p.1, and the remaining factors are bounded w.p.1 for sufficiently large $N$. This can be shown using (45), (46), and (40), but we need additionally that $M(\hat{\eta}_N)$ is invertible w.p.1 for sufficiently large $N$.

With this purpose, we write

$$||M(\hat{\eta}_N) - M(\eta_0^{(N)})|| = ||Q^\top_n(\hat{\eta}_N)Q_n(\hat{\eta}_N) - Q^\top_n(\eta_0^{(N)})Q_n(\eta_0^{(N)})||$$

Using (41), (43), (45), and Proposition 1, and because $M(\eta_0^{(N)}) \to M(\eta_0)$ as $n \to \infty$, we have that

$$M(\hat{\eta}_N) \to M(\eta_0) \text{ as } N \to \infty \text{ w.p.1.}$$

(57)

As $M(\eta_0)$ is invertible (Lemma 1), by (57) and because the map from the entries of a matrix to its eigenvalues is continuous, there is $N$ such that $M(\hat{\eta}_N)$ is invertible for all $N > N$ w.p.1.

Returning to (56), we may now write

$$||\hat{\theta}^S_N - \theta_0|| \leq C||\hat{\eta}_N - \eta_0^{(N)}|| \forall N > \bar{N}.$$

$$\to 0 \text{ as } N \to \infty \text{ w.p.1.}$$

(58)

Moreover, using (39), we can re-write (58) as

$$||\hat{\theta}^S_N - \theta_0|| \leq C||\hat{\eta}_N - \eta_0^{(N)}|| + ||\eta_0^{(N)} - \eta_0^{(N)}||.$$

From [46, Lemma 5.1], we have $||\eta_0^{(N)} - \eta_0^{(N)}|| \leq Cd(N)$, which approaches zero faster than $||\hat{\eta}_N - \eta_0^{(N)}||$, whose decay rate is according to [46, Theorem 5.1]. For the decay rate of $||\hat{\theta}^S_N - \theta_0||$, it suffices to take the rate of the slowest-decaying term, which is given by (35), as we wanted to show. □

**APPENDIX II
CONSISTENCY OF STEP 3**

The main purpose of this appendix is to prove Theorem 2. However, before we do so, we introduce some results regarding the norm of some vectors and matrices.

• $||R_n^S||$ is bounded for all $n$ and sufficiently large $N$ w.p.1

Let $R_n^S$ be defined as in (14). Then, from [46, Lemma 4.2], there exists $\bar{N}$ (a random value) such that w.p.1,

$$||R_n^S|| \leq C \forall n, \forall N > \bar{N}.$$

(59)

• $||T_n^1(\theta_0)||$ is bounded for all $n$

With $T_n(\theta)$ given by (26), the inverse of $T_n(\theta_0)$ is given by

$$T_n^{-1}(\theta) = \begin{bmatrix} T_n(\theta)^{-1} & 0 \\ T_n^T(\theta)^{-1}T_n(\theta)T_n^T(\theta)^{-1} & T_n^T(\theta)^{-1} \end{bmatrix}$$

evaluated at the true parameters $\theta_0$. Then,

$$||T_n^{-1}(\theta_0)|| \leq ||T_n^T(\theta_0)^{-1}|| + ||T_n(\theta_0)^{-1}||$$

$$+ ||T_n^T(\theta_0)^{-1}T_n(\theta_0)T_n^T(\theta_0)^{-1}||$$

$$\leq ||1/F(\theta)||_{\mathcal{H}_w} + ||1/C(\theta)||_{\mathcal{H}_w}$$

(61)

$$+ ||L(\theta)/[F(\theta)C(\theta)]||_{\mathcal{H}_w} \leq C \forall n,$$

where the last two inequalities follow from [56, Theorem 3] and from asymptotic stability of $1/F(\theta), 1/C(\theta)$, and $L(\theta)$.

• $||T_n^1(\hat{\theta}^S_N)||$ is bounded for all $n$ and sufficiently large $N$

Consider the matrix $T_n^{-1}(\hat{\theta}^S_N)$, given by (60) evaluated at $\hat{\theta}^S_N$. Proceeding as in (61), the term $||T_n^1(\hat{\theta}^S_N)||$ is bounded if $F(q, \hat{\theta}^S_N)$ and $C(q, \hat{\theta}^S_N)$ have all poles strictly inside the unit circle. From Theorem 1 and stability of the true system by Assumption 1, we conclude that there exists $N$ such that $F(q, \hat{\theta}^S_N)$ and $C(q, \hat{\theta}^S_N)$ have all roots strictly inside the unit circle for all $N > N$. Thus, we have

$$||T_n^{-1}(\hat{\theta}^S_N)|| \leq C \forall n, \forall N > \bar{N} \text{ w.p.1.}$$

(62)

• $||T_n^1(\hat{\theta}^S_N) - T_n^{-1}(\theta_0)||$ tends to zero as $N$ tends to infinity w.p.1

For the term $||T_n^1(\hat{\theta}^S_N) - T_n^{-1}(\theta_0)||$, with $n = n(N)$, we have

$$||T_n^1(\hat{\theta}^S_N) - T_n^{-1}(\theta_0)|| \leq ||T_n^1(\hat{\theta}^S_N) - T_n^1(\theta_0)|| + ||T_n^1(\theta_0) - T_n^{-1}(\theta_0)||$$

(63)
Because $||X|| \leq \sqrt{||X|| ||X||_\infty}$, where $X$ is an arbitrary matrix, we have that

$$
||T_n(\hat{\theta}_N^{LS}) - T_n(\theta_0)|| \leq \sum_{k=1}^{m} \left| \hat{\theta}_N^{LS} - \theta_0^k \right|
\leq C \left| \hat{\theta}_N^{\theta_0} - \theta_0 \right|,
$$

with superscript $k$ denoting the $k^{th}$ element of the vector, we can use Theorem 1 to show that

$$
||T_n(\hat{\theta}_N^{LS}) - T_n(\theta_0)|| = O\left(\sqrt{n(N) \log N} (1 + d(N)) \right). \quad (64)
$$

From Condition D3 in Assumption 4 and (32),

$$
\sqrt{n(N) \log N} (1 + d(N)) \to 0, \text{ as } N \to \infty,
$$

and thus $||T_n(\hat{\theta}_N^{LS}) - T_n(\theta_0)|| \to 0$ as $N \to \infty$ w.p.1. Together with (61), (62), and (63), this implies that

$$
||T_n^{-1}(\hat{\theta}_N^{LS}) - T_n^{-1}(\theta_0)|| \to 0, \text{ as } N \to \infty \text{ w.p.1.} \quad (65)
$$

The following two lemmas are useful for the invertibility of the weighted least-squares problem (30).

**Lemma 2.** Let Assumption 1 hold and

$$
\hat{M}(\eta_0, \theta_0) := \lim_{n \to \infty} Q_n^T(\eta_0^n) \hat{W}_n(\theta_0) Q_n(\eta_0^n),
$$

where $\hat{W}_n(\theta_0)$ is given by (29), and $Q_n(\eta_0^n)$ is defined by (22) at the true parameters $\eta_0^n$. Then, $\hat{M}(\eta_0, \theta_0)$ is invertible.

**Proof.** Using (15) and (29), we re-write (66) as

$$
\hat{M}(\eta_0, \theta_0) = \lim_{n \to \infty} Q_n^T(\eta_0^n)T_n^{-1}(\theta_0)Q_n(\eta_0^n). \quad (67)
$$

Re-writing $\varphi_n$, defined in (12), as

$$
\varphi_n = \begin{bmatrix} -F_n y_t \\ \Gamma_n u_t \end{bmatrix} = \begin{bmatrix} -\Gamma_n G_0(q) \\ -\Gamma_n H_0(q) \end{bmatrix},
$$

we can then write

$$
\hat{M}(\eta_0, \theta_0) = \lim_{n \to \infty} Q_n^T(\eta_0^n)T_n^{-1}(\theta_0)Q_n(\eta_0^n),
$$

where

$$
\Lambda_n(q) = \begin{bmatrix} -\Gamma_n G_0(q) & -\Gamma_n H_0(q) \end{bmatrix}.
$$

Then, we can re-write (67) as

$$
\hat{M}(\eta_0, \theta_0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Lambda_n(e^{i\omega}) \Phi_2 \Lambda_n^*(e^{i\omega}) d\omega,
$$

where $\Lambda_n(q)$ tends to zero. For the first, we have that

$$
||\hat{W}_n(\theta_0)|| \leq ||T_n^{-1}(\theta_0)|| ||\hat{R}|| \leq C, \quad (72)
$$

following from (59) and (61). For the second, with $\hat{W}_n(\hat{\theta}_N^{LS})$ given by (31) and $\hat{W}_n(\theta_0)$ by (29), conditions in Proposition 1 are satisfied using (59), (61), (65) and [46, Lemma 4.1], from which it follows that

$$
||\hat{W}_n(\hat{\theta}_N^{LS}) - \hat{W}_n(\theta_0)|| \to 0 \text{ as } N \to \infty \text{ w.p.1.} \quad (73)
$$

Having shown (72) and (73), the assumptions of Proposition 1 are verified, from which (71) follows and implies (70). \hfill \Box

We now have the necessary results to prove Theorem 2.

**Proof of Theorem 2:** Similarly to (55), we write

$$
\hat{\theta}_N^{WLS} - \theta_0 = \hat{M}^{-1}(\hat{\eta}_N, \hat{\theta}_N^{LS})Q_n^T(\hat{\eta}_N^\theta_0)W_n(\hat{\theta}_N^{LS})T_n(\hat{\theta}_0)(\hat{\eta}_N - \eta_0^n(N)), \quad (74)
$$

and analyze

$$
||\hat{\theta}_N^{WLS} - \theta_0|| \leq ||\hat{M}^{-1}(\hat{\eta}_N, \hat{\theta}_N^{LS})|| ||Q_n(\hat{\eta}_N)|| \cdot ||W_n(\hat{\theta}_N^{LS})|| ||T_n(\theta_0)|| ||\hat{\eta}_N - \eta_0^n(N)||.
$$

From Lemma 3, $M(\hat{\eta}_N, \hat{\theta}_N^{LS})$ converges to $\hat{M}(\eta_0, \theta_0)$, which is invertible from Lemma 2. Hence, because the map from
the entries of the matrix to its eigenvalues is continuous, $M(ˆ\eta_N, ˆ\theta^{LS}_N)$ is invertible for sufficiently large N, and therefore its norm is bounded, as it is a matrix of fixed dimensions. Also, from (45), $||Q_n(ˆ\eta_N)||$ is bounded for sufficiently large N. Moreover, we have that, making explicit that $n = n(N)$,

$$\left\| W_{n(N)}(\hat{\theta}^{LS}_N) \right\| \leq \left\| T_{n(N)}^{-1}(\hat{\theta}^{LS}_N) \right\|^2 \left\| R^n_{N}(N) \right\|.$$ 

Then, from (62) and (59), $\left\| W_{n(N)}(\hat{\theta}^{LS}_N) \right\| \leq C\sqrt{N} > \tilde{N}$. Finally, using also (43), (44), and (40), we conclude that

$$\left\| \hat{\theta}^{WLS}_N - \theta_0 \right\| \to 0 \text{ as } N \to \infty \text{ w.p.1.} \Box$$

**APPENDIX III**

**ASYMPTOTIC DISTRIBUTION AND COVARIANCE OF STEP 3**

The purpose of this appendix is to prove Theorem 3: asymptotic distribution and covariance of

$$\sqrt{N}(\hat{\eta}_N, \hat{\theta}^{LS}_N) = \sqrt{N}^{\eta} N(\hat{\eta}_N, \hat{\theta}^{LS}_N) (\hat{\eta}_N - \eta_0^{(N)}),$$

which is re-written from (74), where

$$Y^{\eta} N(\hat{\eta}_N, \hat{\theta}^{LS}_N) = M^{-1}(\hat{\eta}_N, \hat{\theta}^{LS}_N) Q^n_{\eta} N(\hat{\eta}_N) W_{n}(\hat{\theta}^{LS}_N) T_n(\theta_0).$$

If $Y^{\eta} N(\hat{\eta}_N, \hat{\theta}^{LS}_N)$ were of fixed dimensions, the standard idea would be to show that $Y^{\eta} N(\hat{\eta}_N, \hat{\theta}^{LS}_N)$ converges w.p.1 to a deterministic matrix, as consequence of $\hat{\eta}_N$ and $\hat{\theta}^{LS}_N$ being consistent estimates of $\eta_0$ and $\theta_0$, respectively. Then, for computing the asymptotic distribution and covariance of (75), one can consider the asymptotic distribution and covariance of $\sqrt{N}(\hat{\eta}_N - \eta_0^{(N)})$ while $Y^{\eta} N(\hat{\eta}_N, \hat{\theta}^{LS}_N)$ can be replaced by the deterministic matrix it converges to. This is a standard result (e.g., Slutsky’s theorem [57]), but it is not applicable here because the dimensions of $Y^{\eta} N(\hat{\eta}_N, \hat{\theta}^{LS}_N)$ and $\eta_0^{(N)}$ are not fixed. In this scenario, [46, Theorem 7.3] must be used instead. However, (75) is not ready to be used with [46, Theorem 7.3], because it requires $\hat{\eta}_N - \eta_0^{(N)}$ to be pre-multiplied by a deterministic matrix. The key idea of proving Theorem 3 is to show that (75) has the same asymptotic distribution and covariance as an expression of the form $Y^{\eta} N(\hat{\eta}_N, \eta_0^{(N)})$, where $Y^{\eta}$ is a deterministic matrix, and then apply [46, Theorem 7.3]. The following result will be useful for this purpose.

**Proposition 2.** Let $\hat{x}_N = \sqrt{N} \hat{A}_N \hat{B}_N \hat{\delta}_N$ be a finite-dimensional vector, where $\hat{A}_N$ and $\hat{B}_N$ are stochastic matrices and $\hat{\delta}_N$ is a stochastic vector of compatible dimensions. The dimensions may increase to infinity as function of N, except for the number of rows of $\hat{A}_N$, which is fixed. We assume that there is $\tilde{N}$ such that $||\hat{A}_N|| < C$ for all $N > \tilde{N}$, there is $B$ such that $||\hat{B}_N - \hat{B}_0|| \to 0$ as $N \to \infty$ w.p.1, and $||\hat{\delta}_N|| \to 0$ as $N \to \infty$ w.p.1. Then, if $\sqrt{N} ||\hat{B}_N - \hat{B}_0|| ||\hat{\delta}_N|| \to 0$ as $N \to \infty$ w.p.1, $\hat{x}_N$ and $\sqrt{N} \hat{A}_N \hat{B}_N \hat{\delta}_N$ have the same asymptotic distribution and covariance.

**Proof.** We begin by writing

$$\hat{x}_N = \sqrt{N} \hat{A}_N \hat{B}_N \hat{\delta}_N + \sqrt{N} \hat{A}_N (\hat{B}_N - B) \hat{\delta}_N.$$  

(76) Although some of the matrix and vector dimensions may increase to infinity with N, the number of rows of $\hat{A}_N$ is fixed, which makes $\hat{x}_N$ finite dimensional, to which Slutsky’s theorem may be applied. Then, this implies that $\hat{x}_N$ and $\sqrt{N} \hat{A}_N \hat{B}_N \hat{\delta}_N$ have the same asymptotic distribution and covariance if the second term on the right side of (76) tends to zero with probability one. By assumption, we have

$$\left| \frac{\sqrt{N} \hat{A}_N (\hat{B}_N - B) \hat{\delta}_N}{\sqrt{N} \hat{A}_N} \right| \leq \sqrt{N} ||\hat{A}_N|| ||\hat{B}_N - B|| ||\hat{\delta}_N|| \to 0$$

as $N \to \infty$ w.p.1. \Box

We now have the necessary results to prove Theorem 3.

**Proof of Theorem 3:** We start by re-writing (75) as

$$\sqrt{N}(\hat{\theta}^{WLS}_N - \theta_0) = M^{-1}(\hat{\eta}_N, \hat{\theta}^{LS}_N) x(\hat{\eta}_N, \hat{\theta}^{LS}_N),$$

where $M(\hat{\eta}_N, \hat{\theta}^{LS}_N)$ is defined in Lemma 3 and

$$x(\hat{\eta}_N, \hat{\theta}^{LS}_N) = \sqrt{N} Q_n(\hat{\eta}_N) W_n(\hat{\theta}^{LS}_N) T_n(\theta_0)(\hat{\eta}_N - \eta_0^{(N)}).$$

Both $M(\hat{\eta}_N, \hat{\theta}^{LS}_N)$ and $x(\hat{\eta}_N, \hat{\theta}^{LS}_N)$ are of fixed dimension, and from (69) and (70) $M^{-1}(\hat{\eta}_N, \hat{\theta}^{LS}_N) \to M^{-1}_C$ as $N \to \infty$ w.p.1. Then, if we assume that

$$x(\hat{\eta}_N, \hat{\theta}^{LS}_N) \sim ASN(0, P),$$

we have that, from [58, Lemma B.4],

$$\sqrt{N}(\hat{\theta}^{WLS}_N - \theta_0) \sim ASN\left(0, M^{-1} C P M^{-1} C \right).$$

(78) We will proceed to show that (78) is verified with

$$P = \sigma_0^2 \lim_{n \to \infty} Q_n(\eta_0^{(N)}) W_n(\theta_0) Q_n(\eta_0^{(N)}) = \sigma_0^2 M_{C},$$

(79) where the second equality follows directly from (66) and (69). We now proceed to show the first equality.

In the following arguments, we will apply Proposition 2 repeatedly to $x(\hat{\eta}_N, \hat{\theta}^{LS}_N)$. This requires the boundedness of some matrices; however, because all the matrices in $x(\hat{\eta}_N, \hat{\theta}^{LS}_N)$ have been shown to be bounded for sufficiently large N w.p.1, for readability we will refrain from referring to this every time Proposition 2 is applied.

Because it is more convenient to work with $\hat{\eta}_N^{(N)}$ than $\eta_0^{(N)}$, we start by re-writing $x(\hat{\eta}_N, \hat{\theta}^{LS}_N)$ as

$$x(\hat{\eta}_N, \hat{\theta}^{LS}_N) = \sqrt{N} Q_n(\hat{\eta}_N) W_n(\hat{\theta}^{LS}_N) T_n(\theta_0)(\hat{\eta}_N - \hat{\eta}_N^{(N)}) + Q_n^T(\hat{\eta}_N) W_n(\hat{\theta}^{LS}_N) T_n(\theta_0) \sqrt{N}(\eta_0^{(N)} - \hat{\eta}_N^{(N)}).$$

For sufficiently large N w.p.1, we have [46, Lemma 5.1]

$$\left\| Q_n^T(\hat{\eta}_N) W_n(\hat{\theta}^{LS}_N) T_n(\theta_0) \right\| \leq C \sqrt{N} d(\hat{\eta}_N^{(N)} - \hat{\eta}_N^{(N)})$$

(80) have the same asymptotic distribution and covariance, so we will analyze (80) instead.

Expanding $W_n(\hat{\theta}^{LS}_N)$ in (80), we obtain

$$\sqrt{N} Q_n^T(\hat{\eta}_N) W_n(\hat{\theta}^{LS}_N) T_n(\theta_0)(\hat{\eta}_N - \hat{\eta}_N^{(N)})$$

(81)

$$= \sqrt{N} Q_n^T(\hat{\eta}_N) T_n^{-1}(\hat{\eta}_N) R_n^{\theta} T_n^{-1}(\hat{\theta}^{LS}_N) T_n(\theta_0)(\hat{\eta}_N - \hat{\eta}_N^{(N)}).$$

Using Proposition 2, we conclude that (81) and

$$\sqrt{N} Q_n^T(\hat{\eta}_N) T_n^{-1}(\hat{\theta}^{LS}_N) R_n^{\theta} T_n^{-1}(\hat{\theta}^{LS}_N) T_n(\theta_0)(\hat{\eta}_N - \hat{\eta}_N^{(N)})$$

(82)
have the same asymptotic properties if
\[
\sqrt{N} \left| T_n^{-1}(\bar{\theta}_N^S) - T_n^{-1}(\theta_0) \right| \left\| \bar{\eta}_N - \bar{\eta}^n(N) \right\| \to 0
\]
as \( N \to \infty \) w.p.1. Using (61), (62) and (63) to write
\[
\sqrt{N} \left| T_n^{-1}(\bar{\theta}_N^S) - T_n^{-1}(\theta_0) \right| \left\| \bar{\eta}_N - \bar{\eta}^n(N) \right\|
\leq CN \left| T_n^{-1}(\bar{\theta}_N^S) - T_n^{-1}(\theta_0) \right| \left\| \bar{\eta}_N - \bar{\eta}^n(N) \right\|
\]
we have from (64) and [46, Theorem 5.1] that
\[
\sqrt{N} \left| T_n(\theta_0) - T_n(\bar{\theta}_N^S) \right| \left\| \bar{\eta}_N - \bar{\eta}^n(N) \right\| = O \left( \frac{n(N) \log N}{\sqrt{N}} \right) \left( 1 + d(N) \right) \right)^2.
\] (83)

where
\[
\frac{n(N) \log N}{\sqrt{N}} = \left( \frac{n^{3+\delta}(N)}{N} \right)^{\frac{1}{2+\delta}} \log N \to 0, \quad \text{as } N \to \infty,
\]
due to Condition D2 in Assumption 4. This implies that (81) and (82), and in turn \( \varepsilon(\bar{\eta}_N, \bar{\theta}_N^S) \), have the same asymptotic distribution and covariance. Repeating this procedure, it can be shown that, in turn, have the same asymptotic distribution and covariance as
\[
\sqrt{N} Q_n^T(\bar{\eta}_N) T_n^{-1}(\theta_0) R_n(\bar{\eta}_N - \bar{\eta}^n(N)). \quad (84)
\]

There are two stochastic matrices left in (84), which we need to replace by deterministic matrices that do not affect the asymptotic properties. Using [46, Lemma 4.1],
\[
\sqrt{N} \left| R_n^{\dagger} - \bar{R}_n^{\dagger} \right| \left\| \bar{\eta}_N - \bar{\eta}^n(N) \right\| =
\left( \frac{n^{3/2}(N)}{N} \right)^{\frac{3}{2+\delta}} \frac{\log N}{\sqrt{N}} \left( 1 + d(N) \right)
\]
\[+ C \sqrt{\frac{n^2(N) \log N}{N} \left( \frac{n^3(N)}{N} \right)^{\frac{1}{2+\delta}} \left( 1 + d(N) \right)}
\]
where the first term tends to zero by applying Condition D2 in Assumption 4 to
\[
\frac{n^{3/2}(N) \log N}{\sqrt{N}} = \left( \frac{n^{4+\delta}(N)}{N} \right)^{\frac{3}{2+\delta}} \frac{\log N}{N^{1/2+\delta}} \to 0, \quad \text{as } N \to \infty,
\]
and the second because of Condition D2 in Assumption 4, and (32). Then, from Proposition 2, we have that (84) and
\[
\sqrt{N} Q_n^T(\bar{\eta}_N) T_n^{-1}(\theta_0) R_n(\bar{\eta}_N - \bar{\eta}^n(N)), \quad (85)
\]
and in turn \( \varepsilon(\bar{\eta}_N, \bar{\theta}_N^S) \), have the same asymptotic distribution and covariance, so we will analyze (85).

Applying again Proposition 2, we have that (85) and
\[
\sqrt{N} Q_n^T(\bar{\eta}^n(N)) T_n^{-1}(\theta_0) R_n(\bar{\eta}_N - \bar{\eta}^n(N)) \quad (86)
\]
have the same asymptotic properties, since
\[
\sqrt{N} \left| Q_n(\bar{\eta}_N) - Q_n(\bar{\eta}^n(N)) \right| \left\| \bar{\eta}_N - \bar{\eta}^n(N) \right\| \leq C \sqrt{N} \left\| \bar{\eta}_N - \bar{\eta}^n(N) \right\|^2
\]
\[= O \left( \frac{n(N) \log N}{\sqrt{N}} \right) \left( 1 + d(N) \right)^2 \right)^2 \]
by using (41) and [46, Theorem 5.1], which tends to zero as \( N \to \infty \), identically to (83).

In (86), the matrix multiplying \( \bar{\eta}_N - \bar{\eta}^n(N) \) is now deterministic, but it will be more convenient to work with \( Q_n(\bar{\eta}^n(N)) \). With this purpose, [46, Lemma 5.1] can be used to show that (86) and
\[
\sqrt{N} Q_n^T(\bar{\eta}_N) T_n^{-1}(\theta_0) R_n(\bar{\eta}_N - \bar{\eta}^n(N)) \quad (87)
\]
have the same asymptotic properties, as
\[
\sqrt{N} \left| Q_n(\bar{\eta}^n(N)) - Q_n(\bar{\eta}^n(N)) \right| \left\| \bar{\eta}_N - \bar{\eta}^n(N) \right\|
\leq C \sqrt{N} \left\| \bar{\eta}_N - \bar{\eta}^n(N) \right\|^2
\]
\[= O \left( \frac{n(N) \log N}{\sqrt{N}} \right) \left( 1 + d(N) \right)^2 \right)^2 \]
which tends to zero due to (32) and Condition D3 in Assumption 4. Thus, \( \varepsilon(\bar{\eta}_N, \bar{\theta}_N^S) \) and (87) have the same asymptotic distribution and covariance, so we will analyze (87) instead.

Let \( T_n^\dagger := Q_n^T(\bar{\eta}_N) T_n^{-1}(\theta_0) R_n^T \). Then, using [46, Theorem 7.3], we have \( \sqrt{N} T_n^\dagger (\bar{\eta}_N - \bar{\eta}^n(N)) \sim AS_N(0, P) \), where P is given by (79). Finally, (36) follows from (69), (78), and (79).

REFERENCES

[1] E. J. Hannan. The asymptotic theory of linear time-series models. Journal of Applied Probability, 10(1):130–145, 1973.
[2] W. Dunsmuir and E. J. Hannan. Vector linear time series models. Advances in Applied Probability, 8(2):339–364, 1976.
[3] W. J. Hannan, W. Dunsmuir, and M. Deistler. Estimation of vector ARMAX models. Journal of Multivariate Analysis, 10(3):275–295, 1980.
[4] E. J. Hannan and M. Deistler. The statistical theory of linear systems. SIAM, 2012.
[5] K. Aström and T. Bohlin. Numerical identification of linear dynamic systems from normal operating records. In IFAC Symposium on Self-Adaptive Systems, Teddington, United Kingdom, 1965.
[6] L. Ljung. On the consistency of prediction error identification methods. In Mathematics in Science and Engineering, volume 126, pages 121–164. Elsevier, 1976.
[7] P. Caines and L. Ljung. Prediction error estimators: Asymptotic normality and accuracy. In 1976 IEEE Conference on Decision and Control, pages 652–658, 1976.
[8] L. Ljung System Identification. Theory for the User. Prentice-Hall, 1999.
[9] T. Söderström and P. Stoica. Instrumental variable methods for system identification. Springer-Verlag, 1983.
[10] P. Stoica and T. Söderström. Optimal instrumental variable estimation and approximate implementations. IEEE Transactions on Automatic Control, vol. 28, 757–772, July 1983.
[11] P. Young. The refined instrumental variable method: Unified estimation of discrete and continuous-time transfer function models. Journal Européen des Systèmes Automatisés, 42:149–179, 2008.
[12] M. Gilson and P. van den Hof. Instrumental variable methods for closed-loop system identification. Automatica, 41(2):241–249, 2005.
[13] S. Y. Kung. A new identification and model reduction algorithm via singular value decomposition. In Asilomar Conference on Circuits, Systems and Computers, pages 705–714, Pacific Grove, USA, 1978.
[14] P. van Overschee and B. de Moor. Subspace identification for linear systems: theory—implementation—applications. Kluwer Academic, Boston, 1996.
[15] M. Verhaegen. Application of a subspace model identification technique to identify LTI systems operating in closed-loop. Automatica, 29(4):1027–1040, 1993.
[16] S. Qin and L. Ljung. Closed-loop subspace identification with innovation estimation. In IFAC Symposium on System Identification, pages 861–866, Rotterdam, The Netherlands, 2003.
[17] M. Jansson. Subspace identification and ARX modeling. In 15th IFAC Symposium on System Identification, Rotterdam, The Netherlands, 2003.
[18] A. Chiuso and G. Picci. Consistency analysis of some closed-loop subspace identification methods. Automatica, 41(3):377–391, 2005.
[19] M. Jansson and B. Wahlberg. A linear regression approach to state-space subspace system identification. Signal Processing, 52:103–129, 1996.
[20] D. Bauer. Asymptotic properties of subspace estimators. Automatica, 41(3):359 – 376, 2005.
C. Sanathanan and J. Koerner. Transfer function synthesis as a ratio of two complex polynomials. *IEEE Transactions on Automatic Control*, 8(1):56–58, 1963.

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.

Y. Bresler and A. Macovski. Exact maximum likelihood parameter estimation of superimposed exponential signals in noise. *IEEE Transactions on Signal Processing*, 34(5):1081–1089, 1986.

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

A. K. Shaw, P. Misra, and R. Kumaresan. Identification of a class of multivariable systems from impulse response data: Theory and computational algorithm. *Circuits, Systems and Signal Processing*, 13(6):759–782, 1994.

P. Lemmerling, L. Vanhamme, S. van Huffel, and B. de Moor. IQML-like algorithms for solving structured total least squares problems: a unified view. *Signal Processing*, 81:1935–1945, 2001.

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

J. H. McClellan and D. Lee. Exact equivalence of the Steiglitz-McBride iteration and IQML. *IEEE Trans. Signal Process.*, 39(2):509–512, 1991.

P. Stoica and T. Söderström. The Steiglitz-McBride identification algorithm revisited-convergence analysis and accuracy aspects. *IEEE Transactions on Automatic Control*, 26(5):712–717, 1981.

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

M. Zorzi. An interpretation of the dual problem of the three-like approaches. *Automatica*, 62:87–92, 2015.

J. Durbin. The fitting of time-series models. *Revue de l'Institut International de Statistique*, pages 233–244, 1960.

J. Durbin. Efficient estimation of parameters in moving-average models. *Biometrika*, 46(3/4):306–316, 1959.

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

E. J. Hannah and L. Kavaleris. Linear estimation of ARMA processes. *Automatica*, 19(4):447–448, 1983.

E. Haman and L. Kavaleris. Multivariate linear time series models. *Advances in Applied Probability*, 16(3):492–561, 1984.

G. Reinsel, S. Basu, and S. Yap. Maximum likelihood estimators in the multivariate autoregressive moving-average model from a generalized least squares viewpoint. *J. Time Series Anal.*, 13(2):133–145, 1992.

D. Poskitt and M. Salau. On the relationship between generalized least squares and gaussian estimation of vector ARMA models. *J. Time Series Anal.*, 16(6):617–645, 1995.

J. M. Dufour and T. Jouini. Asymptotic distributions for quasi-efficient estimators in echelon VARMA models. *Computational Statistics & Data Analysis*, 73:69–86, 2014.

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

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

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

E. Verri, M. Galrinho, and H. Hjalmarsson. Open-loop asymptotically efficient model reduction with the Steiglitz-McBride method. *Automatica*, 89:221–234, 2018.

M. Galrinho, C. R. Rojas, and H. Hjalmarsson. A weighted least-squares method for parameter estimation in structured models. In *IEEE Conf. on Decision and Control*, pages 3322–3327, Los Angeles, USA, 2014.

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

L. Ljung and B. Wahlberg. Asymptotic properties of the least-squares method for estimating transfer functions and disturbance spectra. *Advances in Applied Probabilities*, 24:412–440, 1992.

M. Viberg and B. Ottersten. Sensor array processing based on subspace fitting. *IEEE Trans. on Signal Processing*, 39(5):1110–1121, 1991.

C. Gourieroux and A. Monfort. *Statistics and econometric models*, volume 2. Cambridge University Press, 1995.

T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation*. Prentice-Hall, 2000.

M. Galrinho, C. R. Rojas, and H. Hjalmarsson. Estimating models with high-order noise dynamics using semi-parametric Weighted Null-Space Fitting, *submitted to Automatica* (arXiv: 1708.03947).

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

W. Larimore. System identification, reduced-order filtering and modeling via canonical variate analysis. In *American Control Conference*, 1985.

K. Peternel, W. Scherrer, and M. Deistler. Statistical analysis of novel subspace identification methods. *Signal Processing*, 52:161–177, 1996.

P. van Overschee and B. de Moor. N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, 30:75–93, 1994.

P. Stoica and M. Janssen. MIMO system identification: state-space and subspace approximations versus transfer function and instrumental variables. *IEEE Trans. on Signal Processing*, 48(11):3087–3099, 2002.

C. R. Rojas, T. Oomen, H. Hjalmarsson, and B. Wahlberg. Analyzing iterations in identification with application to nonparametric $H_\infty$-norm estimation. *Automatica*, 48(11):2776–2790, 2012.

A. Gut, *Probability; a graduate course*, volume 75. Springer Science & Business Media, 2013.

T. Söderström and P. Stoica. *System Identification*. Prentice Hall, 1989.

Miguel Galrinho was born in 1988. He received his M.S. degree in aerospace engineering in 2013 from Delft University of Technology, The Netherlands, and the Ph.D. degree in electrical engineering in 2018 from KTH Royal Institute of Technology, Stockholm, Sweden. His research interests are in multi-step least-squares methods for system identification.

Cristian R. Rojas (M’13) was born in 1980. He received the M.S. degree in electronics engineering from the Universidad Técnica Federico Santa María, Valparaiso, Chile, in 2004, and the Ph.D. degree in electrical engineering at The University of Newcastle, NSW, Australia, in 2008.

Since October 2008, he has been with the Royal Institute of Technology, Stockholm, Sweden, where he is currently Associate Professor at the Department of Automatic Control, School of Electrical Engineering. His research interests lie in system identification and signal processing.

Håkan Hjalmarsson (M’98–SM’11–F’13) was born in 1962. He received the M.S. degree in electrical engineering in 1988, and the Licentiate and Ph.D. degrees in automatic control in 1990 and 1993, respectively, all from Linköping University, Linköping, Sweden.

He has held visiting research positions at California Institute of Technology, Louvain University, and at the University of Newcastle, Australia. His research interests include system identification, signal processing, control and estimation in communication networks and automated tuning of controllers.

Dr. Hjalmarsson has served as an Associate Editor for *Automatica* (1996–2001) and for the IEEE Transactions on Automatic Control (2005–2007), and has been Guest Editor for the European Journal of Control and Control Engineering Practice. He is a Professor at the School of Electrical Engineering, KTH, Stockholm, Sweden. He is a Chair of the IFAC Coordinating Committee CC1 Systems and Signals. In 2001, he received the KTH award for outstanding contribution to undergraduate education. He is co-recipient of the European Research Council advanced grant.