Linear Prediction Error Methods for Stochastic Nonlinear Models ⋆

Mohamed Rasheed-Hilmy Abdalmoaty, Håkan Hjalmarsson

Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Malvinas väg 10, floor 6, SE-10044 Stockholm, Sweden

Abstract

The estimation problem for stochastic parametric nonlinear dynamical models is recognized to be challenging. The main difficulty is the intractability of the likelihood function and the optimal one-step ahead predictor. In this paper, we present relatively simple prediction error methods based on non-stationary predictors that are linear in the outputs. They can be seen as extensions of the linear prediction error methods for the case where the hypothesized model is stochastic and nonlinear. The resulting estimators are defined by analytically tractable objective functions in several common cases. It is shown that, under certain identifiability and standard regularity conditions, the estimators are consistent and asymptotically normal. We discuss the relationship between the suggested estimators and those based on second-order equivalent models as well as the maximum likelihood method. The paper is concluded with a numerical simulation example as well as a real-data benchmark problem.

Key words: Parameter estimation; System identification; Stochastic systems; Nonlinear models; Prediction error methods.

1 Introduction

System identification of linear dynamical systems is a well-developed and well-understood subject. During the last five decades, methods and algorithms based on stochastic as well as deterministic frameworks have been developed and used. The availability of many devoted monographs [22,27,64,44,40,54] as well as software packages [36,48,31,52] is a clear indication of the maturity of the subject. In principle, linear system identification may be used to construct linear models even when the underlying system is nonlinear [41,17,57,58]; however, when the results are not satisfactory, nonlinear models have to be identified.

Unfortunately, the estimation problem for stochastic nonlinear models can be quite challenging. General nonlinear transformations of unobserved disturbances render commonly used estimation methods—such as the Maximum Likelihood (ML) method—analytically intractable. Until recently, the main body on system identification of nonlinear models considered model structures with an explicit correspondence between observations and innovations such that predictors and likelihood functions are (relatively) easy to compute; for example, NARX and NARMAX models belong to this type of model structures. Several methods have been developed under that assumption to address problems such as model structure selection, parameterization, and initialization; see the surveys [7,24,63,42,60], the articles [29,61,53,62] and the books [47,20,8,45]. It was not until the last decade that such a restrictive assumption on the model was relaxed.

It has been shown in [25] that estimators obtained by ignoring disturbances passing through the nonlinear system may not be consistent. Since then, there has been a growing interest in the estimation problem for stochastic nonlinear models. An ML method and a Prediction Error Method (PEM) for stochastic Wiener models, when the unobserved disturbance is independent over time, have been developed in [26] and [67], respectively. In [66] a performance analysis and approximate estimation methods based on Taylor approximations were considered; however, due to this approximation, the obtained estimators may not be consistent. A solution to the ML problem for general stochastic nonlinear state-space models was proposed in [56]. It relied on a Monte Carlo Expectation-Maximization (MCEM) algorithm [68] where the E-step was approximated by a Sequential

⋆ This work was supported by the Swedish Research Council via the projects NewLEADS (contract number: 2016-06079) and System identification: Unleashing the algorithms (contract number: 2015-05285). Parts of this paper have appeared in [1] and [2]. Corresponding author M. Abdalmoaty.

Email addresses: abda@kth.se (Mohamed Rasheed-Hilmy Abdalmoaty), hjalmars@kth.se (Håkan Hjalmarsson).

Postprint of an article published in Automatica, 105:49-63, 2019. Please cite the published version.

Monte Carlo (SMC) smoother [16] (also known as particle smoother). A PEM estimator based on the optimal Mean-Square Error (MSE) one-step ahead predictor was suggested in [49]. In [70], a MCEM algorithm, in the same spirit as [56], was used; but this time a rejection sampling based particle smoother [14] was employed. These methods, however, can be computationally expensive: to be convergent, they require the number of used particles in the SMC smoother to increase with the iterations of the optimization algorithm [18].

The current state-of-the-art algorithm for off-line ML estimation of general nonlinear state-space models was outlined in [34]. It is based on a combination of a stochastic approximation Expectation-Maximization algorithm [13] and an SMC smoother known as the conditional particle filter with ancestor sampling (CPF-AS) [35]. The CPF-AS is an SMC sampler similar to a standard auxiliary particle filter [16] with the difference that one particle at each time step is set deterministically. The resulting method is asymptotically efficient and convergence to an ML estimate can be established [32]. More recently, an algorithm for on-line ML estimation has been proposed in [51]. It employs a recently developed on-line SMC smoother [50] to approximate the gradient of the predictive densities of the state (also known as tangent filters/filter sensitivity [11, Section 10.2.4]) which are then used to update the parameter estimate.

These methods have been shown to provide interesting results on several benchmark problems. However, their application is so far limited to cases where fundamental limitations of SMC algorithms—such as particle degeneracy (see [15,16])—can be avoided. For example, they are not directly applicable when the measurement noise variance is small; in this case a modified algorithm has to be used [65]. Furthermore, the convergence of the Expectation-Maximization algorithm may be very slow if the variance of the latent process is small. Moreover, the estimation of high-dimensional models is still out of reach. These limitations are currently the topic of active research within different communities including system identification; see for example [46] and [69].

1.1 Contributions

In this paper, we introduce and analyze a PEM based on predictors that are linear in the past outputs. The use of these predictors can be motivated by Wold's decomposition of general second-order non-stationary processes (see Appendix A). It has been noticed in [1] that their use corresponds to a partial probabilistic model. They rely on the second-order properties of the model and the computations of the exact likelihood function are not required. Therefore, they are relatively easy to compute, and can be highly competitive in this respect compared to estimators based on SMC smoothing algorithms. We show that they may be given in terms of closed-form expressions for several common cases, and Monte Carlo approximations are not necessarily required. The difference between the proposed predictors and linear predictors based on second-order equivalent models [41,17] is described. Furthermore, the convergence and consistency of the resulting PEM estimators is established under standard regularity and certain identifiability conditions. The price paid for bypassing the computations of the likelihood function is a loss of statistical asymptotic efficiency. Nevertheless, it is possible to improve the asymptotic properties of the resulting estimator by one iteration of a Newton-Raphson scheme. This requires the evaluation of the gradient vector and the Hessian matrix of the log-likelihood function, and may be achieved by *a single run* of a particle smoothing algorithm, e.g., a conditional particle filter [35]. As is well known, this refined estimator is asymptotically first-order equivalent to the maximum likelihood estimator [33, Chapter 6].

1.2 Paper outline

We start in Section 2 by introducing a stochastic framework and formulating the main problem. In Section 3, we introduce one-step ahead optimal and suboptimal linear predictors for a general class of nonlinear stochastic models. The relationship between these predictors and predictors obtained using second-order equivalent models is discussed. In Section 4, linear PEM estimators are defined; their consistency and asymptotic normality are established under standard conditions in Section 5 and Section 6. A maximum likelihood interpretation is given in Section 7. In Section 8, a numerical simulation example as well as a recent real-data benchmark problem are used to demonstrate the performance of the proposed estimators. The paper is concluded in Section 9. Finally, Appendix A gives a brief overview of Wold's decomposition of second-order non-stationary processes.

1.3 Notations

Bold font is used to denote random quantities and regular font is used to denote realizations thereof. The triplet $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$ denotes a generic underlying probability space on which the output process $\boldsymbol{y}$ is defined; here, Ω is the sample space, $\mathcal{F}$ is the basic σ-algebra, and $\mathbb{P}_\theta$ is a probability measure parameterized by a finite-dimentional real vector θ and an a priori known input signal u. The symbols $\mathbb{E}[\cdot;\theta]$, $\mathbf{var}(\cdot;\theta)$ and $\mathbf{cov}(\cdot,\cdot;\theta)$ denote the mathematical expectation, variance and covariance operators with respect to $\mathbb{P}_\theta$. The space $\mathsf{L}_2^n(\Omega, \mathcal{F}, \mathbb{P}_\theta)$ is the Hilbert space of $\mathbb{R}^n$-valued random vectors with finite second moments [9, Chapter 2]; for brevity, we simply use L_2^n. The notation $\boldsymbol{x} \sim p(\boldsymbol{x})$ is used to mean that the random variable $\boldsymbol{x}$ is distributed according to the probability density function $p(\boldsymbol{x})$. For a matrix M, the notation $[M]_{ij}$ denotes the ij^{th}-entry of M, and when M is real and symmetric, $M \succ 0$ means that M is positive definite. Finally for any vector v, $\|v\|_M^2 := v^\top M v$.

2 Problem Formulation

In this section, we define the used stochastic framework and formulate the main problem of the paper.

2.1 Signals

The outputs and disturbances are all modeled using discrete-time stochastic processes. In other words, we assume that the observed data is embedded in an infinite sequence of potential observations. The output signal $\boldsymbol{y} := \{\boldsymbol{y}_t : t \in \mathbb{Z}\}$ is modeled as an $\mathbb{R}^{d_y}$-valued discrete-time stochastic process defined over $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$, $d_y \in \mathbb{N}$. The probability measure $\mathbb{P}_\theta$ is parameterized by a finite-dimensional parameter θ, assuming values in a compact subset $\Theta \subset \mathbb{R}^d$, and an *a priori known* d_u-dimensional input signal $u := \{u_t : t \in \mathbb{Z}\}$, $d_u \in \mathbb{N}$. Hence, the underlying dynamical system is necessarily operating in open-loop, and all unknown disturbances are stochastic processes. The models to be developed are deterministic functions that define a mapping between these processes such that they completely specify $\mathbb{P}_\theta$. We will only consider purely non-deterministic processes that have finite second-order moments: $\boldsymbol{y} \subset \mathsf{L}_2^{d_y}$; see Appendix A.

One of the simplest second-order stochastic processes is white noise. In this paper, white noise is defined as a sequence of uncorrelated random variables with zero mean and finite variance. This definition is quite weak: it does not specify the distribution of the process and neither stationarity nor independence are assumed. However, it is sufficient for our purposes since the proposed methods do not require the use of a full probabilistic model.

2.2 Mathematical Models

We consider the class of discrete-time causal dynamical models given by the following definition.

Definition 1 (Stochastic parametric nonlinear model) *A stochastic parametric nonlinear model is defined by the relations*

$$\boldsymbol{y}_t = f_t(\{u_k\}_{k=1}^{t-1}, \{\boldsymbol{\zeta}_k\}_{k=1}^{t}; \theta), \tag{1}$$

in which $t = 1, \ldots, N$, and $\theta \in \Theta$ is a parameter to be identified, and $\{\boldsymbol{\zeta}_k\}_{k=1}^{t}$ is a subsequence of an unobserved $\mathbb{R}^{d_\zeta}$-valued stochastic process, whose distribution may be parameterized by θ, such that $\{\boldsymbol{y}_k\}_{k=1}^{N}$ is a subsequence of a second-order stochastic process $\boldsymbol{y}$.

This definition emphasizes the input-output nature of our approach. The resulting model class is fairly general: it covers a wide range of static models as well as most of the commonly used dynamic model structures. Consider, for example, a stochastic nonlinear time-varying state-space model [40, Section 5.3] defined by the relations

$$\begin{aligned} \boldsymbol{x}_{t+1} &= h_t(\boldsymbol{x}_t, u_t, \boldsymbol{w}_t; \theta), \quad \boldsymbol{x}_1 \sim p_{\boldsymbol{x}_1}(\theta), \quad \boldsymbol{w}_t \sim p_{\boldsymbol{w}_t}(\theta), \\ \boldsymbol{y}_t &= g_t(\boldsymbol{x}_t, \boldsymbol{v}_t; \theta), \qquad \boldsymbol{v}_t \sim p_{\boldsymbol{v}_t}(\theta), \quad t \in \mathbb{N}, \end{aligned}$$

in which $\boldsymbol{x}$ is the state process, and $\boldsymbol{w}$ and $\boldsymbol{v}$ are unobserved disturbances and noise. Define the process $\boldsymbol{\zeta}$ as

$$\boldsymbol{\zeta}_1 := \begin{bmatrix} \boldsymbol{x}_1^\top & \boldsymbol{v}_1^\top \end{bmatrix}^\top, \qquad \boldsymbol{\zeta}_t := \begin{bmatrix} \boldsymbol{w}_{t-1}^\top & \boldsymbol{v}_t^\top \end{bmatrix}^\top \quad \forall t > 1.$$

Then the functions $\{h_t\}$ and $\{g_t\}$ determine a model of the form in (1) as follows

$$\begin{aligned} f_1(\boldsymbol{\zeta}_1; \theta) &:= g_1(\boldsymbol{x}_1, \boldsymbol{v}_1; \theta), \\ f_2(u_1, \{\boldsymbol{\zeta}_k\}_{k=1}^{2}; \theta) &:= g_2(h_1(\boldsymbol{x}_1, u_1, \boldsymbol{w}_1; \theta), \boldsymbol{v}_2; \theta), \\ &\vdots \\ f_t(\{u_k\}_{k=1}^{t-1}, \{\boldsymbol{\zeta}_k\}_{k=1}^{t}; \theta) &:= \\ g_t(h_{t-1}(\ldots h_1(\boldsymbol{x}_1, u_1, \boldsymbol{w}_1; \theta) \ldots, u_{t-1}, \boldsymbol{w}_{t-1}; \theta), \boldsymbol{v}_t; \theta). \end{aligned}$$

2.3 The problem

Define the data set

$$\boldsymbol{D}_t := \{(\boldsymbol{y}_k, u_k) : k = 1, \ldots, t\}, \tag{2}$$

that contains pairs of inputs and outputs up to time $t \in \mathbb{N}$. We will assume that the data is generated by a known model structure, i.e., known functions $\{f_t\}$ in (1) parameterized by an unknown *true parameter* $\theta^\circ \in \Theta$. Thus, we will not be concerned here with the important problem of model structure selection [1].

Assumption 2 (True system) *The sequence of data sets $\{\boldsymbol{D}_t\}_{t=1}^{\infty}$ follows a known model structure (1) with an unknown parameter $\theta^\circ \in \Theta$.*

The problem studied in the paper is the construction of a point estimate $\hat{\theta}$ of the parameter vector θ° based on a given realization of the data set $\boldsymbol{D}_N$.

One of the favored and commonly used point estimators is the ML Estimator (MLE) whose computations require the evaluation of the likelihood function of θ at the observed data. While this can be done efficiently for Gaussian linear models, see for example [4], the likelihood function for the model in (1) is, in general, analytically intractable. Let us define the vectors

$$\boldsymbol{Y} := \begin{bmatrix} \boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_N^\top \end{bmatrix}^\top, \quad \boldsymbol{Z} := \begin{bmatrix} \boldsymbol{\zeta}_1^\top, \ldots, \boldsymbol{\zeta}_N^\top \end{bmatrix}^\top,$$

and assume the existence of a known joint probability density function $p(\boldsymbol{Y}, \boldsymbol{Z}; \theta)$. Then the likelihood function of θ is given by the high-dimensional marginalization integral

$$p(Y; \theta) = \int_{\mathbb{R}^{d_z}} p(Y, Z; \theta)\, \mathrm{d}Z, \tag{3}$$

in which $d_z = d_\zeta N$ is the dimension of Z. An alternative to the ML method is a PEM based on the optimal MSE one-step ahead predictor. Unfortunately, for the stochastic nonlinear model in (1), such a predictor is in general analytically intractable. It is given by

$$\hat{\boldsymbol{y}}_{t|t-1}(\theta) = \int_{\mathbb{R}^{d_y}} y_t\, p(y_t | \boldsymbol{Y}_{t-1}; \theta)\, \mathrm{d}y_t \quad \forall t \in \mathbb{Z}, \tag{4}$$

[1] The convergence of the proposed estimators can be established even when θ° does not exist or $\theta^\circ \notin \Theta$; see Section 5.

where $p(\boldsymbol{y}_t|\boldsymbol{Y}_{t-1};\theta)$ is the predictive density of $\boldsymbol{y}_t$, $\boldsymbol{Y}_{t-1} := \left[\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_{t-1}^\top\right]^\top$, and $\boldsymbol{Y}_0$ is defined as the empty set. Observe that by Bayes' theorem (see [28, page 39]),

$$\begin{aligned} p(y_t|Y_{t-1};\theta) &= \frac{p(Y_t;\theta)}{\int_{\mathbb{R}^{d_y}} p(Y_t;\theta)\,\mathrm{d}y_t} \\ &= \frac{\int_{\mathbb{R}^{d_t}} p(Y_t, Z_t;\theta)\,\mathrm{d}Z_t}{\int_{\mathbb{R}^{d_y}}\int_{\mathbb{R}^{d_t}} p(Y_t, Z_t;\theta)\,\mathrm{d}Z_t\,\mathrm{d}y_t}. \end{aligned} \tag{5}$$

where $d_t = d_\zeta t$, and thus, except in very few cases, are analytically intractable. Hence, it seems that a PEM based on the optimal MSE one-step ahead predictor does not have any computational advantage over the asymptotically efficient ML method. Both the MLE and the conditional mean of the output require the solution of similar intractable marginalization integrals. While ignoring the unobserved disturbance may lead to closed form predictors, it is well known that the resulting PEM estimator is not guaranteed to be consistent [26]. For this reason, most of the recent research efforts found in the system identification literature target the MLE.

In this contribution, we consider PEMs based on relatively simple suboptimal predictors; they are used to construct consistent and asymptotically normally distributed estimators. The obtained results can be seen as extensions of the linear case and can be motivated by Wold's decomposition (see Appendix A).

3 Linear Predictors for Nonlinear Models

The general prediction problem can be described as follows: at time $t-1$, we have observed the outputs $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}$ for some $t \in \mathbb{N}$ and wish to estimate a value for, the next output, $\boldsymbol{y}_t$. In general, for a known input u and a given θ, a one-step ahead predictor may be defined as a measurable function of $\boldsymbol{Y}_{t-1}$, usually chosen to minimize some criteria. As pointed out above, the optimal MSE predictor is a common choice; however, in general, it is given by the intractable integral (4). Instead, in this paper, we consider a class of predictors that are linear in the past outputs and has the form

$$\hat{\boldsymbol{y}}_{t|t-1}(\theta) = \tilde{\mu}_t(U_{t-1};\theta) + \sum_{k=1}^{t-1} \tilde{l}_{t-k}(t, U_{t-1};\theta)\boldsymbol{y}_k, \quad t \in \mathbb{N},$$

where $\tilde{\mu}_t$ and $\tilde{l}_{t-k}$ are, possibly nonlinear, functions in θ and the known vector of inputs

$$U_{t-1} := \left[u_1^\top, \ldots, u_{t-1}^\top\right]^\top.$$

Observe that the dependence of the predictor on u is implicit in the notation.

Linear predictors are much easier to work with; a unique linear Minimum MSE (MMSE) predictor for any second-order process always exists among the set of linear predictors (see Lemma 4 below). The computations are also straightforward, and closed-form expressions for the predictors may be available in several common cases.

3.1 The Optimal Linear Predictor (OL-predictor)

By considering the outputs of the model in (1) as elements of the Hilbert space $\mathsf{L}_2^{d_y}$, the projection theorem (see [72] or [3]) can be used to define the linear MMSE one-step ahead predictor. This is a standard result of Hilbert spaces; the key idea is that such a predictor can be thought of as the unique orthogonal projection of $\boldsymbol{y}_t$ onto the closed subspace spanned by the entries of $\boldsymbol{Y}_{t-1}$ when the MSE is used as an optimality criterion.

Definition 3 (Linear MMSE one-step ahead predictor) *Let $\mathcal{S} \subset \mathsf{L}_2^{d_y}$ be the closed subspace spanned by the entries of $\boldsymbol{Y}_{t-1}$. Then, a linear Minimum MSE (MMSE) predictor of $\boldsymbol{y}_t$ in $\mathcal{S}$ is defined as a vector $\hat{\boldsymbol{y}}_{t|t-1} \in \mathcal{S}$ such that*

$$\mathbb{E}\left[\|\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}\|_2^2;\theta\right] \leq \mathbb{E}\left[\|\boldsymbol{y}_t - \tilde{\boldsymbol{y}}\|_2^2;\theta\right] \quad \forall \tilde{\boldsymbol{y}} \in \mathcal{S}.$$

A characterization of such a predictor is given in the following classical lemma. Note that all the expectations are functions of the input which is assumed to be known and deterministic.

Lemma 4 (Existence and uniqueness) *The linear MMSE one-step ahead predictor defined in Definition 3 exists and is unique. It is given by*

$$\hat{\boldsymbol{y}}_{t|t-1}(\theta) = \mathbb{E}[\boldsymbol{y}_t;\theta] + \Psi_t(U_{t-1};\theta)\left(\boldsymbol{Y}_{t-1} - \mu_{t-1}(U_{t-1};\theta)\right), \tag{6}$$

for $1 < t \leq N$, where $\mu_{t-1}(U_{t-1};\theta) := \mathbb{E}[\boldsymbol{Y}_{t-1};\theta]$, $\boldsymbol{Y}_{t-1} = [\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_{t-1}^\top]^\top$ and $\Psi_t(U_{t-1};\theta)$ is given by any solution to the normal equations

$$\Psi_t(U_{t-1};\theta)[\mathbf{cov}(\boldsymbol{Y}_{t-1}, \boldsymbol{Y}_{t-1};\theta)] = \mathbf{cov}(\boldsymbol{y}_t, \boldsymbol{Y}_{t-1};\theta). \tag{7}$$

Furthermore, $\hat{\boldsymbol{y}}_{1|0}(\theta) = \mathbb{E}[\boldsymbol{y}_1;\theta]$.

PROOF. See [72].

For brevity, we will refer to the linear MMSE predictor in (6) as *the Optimal Linear predictor (OL-predictor).*

Remark 5 *Observe that the coefficients in (7), which are used in the expression of the OL-predictor, depend only on the unconditional first and second moments of $\boldsymbol{y}$ up to time t. therefore, the computations of the OL-predictor can be simpler than that of the unrestricted optimal predictor (the conditional mean) that, as shown in Section 2.3, required computing the integrals (4) and (5).*

To connect Lemma 4 to Wold's decomposition of $\boldsymbol{y}$, note that the predictor in (6) would be easy to compute if the

matrices $\mathbf{cov}(\boldsymbol{Y}_{t-1}, \boldsymbol{Y}_{t-1}; \theta)$ were diagonal. This holds only if the output vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}$ are orthogonal (uncorrelated), which is rarely the case in most applications. Nevertheless, the Gram-Schmidt procedure (see [30]) can be used to (causally) transform the output vectors into a set of orthogonal vectors $\{\boldsymbol{\varepsilon}_k\}$ such that

$$\begin{aligned} \boldsymbol{\varepsilon}_t(\theta) &:= \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}(\theta), \quad 1 \le t \le N, \\ &= \boldsymbol{y}_t - \mathbb{E}[\boldsymbol{y}_t; \theta] - \sum_{k=1}^{t-1} \mathbf{cov}(\boldsymbol{y}_t, \boldsymbol{\varepsilon}_k; \theta) \lambda_{\boldsymbol{\varepsilon}_k}^{-1} \boldsymbol{\varepsilon}_k(\theta), \end{aligned} \tag{8}$$

with $\boldsymbol{\varepsilon}_1(\theta) = \boldsymbol{y}_1 - \mathbb{E}[\boldsymbol{y}_1; \theta]$, and $\lambda_{\boldsymbol{\varepsilon}_k} = \mathbf{cov}(\boldsymbol{\varepsilon}_k, \boldsymbol{\varepsilon}_k; \theta)$.

Let $\boldsymbol{\mathcal{E}}_{t-1} := [\boldsymbol{\varepsilon}_1^\top \ldots \boldsymbol{\varepsilon}_{t-1}^\top]^\top$. Then, for linear prediction, the vectors $\boldsymbol{\mathcal{E}}_{t-1}$ and $\boldsymbol{Y}_{t-1}$ are equivalent in the sense that they span the same subspaces. Thus, under the assumption that all signals are known to be zero for $t \le 0$, the above construction is identical to Wold's decomposition (see the third row of (A.2) and compare to (8)).

The vector $\boldsymbol{\varepsilon}_t$ is known as the innovation in $\boldsymbol{y}_t$ (see [12]).

Definition 6 (The (linear) innovation process) *The linear innovation process of $\boldsymbol{y}$ is defined as*

$$\boldsymbol{\varepsilon}_t(\theta) := \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}(\theta), \quad t \in \mathbb{Z},$$

where $\hat{\boldsymbol{y}}_{t|t-1}(\theta)$ is the OL-predictor defined in (6).

The next lemma concerns the computations of the OL-predictor. It shows that finding the predictors and the innovations corresponds to a (block) $LDL^\top$ factorization (see [21, Chapter 4]) of the covariance matrix of $\boldsymbol{Y} := [\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_N^\top]^\top$. We will use the notation $U := U_N$.

Lemma 7 (Computations of the OL-predictor) *Consider the general nonlinear model in (1) such that $\boldsymbol{y}_t = 0 \; \forall t \le 0$. Suppose that*

$$\begin{aligned} \mu(U; \theta) &:= \mathbb{E}[\boldsymbol{Y}; \theta], \\ \Sigma(U; \theta) &:= \mathbf{cov}(\boldsymbol{Y}, \boldsymbol{Y}; \theta) \succ 0 \end{aligned} \tag{9}$$

are given. Then the unique OL-predictor of $\boldsymbol{y}_t$, $t = 1, \ldots, N$, is given by

$$\hat{\boldsymbol{y}}_{t|t-1}(\theta) = \mathbb{E}[\boldsymbol{y}_t; \theta] + \sum_{k=1}^{t-1} \tilde{l}_{t-k}(t, U_{t-1}; \theta) \left(\boldsymbol{y}_k - \mathbb{E}[\boldsymbol{y}_k; \theta]\right) \tag{10}$$

in which $\tilde{l}_j(t, U_{t-1}; \theta) := \left[L^{-1}(U; \theta)\right]_{tj}$, where the matrix $L(U; \theta)$ is the unique (block) lower unitriangular matrix[2] given by the (block) $LDL^\top$ factorization of Σ; that is,

$$\Sigma(U; \theta) =: L(U; \theta) \Lambda(U; \theta) L^\top(U; \theta). \tag{11}$$

Moreover, $\hat{\boldsymbol{y}}_{1|0}(\theta) = \mathbb{E}[\boldsymbol{y}_1; \theta]$ and the vector of OL-predictors is given by

$$\begin{aligned} \widehat{\boldsymbol{Y}}(\theta) &:= \left[\hat{\boldsymbol{y}}_{1|0}^\top(\theta) \ldots \hat{\boldsymbol{y}}_{N|N-1}^\top(\theta)\right]^\top \\ &= \boldsymbol{Y} - L^{-1}(U; \theta)(\boldsymbol{Y} - \mu(U; \theta)). \end{aligned} \tag{12}$$

PROOF. To establish (10), first recall that whenever the covariance matrix Σ is positive definite, the factorization in (11) is unique (see [21, Theorem 4.1.3]). Then observe that, using Wold's decomposition or (8), we may write

$$\boldsymbol{Y} = \mu(U; \theta) + \tilde{L}(U; \theta) \boldsymbol{\mathcal{E}}, \tag{13}$$

$$\tilde{L}(U; \theta) = \begin{bmatrix} I & 0 & \ldots & 0 \\ \mathbf{cov}(\boldsymbol{y}_2, \boldsymbol{\varepsilon}_1) \lambda_{\boldsymbol{\varepsilon}_1}^{-1} & I & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{cov}(\boldsymbol{y}_N, \boldsymbol{\varepsilon}_1) \lambda_{\boldsymbol{\varepsilon}_1}^{-1} & \mathbf{cov}(\boldsymbol{y}_N, \boldsymbol{\varepsilon}_2) \lambda_{\boldsymbol{\varepsilon}_2}^{-1} & \ldots & I \end{bmatrix}.$$

From (13), due to the linearity of the expectation operator, $\mathbf{cov}(\boldsymbol{Y}, \boldsymbol{Y}; \theta) = \tilde{L}(U; \theta) \tilde{\Lambda}(U; \theta) \tilde{L}^\top(U; \theta)$. Consequently, the uniqueness of the factorization in (11) implies that $\tilde{L}(U; \theta) = L(U; \theta)$, and $\tilde{\Lambda}(U; \theta) = \Lambda(U; \theta)$ is a block diagonal matrix of innovation covariances. Now, observe that it is possible to compute the innovations vector by inverting the unitriangular matrix L (which is always invertible for any finite N) to get

$$\boldsymbol{\mathcal{E}}(\theta) = L^{-1}(U; \theta)(\boldsymbol{Y} - \mu(U; \theta)), \tag{14}$$

and by definition (see (8)) we have

$$\boldsymbol{\mathcal{E}}(\theta) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\theta). \tag{15}$$

Therefore the vector of OL-predictors is given by

$$\begin{aligned} \widehat{\boldsymbol{Y}}(\theta) &= \boldsymbol{Y} - L^{-1}(U; \theta)(\boldsymbol{Y} - \mu(U; \theta)) \\ &= (I - L^{-1}(U; \theta))\boldsymbol{Y} + L^{-1}(U; \theta)\mu(U; \theta) \end{aligned}$$

from which (10) follows after making use of the unitriangular form of $L^{-1}(U; \theta)$.

The computations of the innovations in Lemma 7 are similar to that of the standard Kalman filter in the linear case (see [30]); however, an important difference here is the dependence on the used input in L and Λ.

Remark 8 *Wold's decomposition implies that $\hat{\boldsymbol{y}}_{t|t-1}$ is well defined in terms of the innovation process as $t \to \infty$ (Theorem 28). However, an invertibility condition on $\boldsymbol{y}$ with respect to the linear innovations need to be imposed in order to be able to compute the OL-predictor in terms of the data as $N \to \infty$ (see Section 5.1.1).*

3.2 The Output-Error predictor (OE-predictor)

In order to define a sensible linear predictor without using an optimality criteria, we first recall how a suboptimal predictor may be defined in the linear case. Suppose that $\boldsymbol{y}_t = G(\mathrm{q}; \theta) u_t + H(\mathrm{q}) \boldsymbol{\varepsilon}_t$, where $G(\mathrm{q}; \theta)$ is a stable transfer operator, $\boldsymbol{\varepsilon}$ is white noise, and q is the forward-shift operator (see [5]). Then, it is well known that if

[2] A lower unitriangular matrix is a lower triangular matrix whose main diagonal entries are equal to the identity matrix.

the data is collected in open-loop, and when standard regularity and identifiability conditions hold, a PEM estimator based on the Output-Error (OE) predictor, $\hat{y}_t(\theta) = G(\mathrm{q};\theta)u_t$, is consistent [40, Theorem 8.4]. Notice that this predictor neither requires the specification of the exact noise model $H(\mathrm{q})$, nor the distribution of $\boldsymbol{\varepsilon}$. The only used information regarding the probabilistic structure of the model is the mean of its output. It is thus possible to generalize the above observation to a large class of stochastic nonlinear models whose output has a finite mean, such as the model in (1). This leads us to the following definition.

Definition 9 (The OE-predictor) *Consider the general model in (1). The Output-Error predictor (OE-predictor) of $\boldsymbol{y}_t$ is defined as the deterministic quantity*

$$\hat{y}_t(\theta) := \mathbb{E}[\boldsymbol{y}_t;\theta], \quad t \in \mathbb{N}. \tag{16}$$

The predictor in (16), although deterministic and independent of $\boldsymbol{Y}_{t-1}$, is different from the "nonlinear simulation predictor" [40, Section 5.3, page 147] which is defined by fixing $\{\boldsymbol{\zeta}_k\}_{k=1}^t$ in (1) to zero and taking $\hat{y}_t(\theta) = f_t(\{u_k\}_{k=1}^{t-1}, 0\,;\theta)$. Instead, the OE-predictor (16) averages the output over all *possible values of the unobserved disturbances.*

Both the OL-predictor and the OE-predictor may be computed in terms of closed-form expressions in several common cases—which are usually considered challenging—as illustrated in the following example.

Example 10 (Linear predictors for a scalar stochastic Wiener model) *Consider a stochastic Wiener model defined by the relations*

$$\boldsymbol{y}_t = \beta(u_t + \boldsymbol{w}_t)^2 + \frac{1}{1 - a\mathrm{q}^{-1}}\boldsymbol{v}_t - 2\beta, \tag{17}$$

$t = 1, \dots, N$, in which $\beta \in \mathbb{R}$, u is a known input signal, and $\boldsymbol{w}$ and $\boldsymbol{v}$ are unobserved independent white noises with time-independent variances denoted by λ_w and λ_v respectively. Let $\theta := [\beta\ \lambda_w\ \lambda_v]^\top$ and suppose that a is known such that $|a| < 1$ and that all signals are scalars.

Observe that the full distribution of $\boldsymbol{v}$ is not specified in the model. However, for the clarity of the exposition, assume that $\boldsymbol{w}$ is a Gaussian process and let $\boldsymbol{w}$ and $\boldsymbol{v}$ be mutually independent. Moreover, note that even when a full probabilistic model is hypothesized, both the likelihood function of θ and the optimal MSE predictor of $\boldsymbol{y}$ are analytically intractable (see e.g. [16]). However, as we now show, the mean and the covariance of the model's output can be computed in terms of closed-form expressions.

The model in (17) may be written in vector form as

$$\boldsymbol{Y} = \beta(U + \boldsymbol{W})^2 + H\boldsymbol{V} - 2\beta\mathbf{1}$$

in which $\mathbf{1}$ denotes a vector of ones,

$$\boldsymbol{W} := \begin{bmatrix} \boldsymbol{w}_1 \\ \vdots \\ \boldsymbol{w}_N \end{bmatrix}, \quad \boldsymbol{V} := \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_N \end{bmatrix}, \quad H := \begin{bmatrix} 1 & 0 & \dots & 0 \\ a & 1 & \dots & 0 \\ \vdots & & \ddots & \\ a^{N-1} & a^{N-2} & \dots & 1 \end{bmatrix},$$

and the exponent is applied entry-wise; i.e., for a vector X we define $X^2 := \left[x_1^2, \dots, x_N^2\right]^\top$. Then, it is straightforward to see that the mean of $\boldsymbol{Y}$ is given by

$$\begin{aligned} \mu(U;\theta) = \mathbb{E}[\boldsymbol{Y};\theta] &= \mathbb{E}[\beta(U + \boldsymbol{W})^2;\theta] - 2\beta\mathbf{1} \\ &= \beta(U^2 + \lambda_w\mathbf{1}) - 2\beta\mathbf{1}. \end{aligned} \tag{18}$$

Because $\boldsymbol{W}$ and $\boldsymbol{V}$ are independent, the covariance matrix of $\boldsymbol{Y}$ is given by

$$\begin{aligned} \Sigma(U;\theta) &= \mathbf{cov}(\boldsymbol{Y}, \boldsymbol{Y};\theta) \\ &= \beta^2\mathbf{cov}\left((U + \boldsymbol{W})^2, (U + \boldsymbol{W})^2\right) + \mathbf{cov}(H\boldsymbol{V}, H\boldsymbol{V}) \\ &= D(U;\theta) + \lambda_v HH^\top, \end{aligned} \tag{19}$$

where $D(U;\theta)$ is a diagonal matrix with entries

$$[D(U;\theta)]_{tt} = 2\beta^2\lambda_w(2u_t^2 + \lambda_w), \qquad t = 1, \dots N, \tag{20}$$

because of the assumption that $\boldsymbol{W} \sim \mathcal{N}(0, \lambda_w I_N)$. Therefore, the vector of OE-predictors is given by

$$\widehat{Y}(\theta) := \beta(U^2 + (\lambda_w - 2)\mathbf{1}), \tag{21}$$

and, by Lemma 7, the vector of OL-predictors is given by (12) which, using (18)-(20), is equal to

$$\widehat{\boldsymbol{Y}}(\theta) = \boldsymbol{Y} - L^{-1}(U;\theta)(\boldsymbol{Y} - \beta(U^2 + (\lambda_w - 2)\mathbf{1})),$$

where $L(U;\theta)$ is given by the $LDL^\top$ factorization of $\Sigma(U;\theta)$.

Observe that due to the nonlinearity of the model, the covariance matrix of $\boldsymbol{Y}$ depends on the input (unlike the case of linear models). Moreover, note that the predictors are parameterized by β as well as λ_w and λ_v.

A straightforward extension of the model in (17)—that does not affect the discussion—is to let $\boldsymbol{w}$ be a linearly filtered Gaussian white noise and assume a parameterized input; for example, $u_t(\theta) := G(\mathrm{q};\theta)\tilde{u}_t$ for some transfer operator G and known signal $\tilde{u}$.

In the following section, we discuss the relationship between the proposed predictors and linear predictors obtained based on LTI second-order equivalent models.

3.3 Relation to LTI Second-Order Equivalent Models

Linear time-invariant approximations of nonlinear systems are usually considered under different sets of assumptions and objectives [41,17,58]. They are generally

studied in an MSE framework where assumptions and restrictions on the systems to be approximated are implicitly given as assumptions on the input and output signals. It is commonly assumed that the inputs and the outputs are zero mean stationary stochastic processes, such that the input belongs to a certain class; for example, a class of periodic processes, or processes that have a specific spectrum. In such a framework, explicit assumptions on the underlying data generating mechanism (such as a parametric nonlinear model) are not necessarily used or required. The goal there is to use spectral assumptions on the data to obtain an LTI model—*linear in both* $\boldsymbol{y}$ *and* $\boldsymbol{u}$—that approximate the behavior of the underlying nonlinear system. Once a model is computed, it might be used to construct a predictor of $\boldsymbol{y}_t$ that is *linear in the past inputs and outputs.*

An Output-Error LTI Second-Order Equivalent (OE-LTI-SOE) model is defined in [17, Section 4.2] as

$$G_{\mathrm{OE}}(\mathrm{q}) := \arg\min_{G\in\mathcal{G}} \mathbb{E}\left[\|\boldsymbol{y}_t - G(\mathrm{q})\boldsymbol{u}_t\|^2\right], \tag{22}$$

where $\mathcal{G}$ is the set of stable and causal LTI models, the expectation operator is with respect to the *joint distribution* of $\boldsymbol{y}$ and $\boldsymbol{u}$. When the stability and causality constraints are dropped, the minimizer is called the best linear approximation (BLA) [54]. Note that an OE-LTI-SOE model only captures the causal part of the cross-covariance function between $\boldsymbol{y}$ and $\boldsymbol{u}$. A better approximation is obtained by a General-Error LTI-SOE (GE-LTI-SOE) model defined as (see [17, Section 4.4] or [41])

$$\begin{aligned}(G_{\mathrm{GE}}, H_{\mathrm{GE}}) := \arg\min_{G,H} \mathbb{E}\left[\|H^{-1}(\mathrm{q})(\boldsymbol{y}_t - G(\mathrm{q})\boldsymbol{u}_t)\|^2\right] \\ \text{such that} \quad H^{-1}(\mathrm{q}),\ H^{-1}(\mathrm{q})G(\mathrm{q}) \in \mathcal{G}.\end{aligned} \tag{23}$$

It captures the second-order properties in terms of the covariance function of $\boldsymbol{y}$ and the cross-covariance function between $\boldsymbol{y}$ and $\boldsymbol{u}$. In other words, the process

$$\tilde{\boldsymbol{y}}_t := G_{\mathrm{GE}}(\mathrm{q})\boldsymbol{u}_t + H_{\mathrm{GE}}(\mathrm{q})\tilde{\boldsymbol{\varepsilon}}_t \tag{24}$$

has exactly the same spectrum as $\boldsymbol{y}$, where $\tilde{\boldsymbol{\varepsilon}}$ is a stationary white noise with variance

$$\lambda_0 := \mathbb{E}\left[\|H_{\mathrm{GE}}^{-1}(\mathrm{q})(\boldsymbol{y}_t - G_{\mathrm{GE}}(\mathrm{q})\boldsymbol{u}_t)\|^2\right].$$

By definition, LTI-SOE models depend on the assumed distribution of the input process. Notice that the models in (22) and (23) are defined by averaging, not only over $\boldsymbol{y}$, *but also over all realizations of the input* $\boldsymbol{u}$. Therefore, one has to speak of an LTI-SOE model "with respect to a certain class of input signals". In this contribution, by contrast, the inputs are assumed fixed and known. They are used to describe the mean and covariance functions of $\boldsymbol{y}$, which is not necessarily stationary, and therefore all the computations are conditioned on the given input. To further clarify these important remarks, we have the following example.

Example 11 (LTI-SOE predictor models) *Consider the model of Example 10 and let* $a = 0$ *so that*

$$\boldsymbol{y}_t = \beta(\boldsymbol{u_t} + \boldsymbol{w}_t)^2 + \boldsymbol{v}_t - 2\beta. \tag{25}$$

Suppose that $\boldsymbol{u}$, $\boldsymbol{w}$ *and* $\boldsymbol{v}$ *are independent and mutually independent zero mean stationary Gaussian processes with unit variances. Then* $\boldsymbol{y}$ *is a zero mean stationary process. Now observe that due to the independence assumptions*

$$\begin{aligned}\mathbb{E}[\boldsymbol{y}_t\boldsymbol{u}_{t-\tau}] &= 0 \ \ \forall \tau > 1,\\ \mathbb{E}[\boldsymbol{y}_t\boldsymbol{u}_t] &= \mathbb{E}[\beta\boldsymbol{u}_t^3 + \beta\boldsymbol{u}_t\boldsymbol{w}_t^2 + 2\beta\boldsymbol{u}_t^2\boldsymbol{w}_t + \boldsymbol{u_t}\boldsymbol{v}_t - 2\beta\boldsymbol{u_t}]\\ &= 0,\end{aligned}$$

and therefore the cross-spectrum between $\boldsymbol{y}$ *and* $\boldsymbol{u}$ *is* $\Phi_{\boldsymbol{yu}}(z) = 0$. *Moreover, straightforward calculations show that the spectra of* $\boldsymbol{u}$ *and* $\boldsymbol{y}$ *are given by* $\Phi_{\boldsymbol{u}}(z) = 1$, *and* $\Phi_{\boldsymbol{y}}(z) = 8\beta^2 + 1$. *Consequently, the OE-LTI-SOE model in (22) is given by [17, Corollary 4.1]*

$$G_{OE}(\mathrm{q}) = \frac{\Phi_{\boldsymbol{yu}}(\mathrm{q})}{\Phi_{\boldsymbol{u}}(\mathrm{q})} = 0,$$

which is independent of β *and* $\Phi_{\boldsymbol{u}}(\mathrm{q})$. *Similarly, the GE-LTI-SOE model in (23) is given by [17, Theorem 4.5]*

$$G_{GE}(\mathrm{q}) = 0, \quad H_{GE}(\mathrm{q}) = 1,$$

and $\lambda_0 = 8\beta^2 + 1$. *Therefore, in this case,* $\tilde{\boldsymbol{y}}_t = H_{GE}(\mathrm{q})\tilde{\boldsymbol{\varepsilon}}_t$ *has exactly the same spectrum as* $\boldsymbol{y}$, *and the optimal linear predictor constructed based on either LTI-SOE models is independent of* β *and* $\boldsymbol{u}$; *it is given by*

$$\hat{y}_{t|t-1} = (1 - H_{GE}^{-1}(\mathrm{q}))\boldsymbol{y}_t + H_{GE}^{-1}(\mathrm{q})G_{GE}(\mathrm{q})\boldsymbol{u}_t = 0.$$

On the other hand, the OL-predictor and the OE-predictor suggested in this paper are defined by conditioning on the assumed known (realization of the) input. As shown in Example 10, assuming that u *is known, the mean and the covariance of the model's output are given by (18) and (19). When* $a = 0$ *and* $\lambda_w = \lambda_v = 1$, *the mean of* $\boldsymbol{y}_t$ *becomes* $\mathbb{E}[\boldsymbol{y}_t; \theta] = \beta(u_t^2 - 1)$, *and its variance becomes* $\mathbf{var}(\boldsymbol{y}_t) = 2\beta^2(2u_t^2 + 1) + 1$. *Hence, Wold's decomposition of* $\boldsymbol{y}$ *(given* u*) is*

$$\boldsymbol{y}_t = \beta(u_t^2 - 1) + H_t(\mathrm{q};\beta)\boldsymbol{\varepsilon}_t, \ \ \mathbf{var}(\boldsymbol{\varepsilon}_t) = 1, \tag{26}$$

in which $H_t(\mathrm{q};\beta)$ *is a time-varying filter with impulse response coefficients*

$$h_k(t) = 0 \ \ \forall k \geq 1, \ \ h_0(t) = \sqrt{2\beta^2(2u_t^2 + 1) + 1} \ \ \forall t \in \mathbb{Z}.$$

Note that here, because $a = 0$, $\boldsymbol{y}$ *is an independent process and the OL-predictor coincides with the unrestricted optimal MSE predictor as well as the unconditional mean:*

$$\hat{y}_{t|t-1}(\beta) = \mathbb{E}[\boldsymbol{y}_t; \beta] = \mathbb{E}[\boldsymbol{y}_t|\boldsymbol{Y}_{t-1}; \beta] = \beta(u_t^2 - 1),$$

which is nonlinear in u_t.

Thus, the main difference between the models in (24) and (26) is how the input is handled. While LTI-SOE models are defined by averaging over a stationary input, the model in (26) is obtained by conditioning on a given realization, typically leading to a non-stationary model.

4 Linear PEM Estimators

We now define four PEM estimators based on the predictors defined in the previous section. Their asymptotic analysis is given in Section 5 and Section 6.

The first estimator is based on the OL-predictor and the squared Euclidean norm; we will refer to it as the OL-QPEM (OL-predictor Quadratic PEM) estimator.

Definition 12 (The OL-QPEM estimator) *The OL-QPEM estimator is defined as*

$$\begin{aligned}\hat{\theta}(\boldsymbol{D}_N) &= \arg\min_{\theta\in\Theta} \ \|\boldsymbol{Y}-\widehat{\boldsymbol{Y}}(\theta)\|^2 \\ &\textit{where} \quad \widehat{\boldsymbol{Y}}(\theta)=\boldsymbol{Y}-L^{-1}(U;\theta)(\boldsymbol{Y}-\mu(U;\theta)),\end{aligned} \tag{27}$$

in which μ and L are defined in (9) and (11).

Note that, by the definition of the OL-predictor, it holds that the expectation of the criterion function in (27) is minimized at θ°, i.e.,

$$\mathbb{E}[\|\boldsymbol{Y}-\widehat{\boldsymbol{Y}}(\theta)\|_2^2;\theta^\circ] \geq \mathbb{E}[\|\boldsymbol{Y}-\widehat{\boldsymbol{Y}}(\theta^\circ)\|_2^2;\theta^\circ] \quad \forall\theta\in\Theta,$$

and whenever U and the parameterization of μ and L is such that $\widehat{\boldsymbol{Y}}(\theta^\circ)=\widehat{\boldsymbol{Y}}(\theta) \implies \theta^\circ=\theta$, the true parameter θ° is a unique minimizer.

Observe that in the classical case of LTI models, the OL-QPEM estimator is nothing more than the commonly used PEM estimator defined by the Euclidean norm and the optimal linear one-step ahead predictor

$$\begin{aligned}\hat{\boldsymbol{y}}_{t|t-1}(\theta) &= \boldsymbol{y}_t - H^{-1}(\mathrm{q};\theta)(\boldsymbol{y}_t - G(\mathrm{q};\theta)u_t), \\ &= G(\mathrm{q};\theta)u_t + \sum_{k=1}^{t-1}\tilde{h}_{t-k}(\theta)\,(\boldsymbol{y}_k - G(\mathrm{q};\theta)u_k)\end{aligned} \tag{28}$$

where $G(\mathrm{q};\theta)$ is the plant model and $\{\tilde{h}_k(\theta)\}$ is the impulse response of the inverted noise model $H^{-1}(\mathrm{q};\theta)$ [40, Chapter 3]. In that case, $[\mu(U;\theta)]_t = G(\mathrm{q};\theta)u_t$ and $[L^{-1}(U;\theta)]_{ij} = \tilde{h}_{|i-j|}(\theta)$. Furthermore, conditions on u and the parameterization are given by the concepts of informative experiment and identifiability [40,64].

The second estimator is based on the OL-predictor and a weighted time- and θ-*dependent* criterion function; we will refer to it as the OL-GPEM (OL-predictor Gaussian PEM) estimator because the criterion function is on the form of a Gaussian log-likelihood function.

Definition 13 (The OL-GPEM estimator) *The OL-GPEM estimator is defined as*

$$\begin{aligned}\hat{\theta}(\boldsymbol{D}_N) &= \arg\min_{\theta\in\Theta} \ \|\boldsymbol{Y}-\widehat{\boldsymbol{Y}}(\theta)\|^2_{\Lambda^{-1}(U;\theta)} + \log\det\Lambda(U;\theta) \\ &\textit{where} \quad \widehat{\boldsymbol{Y}}(\theta)=\boldsymbol{Y}-L^{-1}(U;\theta)(\boldsymbol{Y}-\mu(U;\theta)),\end{aligned} \tag{29}$$

in which μ, L and Λ are defined in (9) and (11).

Observe that the used criterion function is both input- and θ-dependent via the *linear* innovation covariance matrices. The log det term is important for the consistency of the estimator due to the dependence of the weighting matrix $\Lambda(U;\theta)$ on θ. As with the OL-QPEM problem, the properties of the OL-predictor imply that the expected value of the criterion function in (29) is minimized at θ° (see, for example, [10, (3.1) and (3.2)]).

The third estimator is based on the OE-predictor and the squared Euclidean norm; we will refer to it as the OE-QPEM (OE-predictor Quadratic PEM) estimator.

Definition 14 (The OE-QPEM estimator) *The OE-QPEM estimator is defined as*

$$\begin{aligned}\hat{\theta}(\boldsymbol{D}_N) &= \arg\min_{\theta\in\Theta} \quad \|\boldsymbol{Y}-\widehat{Y}(\theta)\|^2 \\ &\textit{where} \quad \widehat{Y}(\theta)=\mu(U;\theta),\end{aligned} \tag{30}$$

in which μ is defined in (9).

Once more, the expected value of the criterion function in (30) is minimized at θ°, since $\mu(U;\theta^\circ)$ is the optimal MSE predictor of $\boldsymbol{Y}$ (given zero initial conditions). Note that the criterion function in (30) can be weighted using a θ-independent positive definite matrix to potentially improve the asymptotic properties of the estimator. In that case we refer to it as the OE-WQPEM (OE-predictor Weighted Quadratic PEM) estimator.

Definition 15 (The OE-WQPEM estimator) *Let M be a given θ-independent bounded positive definite matrix. The OE-WQPEM estimator corresponding to M is defined as*

$$\begin{aligned}\hat{\theta}(\boldsymbol{D}_N) &= \arg\min_{\theta\in\Theta} \quad \|\boldsymbol{Y}-\widehat{Y}(\theta)\|^2_M \\ &\textit{where} \quad \widehat{Y}(\theta)=\mu(U;\theta),\end{aligned} \tag{31}$$

in which μ is defined in (9).

In the next two sections, we show that the general asymptotic theory of the PEMs is applicable to the proposed estimators. The asymptotic results are based on the original work of Ljung in [39] and Ljung and Caines in [43] where the dependence structure of the processes is specified in a generic form in terms of an "exponential forgetting" hypothesis [38].

5 Convergence and Consistency

Let us denote the normalized PEM criterion function by

$$\boldsymbol{\mathcal{V}}_N(\theta) := \frac{1}{N}\sum_{t=1}^{N} \ell(\boldsymbol{e}_t(\theta), t; \theta), \tag{32}$$

in which $\boldsymbol{e}(\theta)$ is the Prediction Error (PE) process (the difference between the observed and predicted $\boldsymbol{y}$), $\ell(\boldsymbol{e}_t(\theta), t; \theta) = \|\boldsymbol{e}_t(\theta)\|^2$ for the OL-QPEM and the OE-QPEM,

$$\ell(\boldsymbol{e}_t(\theta), t; \theta) = \boldsymbol{e}_t^\top(\theta)\Lambda_t^{-1}(U_t; \theta)\boldsymbol{e}_t(\theta) + \log\det\Lambda_t(U_t; \theta)$$

for the OL-GPEM (where $\boldsymbol{e} = \boldsymbol{\varepsilon}$, the linear innovations), and $\ell(\boldsymbol{e}_t(\theta), t; \theta) = \boldsymbol{e}_t^\top(\theta)M_t\boldsymbol{e}_t(\theta)$ for the OE-WQPEM where M_t is a known θ-independent bounded positive definite matrix [3]. The corresponding PEM estimators, defined in Section 4, are then given by

$$\hat{\boldsymbol{\theta}}_N = \arg\min_{\theta\in\Theta} \boldsymbol{\mathcal{V}}_N(\theta), \quad N \in \mathbb{N}.$$

The classical asymptotic analysis usually involves the study of the asymptotic behavior of the sequence of criterion functions $\{\boldsymbol{\mathcal{V}}_N(\theta) : N \in \mathbb{N},\ \theta \in \Theta\}$ and the use of a compactness assumption on the parameter set Θ to control the corresponding process of global minimizers $\{\hat{\boldsymbol{\theta}}_N : N \in \mathbb{N},\ \theta \in \Theta\}$. As far as the prediction error framework is concerned, the simplest cases are those involving (quasi-)stationary ergodic processes such that the sequence of criterion functions converges uniformly over Θ to a well-defined deterministic limit; namely,

$$\sup_{\theta\in\Theta}\left|\boldsymbol{\mathcal{V}}_N(\theta) - \bar{\mathcal{V}}(\theta)\right| \xrightarrow{\text{a.s.}} 0 \quad \text{as} \quad N \to \infty, \tag{33}$$

such that the limit $\bar{\mathcal{V}}(\theta)$ is continuous over Θ and has a unique global minimizer θ^*. The symbol $\xrightarrow{\text{a.s.}}$ denotes the almost sure convergence [3]. In general, the limit in (33) depends on the system and the input properties. Under identifiability conditions and a compactness assumption on Θ, it is straightforward to conclude, when (33) holds, that $\hat{\boldsymbol{\theta}}_N \xrightarrow{\text{a.s.}} \theta^* = \theta^\circ$ as $N \to \infty$. These are essentially the arguments used in the convergence and consistency proofs in an ergodic environment (see [40, Chapter 8] for the LTI case).

In a general non-stationary environment however, the sequence of criterion functions does not necessarily converge to any limit and may very well be divergent. These cases are of interest particularly when the predictors are non-stationary or when the user cannot control the identification experiment to ensure their convergence. Nevertheless, it is possible to establish the convergence of the minimizers by showing that $\boldsymbol{\mathcal{V}}_N(\theta)$ asymptotically behaves like the averaged criterion $\mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)]$ uniformly in θ. This is the main idea of the convergence and consistency analysis developed in [37,39]. Below, we discuss sufficient regularity conditions regarding the data, the predictor and the used criterion, given in [39], when applied to the linear PEMs proposed in this paper.

5.1 Conditions on the data generating mechanism

For the convergence of PE methods, it is sufficient that the dependence of the moments of $\boldsymbol{y}$ upon the history of the process decays at an exponential rate. It will be assumed that Assumption 2 holds, and therefore the terms "model" and "system" are used interchangeably.

Definition 16 (r-stability) *A discrete-time causal dynamical model of $\boldsymbol{y}$ is said to be r-stable with some $r > 1$, if for all $s, t \in \mathbb{Z}$ such that $s \le t$ there exist doubly-indexed random variables $\{\boldsymbol{y}_{t,s} : \boldsymbol{y}_{t,t} = 0\}$ such that*

(1) $\boldsymbol{y}_{t,s}$ is a (measurable) function of $\{\boldsymbol{y}_k\}_{k=s+1}^{t}$ and independent of $\{\boldsymbol{y}_k\}_{k=-\infty}^{s}$,
(2) for some positive real numbers $c < \infty$ and $\lambda < 1$, it holds that

$$\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_{t,s}\|^r\right] < c\lambda^{t-s}. \tag{34}$$

The outputs of *r-stable* models form a class of stochastic processes known as *r-mean exponentially stable processes* or *exponentially forgetting processes of order r* [38]. Observe that the definition implies that $|\mathbb{E}[\|\boldsymbol{y}_t\|^r]| < c \ \ \forall t \in \mathbb{Z},\ r > 1$, and therefore the output of an r-stable model must have a bounded mean. Generally speaking, the random variables $\boldsymbol{y}_{t,s}$ can be interpreted as the outputs of the system when the underlying basic stochastic process $\boldsymbol{\zeta}$ is replaced by $\{\boldsymbol{\zeta}_{t,s}\}_{t\in\mathbb{Z}}$ such that $\{\boldsymbol{\zeta}_{t,s}\}_{t<s}$ are given by a value independent of $\{\boldsymbol{\zeta}_t\}_{t<s}$, say zero, but $\boldsymbol{\zeta}_{t,s} := \boldsymbol{\zeta}_t\ \forall t > s$. Note that the above definition of stability includes the conventional stability definition of dynamical systems. For example, in the case of LTI rational models, the output process is exponentially stable when all the poles of the model transfer functions are strictly inside the unit circle.

Models of Definition 1 are quite general and need to be restricted for the results to hold. Observe that not every second-order process is exponentially forgetting, even if the associated linear innovation process is independent. The next proposition clarifies this point.

Proposition 17 *Assume that $\boldsymbol{y}$ is a second-order discrete-time stochastic process with independent linear innovations $\{\boldsymbol{\varepsilon}_t\}$ and no linearly deterministic part; then $\boldsymbol{y}$ is not necessarily exponentially forgetting.*
On the other hand, if $\sup_t \mathbb{E}\left[\|\boldsymbol{\varepsilon}_t\|^4\right] < \infty$ and the sequences $\{h_k(t) : k \in \mathbb{N}_0,\ t \in \mathbb{Z}\}$ in Wold's decomposition (A.1) are uniformly exponentially decaying, i.e., there exist positive constants $c < \infty$ and $0 < \lambda < 1$ such that $|h_k(t)| < c\lambda^k$ for every $k \in \mathbb{N}_0$ and $t \in \mathbb{Z}$; then $\boldsymbol{y}$ is an exponentially forgetting process [4] of order 4.

[3] Note that for the OE-WQPEM problem $\boldsymbol{e}_t(\theta) = \boldsymbol{y}_t - \hat{y}_t(\theta)$, where $\hat{y}_t(\theta)$ is the OE-predictor, only when M in Definition 15 is block diagonal. Otherwise, $\boldsymbol{e}_t(\theta) = [L^{-\top}(\boldsymbol{Y} - \hat{Y}(\theta))]_t$ where L is given by the LDL^T factorization of M^{-1}.

PROOF. The first assertion is straightforward; we only need to find an example of a second-order discrete-time stochastic process whose innovations are independent but which is not exponentially stable. Consider for example the process $\boldsymbol{y}_t := \sum_{k=1}^{\infty} k^{-1}\boldsymbol{\varepsilon}_{t-k}$ where $\{\boldsymbol{\varepsilon}_k\}$ are independent innovations. This is clearly a second-order process that also forgets the remote past, however only linearly. To prove the second part, we use Wold's decomposition of $\boldsymbol{y}$ assuming zero mean, namely $\boldsymbol{y}_t = \sum_{k=0}^{\infty} h_k(t)\boldsymbol{\varepsilon}_{t-k}$, with the hypothesis that the sequence $\{h_k(t) : k \in \mathbb{N}_0, t \in \mathbb{Z}\}$ is uniformly exponentially decaying. Using the triangular inequality, it holds that for every $t \in \mathbb{Z}$ and every $n \in \mathbb{N}$

$$\left\|\sum_{k=1}^{n} h_k(t)\boldsymbol{\varepsilon}_{t-k}\right\|^4 \leq \left(\sum_{k=1}^{n} \|h_k(t)\|\|\boldsymbol{\varepsilon}_{t-k}\|\right)^4$$
$$\leq c^4\left(\sum_{k=1}^{n}\lambda^k\|\boldsymbol{\varepsilon}_{t-k}\|\right)^4 \leq c^4\left(\sum_{k=1}^{n}\lambda^k\right)^3\sum_{k=1}^{n}\lambda^k\|\boldsymbol{\varepsilon}_{t-k}\|^4$$

in which we used Hölder's inequality ([3, page 85] applied to $(\lambda^{\frac{k}{p}})(\lambda^{\frac{k}{q}}\|\boldsymbol{\varepsilon}_{t-k}\|)$ with $p = \frac{4}{3}$, $q = 4$) for the last implication. By applying the expectation operator to both sides and letting $N \to \infty$ we get the inequality

$$\mathbb{E}\left[\|\boldsymbol{y}_t\|^4\right] \leq \frac{c^4}{(1-\lambda)^3}\sum_{k=1}^{\infty}\lambda^k\mathbb{E}\left[\|\boldsymbol{\varepsilon}_{t-k}\|^4\right]. \tag{35}$$

Finally, by defining $\boldsymbol{y}_{t,s} := \sum_{k=0}^{\infty} h_k(t)\boldsymbol{\varepsilon}_{t-k,s}$ such that $\boldsymbol{\varepsilon}_{t,s} = \boldsymbol{\varepsilon}_t$ for $t > s$ and zero otherwise, (35) and the assumption $\sup_t \mathbb{E}\left[\|\boldsymbol{\varepsilon}_t\|^4\right] < \infty$ imply that

$$\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_{t,s}\|^4\right] \leq \frac{c^4}{(1-\lambda)^3}\sum_{k=t-s}^{\infty}\lambda^k\mathbb{E}\left[\|\boldsymbol{\varepsilon}_{t-k}\|^4\right]$$
$$\leq \tilde{c}\lambda^{t-s}, \quad \forall t > s$$

which proves the statement.

More explicit conditions can be given for specific model sets. The next example considers the class of stochastic Wiener models.

Example 18 (Exponentially stable data) *Suppose the system is described by the stochastic Wiener model*

$$\begin{aligned}\boldsymbol{x}_t &= G(\mathrm{q};\theta^\circ)u_t + H(\mathrm{q};\theta^\circ)\boldsymbol{w}_t,\\ \boldsymbol{y}_t &= f(\boldsymbol{x}_t;\theta^\circ) + \boldsymbol{v}_t, \qquad t \in \mathbb{Z},\end{aligned} \tag{36}$$

and suppose that the LTI part of the system is rational and stable; i.e., the poles of $G(z;\theta^\circ)$ and $H(z;\theta^\circ)$ are strictly inside the unit circle. Furthermore, suppose that $\boldsymbol{w}$ and $\boldsymbol{v}$ are independent and mutually independent white noises with bounded moments of all order.

[4] $r = 4$ is sufficient for the analysis of PEMs, see Lemma 21.

Then, in the light of Proposition 17, we see that $\boldsymbol{x}$ is an exponentially forgetting process. Because the nonlinearity f is static, we only need to guarantee that moments of $\boldsymbol{y}$ are bounded and that $f(\boldsymbol{x};\theta^\circ)$ is exponentially decaying whenever $\boldsymbol{x}$ is exponentially decaying. This is the case when f is a polynomial, or bounded, in $\boldsymbol{x}$, for example.

5.1.1 Conditions on the predictor

Conditions on the predictors are mainly used to guarantee that the PE process is exponentially forgetting uniformly in θ. Apart from a differentiability condition with respect to the parameter and a compactness condition on the parameter set, it is required that the remote past observation has little effect on the current output of the predictor and its derivative. From the point view of asymptotic analysis, this means that all the observed outputs, regardless of their order in time, may have a comparable contribution on the choice of the parameter. From the practical point of view, this is required for the numerical stability of the minimization procedure. This reasonable condition means that the used predictors should have a stability property.

Definition 19 (Uniformly stable predictors) *The predictors $\{\hat{\boldsymbol{y}}_{t|t-1}(\theta) = \psi(\boldsymbol{D}_{t-1}, t;\theta),\ \theta \in \Theta\}$, where Θ is compact, are said to be uniformly stable if there exist positive real numbers $c < \infty$ and $\lambda \in (0,1)$ such that the following conditions hold:*

(1) $\theta \mapsto \psi(D_{t-1}, t;\theta)$ is continuously differentiable over an open neighborhood of Θ $\forall t$ and for every data set D_{t-1}.

(2) $\|\xi(0,t;\theta)\| \leq c \quad \forall t,\ \forall\theta$ in an open neighborhood of Θ, where ξ is used to denote both the predictor function ψ and its derivative with respect to θ, and 0 represents a data set of arbitrary inputs and zero outputs of length $t-1$.

(3) $\|\xi(D_{t-1},t;\theta) - \xi(\bar{D}_{t-1},t;\theta)\| \leq c\sum_{k=0}^{t-1}\lambda^{t-k}\|y_k - \bar{y}_k\|$, where θ is in an open neighborhood of Θ, and D_{t-1}, $\bar{D}_{t-1}$ are data sets corresponding to arbitrary realizations, y and $\bar{y}$, of the output, and a fixed arbitrary input u.

First, observe that the OE-predictor is deterministic and depends only on u; therefore, it always satisfies the third condition of the above definition. Moreover, note that the compactness of Θ is part of the definition.

For the OE-predictor and the OL-predictors to be uniformly stable, it is clear that the parameterization of $\mu(U;\theta)$ and $\Sigma(U;\theta)$ is required to be continuously differentiable over Θ; this translates into smoothness conditions on the parameterization of the assumed nonlinear model. To check the remaining conditions, we first recall that the predictors have the form (see (10) and (16))

$$\psi(\boldsymbol{D}_{t-1},t;\theta) = \mathbb{E}[\boldsymbol{y}_t;\theta] + \sum_{k=1}^{t-1}\tilde{l}_{t-k}(t,U_{t-1};\theta)\left(\boldsymbol{y}_k - \mathbb{E}[\boldsymbol{y}_k;\theta]\right),$$

in which $\tilde{l}_j(t, U_{t-1}; \theta) := \left[L^{-1}(U;\theta)\right]_{tj}$, $1 < t \leq N$ for the OL-predictor, and $\tilde{l}_j(t, U_{t-1}; \theta) := 0 \ \forall t, j$ for the OE-predictor. In either case, we observe that the second condition of Definition 19 requires the function $(t, \theta) \mapsto \mathbb{E}[\boldsymbol{y}_t; \theta]$ and its derivative with respect to θ to be uniformly bounded in t and θ.

The third condition of Definition 19 only concerns the OL-predictor. To understand the condition, we invite the reader to compare the OL-predictor (10) to the optimal linear predictor in the LTI case (28). There, the predictors satisfy the required stability property by imposing the assumption that the transfer function $H^{-1}(z;\theta)G(z;\theta)$ is rational and stable for all $\theta \in \Theta$, in addition to the assumption that the noise model $H(z;\theta)$ is rational and causally inversely stable over Θ, see [40, Lemma 4.1 and Lemma 4.2 on pages 109 and 110]. As with the LTI case, we shall here impose the assumption of causal and (exponentially) stable invertibility of $\boldsymbol{y}$ with respect to the linear innovations for all $\theta \in \Theta$. We will assume that, for every t, it is possible to write

$$\boldsymbol{\varepsilon}_t(\theta) = \sum_{k=0}^{\infty} \tilde{l}_k(t, U_{t-1}; \theta)(y_{t-k} - \mathbb{E}[\boldsymbol{y}_{t-k}; \theta]),$$

where the sequence $\{\tilde{l}_k(t, U_{t-1}; \theta) : k \in \mathbb{N}_0, \, t \in \mathbb{N}\}$ and its derivative in θ are uniformly exponentially decaying [5] with $\tilde{l}_0(t;\theta) = I \ \forall t$. Note that in the LTI case this is equivalent to the assumptions on the noise model $H(z;\theta)$; but here it involves the used input. Under these assumptions, the OL-predictor is uniformly stable once the smoothness conditions on the parameterization and the regularity conditions on $\mathbb{E}[\boldsymbol{y}_t; \theta]$ are satisfied.

5.2 Conditions on the identification criterion

The simplest and most commonly used choice for the criterion function is the squared Euclidean norm, i.e., $\ell(e, t; \theta) := \|e\|^2$. In this case, the convergence of the PEM estimators can be established with no further conditions. For the general case, where the criterion function is time- and/or θ-dependent, it is sufficient to require that the functions are quadratically bounded according to the following definition.

Definition 20 (Quadratically bounded criteria) *The family of prediction error criterion functions $\{\boldsymbol{\mathcal{V}}_N(\theta) := \frac{1}{N}\sum_{k=1}^{N} \ell(e_t(\theta), t; \theta) : N \in \mathbb{N}, \ \theta \in \Theta\}$ is quadratically bounded if $\{\ell(\cdot, t; \cdot)\}$ are continuously differentiable for every t, and for some $c_1, c_2 < \infty$ and every $\boldsymbol{e}$ the following conditions hold* [6]

$$\begin{aligned}
&(1)\ \|\ell(0, t; \theta)\| \leq c_1, && \forall \theta \in \Theta, \ \forall t \in \mathbb{N}, \\
&(2)\ \|\tfrac{\partial}{\partial e}\ell(e, t; \theta)\| \leq c_1\|e\|, && \forall \theta \in \Theta, \ \forall t \in \mathbb{N}, \\
&(3)\ \|\tfrac{\partial}{\partial \theta}\ell(e, t; \theta)\| \leq c_1\|e\|^2 + c_2, && \forall \theta \in \Theta, \ \forall t \in \mathbb{N}.
\end{aligned}$$

[5] i.e., $|\tilde{l}_{t-k}(t, U_{t-1}; \theta)| < c\lambda^k$ for some $c < \infty$, $|\lambda| < 1$, every t and θ, and similarily for the derivatives.

It is clear that the squared Euclidean norm (used to define the OL-QPEM and the OE-QPEM estimators) and the weighted norm (used to define the OE-WQPEM estimator) are θ-independent and quadratically bounded. For the case of OL-GPEM, the criterion is defined as

$$\ell(e, t; \theta) = e^\top \Lambda_t^{-1}(\theta)e + \log\det\Lambda_t(\theta).$$

Because this function is quadratic in e, it is only required to verify that

$$\frac{\partial}{\partial\theta}\ell(e, t; \theta) = -e^\top \Lambda_t^{-1}(\theta)\frac{\partial \Lambda_t(\theta)}{\partial\theta}\Lambda_t^{-1}(\theta)e + \mathbf{tr}\left(\Lambda_t^{-1}(\theta)\frac{\partial\Lambda_t(\theta)}{\partial\theta}\right)$$

is well-defined and quadratically bounded. This is a requirement on the parameterization of the covariance matrices $\Lambda_t(\theta)$ of the innovations. Observe that these matrices are defined via an $LDL^\top$ decomposition which is a continuous operation; see [21]. When the parameterization is continuously differentiable such that the covariance matrices are uniformly bounded for all t and θ, the condition is satisfied. Therefore, once more, we end up with conditions on the parameterization of the model.

We are now ready to state the basic convergence result.

Lemma 21 (Convergence of PEM estimators) *Suppose that the nonlinear system generating the data is r-stable with $r = 4$, the used predictor is uniformly stable according to Definition 19, and the identification criterion is quadratically bounded. Then, the sequence $\{\mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)]\}_{N\in\mathbb{N}}$ is equicontinuous on Θ and as $N \to \infty$,*

$$\sup_{\theta\in\Theta} |\boldsymbol{\mathcal{V}}_N(\theta) - \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)]| \xrightarrow{a.s.} 0.$$

Furthermore,

$$\hat{\boldsymbol{\theta}}_N \xrightarrow{a.s.} \mathcal{D}_I := \left\{\theta \in \Theta : \liminf_{N\to\infty} \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)] \leq \min_{\beta\in\Theta} \limsup_{N\to\infty} \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\beta)]\right\},$$

where $\hat{\boldsymbol{\theta}}_N$ denotes the OL-QPEM, the OL-GPEM, the OE-QPEM, or the OE-WQPEM estimator and $\boldsymbol{\mathcal{V}}(\theta)$ denotes the corresponding identification criterion.

PROOF. The proof is due to Ljung in [39]. Observe that the proof there remains valid under the conditions in Definition 20 (compare to condition C1 there). In particular, (A.6) in [39, page 781] still holds true.

This result means that the criterion function becomes arbitrary close to the average criterion function such that, almost surely, for every $\epsilon > 0$ there exist $n \in \mathbb{N}$ such that for all $N > n$, the set $\mathcal{D}_I \cap \{\theta : \|\hat{\theta}_N - \theta\| < \epsilon\} \neq \emptyset$.

[6] Notice that the conditions are slightly different compared to condition C1 in [39, page 775]; here, the conditions are modified to cover the OL-GPEM criterion function which is parameterized by θ (the model parameters).

Observe that the result is given for a fairly general case that includes scenarios where the true system is not in the assumed model set (i.e., there is no true parameter θ°, or $\theta^\circ \notin \Theta$). For all $\theta^* \in \mathcal{D}_I$ it holds that

$$\liminf_{N\to\infty} \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta^*)] \leq \limsup_{N\to\infty} \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)] \quad \forall \theta \in \Theta, \quad (37)$$

and therefore $\mathcal{D}_I$ can be interpreted as the set of parameters that give "the best average prediction" according to the chosen predictor and criterion function. Note that Lemma 21 established the convergence of the process $\{\hat{\boldsymbol{\theta}}_N\}_{N\in\mathbb{N}}$ only to a subset of Θ. However, for cases when $\theta^\circ \in \Theta$ (see Assumption 2), and assuming that an identifiability condition holds such that the limit set is a singleton, $\mathcal{D}_I = \{\theta^\circ\}$, a consistency proof is completed by an application of the lemma. Formally, we will use the following identifiability conditions.

Definition 22 (Identifiable parameterization) *For a model (1) and an input u, we say that Θ constitutes*

- *a* first-order *identifiable parameterization if there exists $\tilde{N} \in \mathbb{N}$, such that $\forall \theta, \tilde{\theta} \in \Theta$, and $N > \tilde{N}$, it holds that*
$$\mu(U;\theta) = \mu(U;\tilde{\theta}) \Leftrightarrow \theta = \tilde{\theta}. \quad (38)$$
- *a* second-order *identifiable parameterization if there exists $\tilde{N} \in \mathbb{N}$, such that $\forall \theta, \tilde{\theta} \in \Theta$, and $N > \tilde{N}$, it holds that*
$$\mu(U;\theta) = \mu(U;\tilde{\theta}),\ \Sigma(U;\theta) = \Sigma(U;\tilde{\theta}) \ \Leftrightarrow \theta = \tilde{\theta}. \quad (39)$$

Note that the definition involves the used input. In the linear case, the first-order condition (38) is analogous to identifiability conditions for OE models, and the second-order condition (39) is analogous to identifiability conditions when the plant model G and the noise model H share parameters (see [40, Chapter 8]).

Next, we state our main consistency theorems.

Theorem 23 (Consistency of the OL-QPEM and the OL-GPEM estimators) *Suppose that the true system, given by Assumption 2, is r-stable with $r = 4$. Assume that the input u is such that Θ constitute a second-order identifiable parameterization, and that the OL-predictor is uniformly stable. Then, the OL-QPEM estimator (27) is strongly consistent; i.e. $\hat{\boldsymbol{\theta}}_N \xrightarrow{a.s.} \theta^\circ$ as $N \to \infty$.*

Moreover if, in addition, the parameterization is such that the linear innovations covariances are continuously differentiable and uniformly bounded in t and θ, then the OL-GPEM estimator (29) is strongly consistent.

PROOF. First observe that the OL-QPEM estimator and the OL-GPEM estimator satisfy (see Section 4)

$$\theta^\circ = \arg\min_{\theta\in\Theta} \ \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)], \quad \forall N \in \mathbb{N}, \quad (40)$$

and due to the identifiability assumption, θ° is a unique minimizer when N is sufficiently large. Moreover, for both estimators, due to the form of $\boldsymbol{\mathcal{V}}_N(\theta)$, the compactness of Θ and the continuity of $\mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)]$ over Θ, it holds that for every $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that

$$\min_{\{\theta\in\Theta:\|\theta-\theta^\circ\|\geq\epsilon\}} \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)] - \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta^\circ)] > \delta_\epsilon \quad (41)$$

for all sufficiently large N. Consequently,

$$\begin{aligned}
&\min_{\{\theta\in\Theta:\|\theta-\theta^\circ\|\geq\epsilon\}} \boldsymbol{\mathcal{V}}_N(\theta) - \boldsymbol{\mathcal{V}}_N(\theta^\circ) \\
&\quad = \min_{\{\theta\in\Theta:\|\theta-\theta^\circ\|\geq\epsilon\}} [\boldsymbol{\mathcal{V}}_N(\theta) - \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)]] \\
&\qquad + \min_{\{\theta\in\Theta:\|\theta-\theta^\circ\|\geq\epsilon\}} [\mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)] - \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta^\circ)]] \\
&\qquad + [\mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta^\circ)] - \boldsymbol{\mathcal{V}}_N(\theta^\circ)] \\
&\quad \geq \min_{\{\theta\in\Theta:\|\theta-\theta^\circ\|\geq\epsilon\}} [\mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)] - \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta^\circ)]] \\
&\qquad - 2\sup_{\theta\in\Theta} |\boldsymbol{\mathcal{V}}_N(\theta) - \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)]| \\
&\quad > 0 \quad \text{a.s. for sufficiently large } N,
\end{aligned} \quad (42)$$

where we used (41) and Lemma 21 to deduce the last inequality. However, by the definition of $\hat{\boldsymbol{\theta}}_N$ and the conditions on ℓ, it holds that

$$\boldsymbol{\mathcal{V}}_N(\hat{\boldsymbol{\theta}}_N) - \boldsymbol{\mathcal{V}}_N(\theta) \leq 0 \quad \forall \theta \in \Theta \ \text{ and every } N,$$

particularly for $\theta = \theta^\circ$. Then, in the view of (42), it must hold that $\|\hat{\boldsymbol{\theta}}_N - \theta^\circ\| < \epsilon$ a.s. for sufficiently large N. But since ϵ is arbitrary, this means that $\hat{\boldsymbol{\theta}}_N \xrightarrow{\text{a.s.}} \theta^\circ$ as $N \to \infty$, and thus $\mathcal{D}_I = \{\theta^\circ\}$ and the estimators are strongly consistent.

In cases where the stronger condition of first-order identifiability holds, consistency of the OE-QPEM and the OE-WQPEM estimators can be established.

Theorem 24 (Consistency of the OE-QPEM and the OE-WQPEM estimators) *Suppose that the true system, given by Assumption 2, is r-stable with $r = 4$. Assume that the input u is such that Θ constitutes a first-order identifiable parameterization, and that the OE-predictor satisfies the first two conditions of Definition 19. Then, the OE-QPEM estimator (30) and the OE-WQPEM estimator (31) are strongly consistent; i.e. $\hat{\boldsymbol{\theta}}_N \xrightarrow{a.s.} \theta^\circ$ as $N \to \infty$.*

PROOF. The proof follows the proof of Theorem 23.

Remark 25 *An equivalent proof for the consistency may be obtained by reducing (37), the defining inequality of $\mathcal{D}_I$, into an equality; e.g., in the OL-QPEM case to the equation $\liminf_{N\to\infty} \frac{1}{N}\mathbb{E}\|\widehat{\boldsymbol{Y}}(\theta^\circ) - \widehat{\boldsymbol{Y}}(\theta)\|^2 = 0$, $\theta \in \mathcal{D}_I$, and then showing, using the identifiability conditions, that it can only be satisfied at θ° (see [37] or [39]).*

6 Asymptotic normality

Subject to a strengthening of the hypotheses of the consistency theorems in Section 5, it is possible to establish that the OL-QPEM, OL-GPEM, OE-QPEM and OE-WQPEM estimators are asymptotically normally distributed around θ°. The additional conditions are used to ensure that $\boldsymbol{\mathcal{V}}''_N(\theta)$ asymptotically behaves as $\mathbb{E}[\boldsymbol{\mathcal{V}}''_N(\theta)]$ uniformly over Θ, and that the derivative $\boldsymbol{\mathcal{V}}'_N(\theta^\circ)$ is asymptotically normal when multiplied by $\sqrt{N}$ and normalized [43].

6.1 Conditions for asymptotic normality

C1. The underlying system is r-stable with $r > 4$.
C2. The predictors are three times continuously differentiable with respect to θ such that the derivatives satisfy the second and third conditions in Definition 19.
C3. The criterion functions are three times continuously differentiable with respect to θ and twice continuously differentiable with respect to e such that for all $\theta \in \Theta$ and $t \in \mathbb{N}$ there exist $c_1, c_2 < \infty$ such that [7]

(1) $\|\frac{\partial^k}{\partial\theta^k}\frac{\partial^2}{\partial e^2}\ell(e,t;\theta)\| \le c_1, \qquad k = 0, 1,$
(2) $\|\frac{\partial^k}{\partial\theta^k}\frac{\partial}{\partial e}\ell(e,t;\theta)\| \le c_1\|e\|, \qquad k = 0, 1, 2,$
(3) $\|\frac{\partial^k}{\partial\theta^k}\ell(e,t;\theta)\| \le c_1\|e\|^2 + c_2, \qquad k = 1, 2, 3.$

Apart from an increased smoothness requirement on the parameterization of the predictor and the criterion function, the additional set of conditions C1-C3 requires that the second and third derivatives of the predictor have the uniform stability property.

We note here that the criterion functions of the PEM instances defined in the previous section are all quadratic in $\boldsymbol{e}$. In the case of OL-GPEM, the criterion is parameterized by θ, and the covariances $\Lambda_t(\theta)$ have to be three times continuously differentiable and uniformly bounded according to the above conditions. This translates into a smoothness requirement on the parameterization of the covariance of the model.

Theorem 26 (Asymptotic normality) *Assume that, in addition to the hypotheses of Theorem 23 and Theorem 24, the set of strengthened conditions C1-C3 holds. Furthermore, let $\mathcal{W}_N(\theta) := \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)]$ and assume that for some $\delta > 0$ and some $N_0 \in \mathbb{N}$,*

$$\mathcal{W}''_N(\theta) \succ \delta I, \quad \forall\theta \in \Theta,\ \forall N > N_0. \tag{43}$$

Introduce the matrices $P_N := [\mathcal{W}''_N(\theta^\circ)]^{-1}\mathcal{Q}_N[\mathcal{W}''_N(\theta^\circ)]^{-1}$, where $\mathcal{Q}_N := \mathbb{E}[N\ \boldsymbol{\mathcal{V}}'_N(\theta^\circ)(\boldsymbol{\mathcal{V}}'_N(\theta^\circ))^\top]$. Assume that $P_N \succ \delta I$ and $\mathcal{Q}_N \succ \delta I$ for some $\delta > 0$ and all sufficiently large N. Then

$$\sqrt{N}P_N^{-\frac{1}{2}}(\hat{\boldsymbol{\theta}}_N - \theta^\circ) \rightsquigarrow \mathcal{N}(0, I_d) \quad as \quad N \to \infty, \tag{44}$$

where the symbol $\rightsquigarrow$ denotes convergence in distribution, I_d denotes the identity matrix of size d, and where $\hat{\boldsymbol{\theta}}_N$ denotes the OL-QPEM, the OL-GPEM, the OE-QPEM or the OE-WQPEM estimator and $\boldsymbol{\mathcal{V}}(\theta)$ denotes the corresponding identification criterion.

PROOF. The proof is due to Ljung and Caines [43]. As with Lemma 21, the proof of the asymptotic normality remains valid under the modified condition C3. In particular, Lemma 3 in [43] still hold true.

In (quasi-)stationary ergodic scenarios where the average criterion $\mathcal{W}_N \to \bar{\mathcal{W}}(\theta)$ as $N \to \infty$ and the matrices $\mathcal{Q}_N \to \bar{\mathcal{Q}}$ as $N \to \infty$ such that the limit is invertible, it is not difficult to show that $\sqrt{N}(\hat{\theta}_N - \theta^\circ) \rightsquigarrow \mathcal{N}(0, P)$ as $N \to \infty$, where $P = [\bar{\mathcal{W}}''(\theta^\circ)]^{-1}\bar{\mathcal{Q}}[\bar{\mathcal{W}}''(\theta^\circ)]^{-1}$ is the asymptotic covariance matrix of the estimator. In such cases, it is possible to derive an expression for P that can be used for the design of an optimal criterion for a given predictor function; i.e., a criterion that leads to a minimal P with respect to the usual partial ordering of positive semidefinite matrices (see [73] for the LTI case). With no (quasi-)stationarity assumptions, an optimality analysis, in the same spirit, can be done by studying the normalizing sequence of matrices $\{P_N : N \in \mathbb{N}\}$. The scalar function ℓ to be preferred is the one corresponding to a minimal normalizing sequence; in other words, the one leading to the largest normalization factors $P_N^{-1/2}$ such that the convergence in (44) still holds. However, computing the expressions of P_N requires the knowledge of up to the fourth order moments of the linear innovation process for $t = 1, \ldots, N$.

7 A Maximum Likelihood Interpretation

It is also possible to arrive at the estimators presented in Section 4 through the use of misspecified models. Let us, incorrectly, assume that

$$\boldsymbol{Y} = \mu(U;\theta) + \boldsymbol{Z}$$

where $\boldsymbol{Z} \sim \mathcal{N}(0, I_N)$. Then the likelihood function is given by

$$\tilde{p}(\boldsymbol{Y};\theta) = \prod_{t=1}^{N} \frac{1}{(2\pi)^{\frac{d_y}{2}}} \exp\left(-\frac{1}{2}\|\boldsymbol{y}_t - \mathbb{E}[\boldsymbol{y}_t;\theta]\|_2^2\right).$$

and we see that the OE-QPEM estimator maximizes this function. Likewise, the more refined model

$$\boldsymbol{Y} = \mu(U;\theta) + L(U;\theta)\Lambda^{\frac{1}{2}}(U;\theta)\boldsymbol{Z},$$

i.e., a Gaussian model with the same first and second moments as the true model, results in that the MLE is given by a OL-GPEM.

From this exercise we obtain some insight into why our estimators are consistent: the distribution of the error

[7] Note that these conditions are modified versions of those in [43, page 33], to allow for the OL-GPEM criterion function.

term is misspecified but, as in the linear case (e.g., for OE models; see Section 3.2), this is not critical for consistency, but only hampers asymptotic efficiency.

Remark 27 *The idea of using a misspecified likelihood function to construct tractable estimators is not new. It can be traced back to [6, Section 3.3] under the name of pseudo-likelihood methods where it was used for data with spatial dependence, when the likelihood function is unavailable. It has also been suggested and studied in Econometrics; for example, the asymptotic properties were investigated in [23] for conditionally independent models.*

8 Numerical Examples

We now demonstrate the performance of the methods proposed in Section 4 on a couple of examples in which the conditions of Theorems 23, 24 and 26 are satisfied. The estimated models are single-input single-output models; but, it should be clear that the methods cover the multiple-input multiple-output case as well.

8.1 First order bi-modal nonlinear state-space model

In this example, we demonstrate the asymptotic properties of the PEM estimators defined in Section 4 and compare their performance to the current state-of-the-art ML method [34]. Suppose that the true model is described by the relations

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \theta^\circ \boldsymbol{x}_t + u_t + \boldsymbol{w}_t, \quad & x_1 &= 0, \\ \boldsymbol{y}_t &= \boldsymbol{x}_t^2 + \boldsymbol{v}_t, & t &\in \mathbb{N}, \end{aligned} \tag{45}$$

where $\theta^\circ = 0.7$, $\boldsymbol{w}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \lambda_w)$, $\boldsymbol{v}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \lambda_v)$ for all t, in which $\lambda_w = 1$, and $\lambda_v = 0.1$ such that $\boldsymbol{w}$ and $\boldsymbol{v}$ are mutually independent. Let the input be a known realization of a standard Gaussian process and assume that λ_w and λ_v are known. To obtain an identifiable parameterization, we let $\Theta := [\epsilon, 1-\epsilon]$ for a small positive ϵ.

The model in (45) is a stochastic Wiener model with a single output; the static nonlinearity is a second order monomial which makes the problem challenging. In particular, note that the posterior distribution of the state is bi-modal, and the optimal MSE predictor as well as the likelihood function are analytically intractable. Nevertheless, as we now show, it is possible to compute the first two moments of $\boldsymbol{y}$ analytically; thus, the OL-predictor and the OE-predictor are given in terms of closed-form expressions. Let us use the vector notation

$$\boldsymbol{Y} = \boldsymbol{X}^2 + \boldsymbol{V} = (F(\theta)\boldsymbol{W} + F(\theta)U)^2 + \boldsymbol{V},$$

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_0 \\ \vdots \\ \boldsymbol{w}_{N-1} \end{bmatrix}, \boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_N \end{bmatrix}, F(\theta) = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \theta & 1 & \dots & 0 \\ \vdots & \ddots & & \\ \theta^{N-1} & \theta^{N-2} & \dots & 1 \end{bmatrix}$$

and recall that we apply the exponent entry-wise. Then,

$$\begin{aligned} \mu(U;\theta) &= \mathbb{E}[\boldsymbol{X}^2;\theta] = \mathbb{E}[(F(\theta)\boldsymbol{W} + F(\theta)U)^2;\theta] \\ &= \mathbb{E}[(F(\theta)\boldsymbol{W})^2 + (F(\theta)U)^2 + 2(F(\theta)\boldsymbol{W}) \circ (F(\theta)U);\theta] \\ &= \lambda_w F(\theta^2)\mathbf{1} + (F(\theta)U)^2, \end{aligned}$$

where the symbol $\circ$ denotes the Hadamard (entry-wise) product, and for the last equality we used $\mathbb{E}[(F(\theta)\boldsymbol{W})^2;\theta] = \lambda_w F(\theta^2)\mathbf{1}$. Moreover,

$$\Sigma(U;\theta) = \mathbf{cov}(\boldsymbol{X}^2, \boldsymbol{X}^2;\theta) + \lambda_v I_N$$

because $\boldsymbol{w}$ and $\boldsymbol{v}$ are mutually independent. Due to the assumption that $\boldsymbol{w}$ is Gaussian, the ij^{th}-entry of $\mathbf{cov}(\boldsymbol{X}^2, \boldsymbol{X}^2)$ is given by

$$\begin{aligned} \lambda_w^2 [M(\theta)]_{ij} + 4\lambda_w \left[(F(\theta) \circ F(\theta)U)(F(\theta) \circ F(\theta)U)^\top\right]_{ij} \\ - \lambda_w^2 \left[(F(\theta^2)\mathbf{1})(F(\theta^2)\mathbf{1})^\top\right]_{ij} \end{aligned}$$

for all $i, j \in \{1, \dots, N\}$, where $M(\theta)$ is a symmetric matrix whose entries are given by

$$[M(\theta)]_{ij} := \begin{cases} 3\left(\dfrac{\theta^{2t}-1}{\theta^2-1}\right)^2, \text{if } i = j = t \\ \dfrac{(1-\theta^{-2i})(3\theta^{2(j+i)} - 2\theta^{2j} - \theta^{2i})}{(\theta^2-1)^2}, \text{if } j > i. \end{cases}$$

We conducted a Monte Carlo simulation using 1000 data sets, corresponding to independent realizations of $\boldsymbol{w}$, $\boldsymbol{v}$ and the input, for values of N between 100 and 2000. For each N and each data set, we computed the OE-QPEM, OL-GPEM, OL-QPEM estimators in addition to the following two versions of the OE-WQPEM estimator:

$$\text{OE-WQPEM: } \arg\min_{\theta\in\Theta} \|\boldsymbol{Y} - \mu(U;\theta)\|^2_{\Sigma^{-1}(U;\hat{\theta}_{\text{OE}})}$$

$$\text{OE-WQPEM(true weight): } \arg\min_{\theta\in\Theta} \|\boldsymbol{Y} - \mu(U;\theta)\|^2_{\Sigma^{-1}(U;\theta^\circ)}$$

where $\hat{\theta}_{\text{OE}}$ denotes the OE-QPEM estimate computed using the same data set. This is to examine the role of weighting on the accuracy. For comparison, we also computed the MLE approximated using a stochastic approximation Expectation-Maximization algorithm based on a Conditional Particle Filter with ancestor sampling (CPF-SAEM) [34]. For the SAEM algorithm, we used 20 particles and 2000 iterations; the step size for the stochastic approximation step is $\gamma_i = 0.98\ \forall i \leq 100$ and $\gamma_i \sim i^{-0.7}$ for $100 < i \leq 2000$. To improve the convergence rate, we applied Polyak averaging (see [55]) over the last 500 iterations. The OL-QPEM and OL-GPEM problems were solved using a quasi-Newton algorithm (`fminunc` in Matlab 2017a). The OE-QPEM and OE-WQPEM problems were solved using the Levenberg-Marquardt algorithm (`lsqnonlin` in Matlab 2017a). To ensure stability, the one-to-one transformation $\theta = 1/(1+\exp(-\tau))$ was used during optimization. In all cases, the problems were initialized at θ°.

The simulation results are reported in Figure 1; they clearly indicate the consistency of all the six estimators. The accuracy of OL-QPEM and OL-GPEM lie between that of OE-QPEM and the MLE. The MSE of OE-WQPEM almost coincides with OE-WQPEM(true weight) and comes very close to that of the asymptotically efficient MLE; the difference at $N = 500$ is 1.3×10^{-4} and at $N = 1000$ is only 8.8×10^{-5}. However, the average time for computing OE-WQPEM is 1.8 seconds, compared to 135 seconds for the MLE (OE-WQPEM is about 70 times faster here) [8]. It is also of interest to observe that the accuracy of OE-WQPEM is better than both OL-QPEM and OL-GPEM. A detailed explanation of this behaviour is considered in a future contribution by the authors.

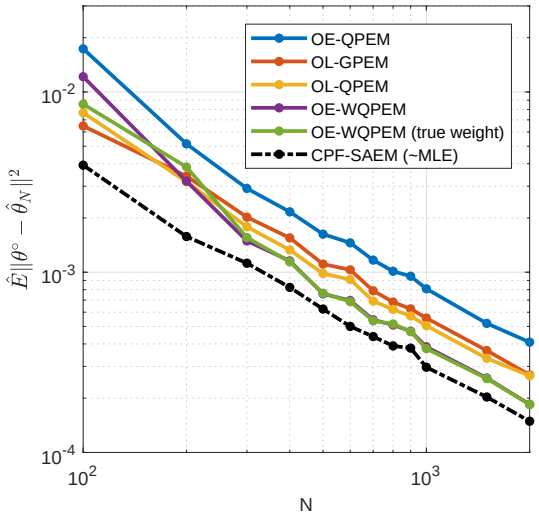

Fig. 1. The MSE of six different estimators approximated using 1000 Monte Carlo realizations.

In the following example, we consider a recently proposed stochastic Wiener-Hammerstein benchmark problem [59]; the results of this example have appeared in [2].

8.2 A Wiener-Hammerstein benchmark problem

Consider the Wiener-Hammerstein benchmark problem defined in [59] where the data was generated using an electronic circuit representing a Wiener-Hammerstein system with a saturation nonlinearity, see Figure 2. The input and output were both measured with small additive noise that may be ignored. The task is to use the measured data to identify a model of the electronic circuit; this poses a challenging problem, particularly due to the presence of a large unobserved colored process disturbance. Several estimation data sets as well as two test data sets (measured when $\boldsymbol{w} = 0$) are provided [9]. Here, we used a multisine estimation data set [10] that contains 10 independent experiments each of which corresponds to two steady-state periods with $N = 8192$.

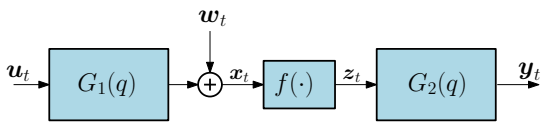

Fig. 2. A stochastic Wiener-Hammerstein model. The blocks G_1, G_2 are LTI models and f is a static nonlinearity.

We used the OE-QPEM estimator to estimate a model

$$\begin{aligned} \boldsymbol{y}_t &= G_2(\mathrm{q};\theta)\boldsymbol{z}_t, \quad t = 1, 2, 3 \dots \\ \boldsymbol{z}_t &= f(\boldsymbol{x}_t;\theta), \\ \boldsymbol{x}_t &= G_1(\mathrm{q};\theta)u_t + \boldsymbol{w}_t, \end{aligned}$$

in which q is the forward-shift operator, G_1 and G_2 are third order, causal, and stable transfer operators [11]. The saturation is modeled by a sigmoid function of the form $f(x;\theta) = \frac{L}{1+\exp(-kx)} - \frac{L}{2}$. It is assumed that the input u is known, the disturbance $\boldsymbol{w}$ is a stationary Gaussian process with zero mean and unknown variance λ_w.

To obtain an initial model, we used the first multisine experiment and an algorithm based on the best split of the BLA model, similar to that in [62]. For each possible split, a model was estimated by estimating the nonlinearity using the OE-QPEM estimator, and the split with the minimum OE-QPEM cost was selected [2]. The following initial values were used in the splitting algorithm: $L = 0.05, k = 1$, and $\lambda_w = 0.25$. The number of poles in G_1 and G_2 was fixed to 3, and only causal splits were allowed. The resulting initial model was then used to initialize the OE-QPEM problem.

To solve the OE-QPEM problem, the estimation data from the 10 multisine experiments were concatenated to obtain a long data set of length $N = 81920$ and the Levenberg-Marquardt algorithm (`lsqnonlin` in Matlab 2017a) was used. Figure 3 compares the simulated output of the final estimated model against the two provided test data sets (those are measured when $\boldsymbol{w} = 0$). The RMS of the residuals is reported in Table 1. The first column gives the RMSE of the BLA, and the second gives the RMSE of the initial model. The values in the third column are obtained when only the 7^{th} multisine experiment is used; this is the smallest RMS value over the 10 multisine experiments. The last column shows the RMSE of the final model, when all the data is used. The estimation time, on a personal laptop with 2.7 GHz Core i7 processor and 8 Gbyte RAM, is about 10 seconds for one experiment (N=8192), and about 30 seconds when all the data is used ($N = 81920$).

[8] These values are obtained using a personal laptop with 2.7 GHz Intel Core i7 processor and 8 Gbyte RAM operated by Windows 7 SP1. They are averages over 100 data sets when $N = 2000$ and θ° is used to initialize the algorithms.

[9] Available at htttp://www.nonlinearbenchmark.org/

[10] Stored in the file `WH_EstimationExample.mat`

[11] This is according to available prior knowledge [59].

Table 1
The residuals RMS of the estimated model (computed according to the instructions in [59])

	BLA	BLA+NL	OE-QPEM *(exp. 7)*	OE-QPEM *(all data)*
Swept-sine	0.0281	0.0127	0.0116	**0.0091**
Multisine	0.0339	0.0261	0.0171	**0.0148**

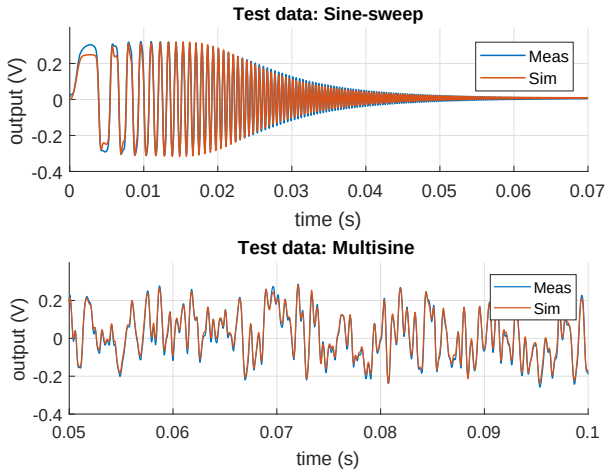

Fig. 3. The first part of the test data: the measured is in blue, and the simulated is in red. Observe how the simulated output follows closely the test data.

These results are quite encouraging. The accuracy of the model obtained using the proposed OE-QPEM estimator is comparable to those reported in [65, Table 1] and [19, Section 4.2] for models estimated using approximate MLEs. Observe that, due to the negligible measurement noise, the above benchmark problem is challenging for methods based on particle smoothing algorithms [65]. While the computational time of the OE-QPEM proposed here is in seconds, the computational time of the approximate ML method reported in [65] is a few hours.

9 Conclusions

In this contribution, we proposed the use of suboptimal one-step ahead predictors in PEMs. These predictors are linear in the observed outputs and may be computed analytically in several cases which are usually considered challenging. The resulting estimators are computationally attractive and their convergence is established under standard regularity conditions. Furthermore, consistency is achieved under certain identifiability conditions. The price paid for bypassing the likelihood function is a loss of statistical efficiency. However, the results of the numerical simulation example as well as the real-data benchmark problem show that the performance of the proposed methods can in some cases be comparable to the current state-of-the-art ML methods with a considerable reduction in the computational time.

References

[1] M. R. Abdalmoaty and H. Hjalmarsson. Simulated pseudo maximum likelihood identification of nonlinear models. *IFAC-PapersOnLine*, 50(1):14058 – 14063, 2017.

[2] M. R. Abdalmoaty and H. Hjalmarsson. Application of a linear PEM estimator to a stochastic Wiener-Hammerstein benchmark problem. *IFAC-PapersOnLine*, 51(15):784 – 789, 2018.

[3] R. B. Ash and C. A. Doléans-Dade. *Probability and Measure Theory*. Academic Press, 2000.

[4] K. J. Åström and T. Bohlin. Numerical identification of linear dynamic systems from normal operating records. In *Theory of Self-Adaptive Control Systems*, pages 96–111. Plenum Press, January 1966.

[5] K. J. Åström and B. Wittenmark. *Computer-Controlled Systems: Theory and Design*. Dover Publications, 2013.

[6] J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society.*, 24(3):179–195, 1975.

[7] S. A. Billings. Identification of nonlinear systems- a survey. *IEE Proceedings D - Control Theory and Applications*, 127(6):272–285, November 1980.

[8] S. A. Billings. *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*. Wiley, 2013.

[9] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer New York, 2009.

[10] P. E. Caines and J. Rissanen. Maximum likelihood estimation of parameters in multivariate Gaussian stochastic processes (corresp.). *IEEE Transactions on Information Theory*, 20(1):102–104, 1974.

[11] O. Cappé. Online sequential Monte Carlo EM algorithm. In *IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 37–40, 2009.

[12] H. Cramér. On some classes of nonstationary stochastic processes. In *Proceedings of the 4th Berkeley Symp. on Math. Statistics and Prob.*, volume 2, pages 57–78, 1961.

[13] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1):94–128, 1999.

[14] R. Douc, A. Garivier, E. Moulines, and J. Olsson. Sequential Monte Carlo smoothing for general state space hidden Markov models. *The Annals of Applied Probability*, 21(6):2109–2145, 2011.

[15] A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, July 2000.

[16] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

[17] M. Enqvist. *Linear Models of Nonlinear Systems*. Dissertation No. 985, Institutionen för systemteknik, Linköping University, Sweden, 2005.

[18] G. Fort and E. Moulines. Convergence of the Monte Carlo Expectation Maximization for curved exponential families. *The Annals of Statistics*, 31(4):1220–1259, 2003.

[19] G. Giordano and J. Sjöberg. Maximum likelihood identification of Wiener-Hammerstein system with process noise. *IFAC-PapersOnLine*, 51(15):401 – 406, 2018.

[20] F. Giri and EW. Bai. *Block-oriented Nonlinear System Identification*. Springer, 2010.

[21] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 2012.

[22] G. C. Goodwin and R. L. Payne. *Dynamic System Identification: Experiment Design and Data Analysis*. Academic Press, 1977.

[23] C. Gouriéroux, A. Monfort, and A. Trognon. Pseudo maximum likelihood methods: Theory. *Econometrica*, 52(3):681–700, 1984.

[24] R. Haber and H. D. Unbehauen. Structure identification of nonlinear dynamic systems - a survey on input/output approaches. *Automatica*, 26(4):651 – 677, 1990.

[25] A. Hagenblad and L. Ljung. Maximum likelihood estimation of Wiener models. In *Proceedings of the 39th IEEE Conference on Decision and Control*, volume 3, pages 2417–2418, 2000.

[26] A. Hagenblad, L. Ljung, and A. Wills. Maximum likelihood identification of Wiener models. *Automatica*, 44(11):2697 – 2705, 2008.

[27] E. J. Hannan and M. Deistler. *The statistical theory of linear systems*. Wiley, 1988.

[28] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Dover Publications, 2007.

[29] A. Juditsky, H. Hjalmarsson, A. Benveniste, B. Delyon, L. Ljung, J. Sjöberg, and Q. Zhang. Nonlinear black-box models in system identification: Mathematical foundations. *Automatica*, 31(12):1725 – 1750, 1995.

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

[31] I. Kollár. Frequency domain system identification toolbox for MATLAB. Budapest, 2004-2018. https://home.mit.bme.hu/∼kollar/fdident/.

[32] E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, 8:115–131, 2004.

[33] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 2011.

[34] F. Lindsten. An efficient stochastic approximation EM algorithm using conditional particle filters. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6274–6278, 2013.

[35] F. Lindsten, M. I. Jordan, and T. B. Schön. Particle Gibbs with ancestor sampling. *Journal of Machine Learning Research*, 15(1):2145–2184, 2014.

[36] L. Ljung. MATLAB System Identification Toolbox User's Guide, Copywrite 1988-2018. The MathWorks, Inc.

[37] L. Ljung. On the consistency of prediction error identification methods. In *System Identification Advances and Case Studies*, volume 126, pages 121 – 164. Elsevier, 1976.

[38] L. Ljung. Some limit results for functionals of stochastic processes. Technical Report 167, Linkping University, 1977.

[39] L. Ljung. Convergence analysis of parametric identification methods. *IEEE Transactions on Automatic Control*, 23(5):770–783, 1978.

[40] L. Ljung. *System Identification: Theory for the User*. Prentice Hall, 2nd edition, 1999.

[41] L. Ljung. Estimating linear time-invariant models of nonlinear time-varying systems. *European Journal of Control*, 7(2):203 – 219, 2001.

[42] L. Ljung. Perspectives on system identification. *Annual Reviews in Control*, 34(1):1 – 12, 2010.

[43] L. Ljung and P. E. Caines. Asymptotic normality of prediction error estimators for approximate system models. *Stochastics*, 3(1-4):29–46, 1980.

[44] M. Milanese, J. Norton, H. Piet-Lahanier, and É. Walter. *Bounding approaches to system identification*. Plenum Press, 1996.

[45] G. Mzyk. *Combined Parametric-Nonparametric Identification of Block-Oriented Systems*. Springer, 2013.

[46] C. A. Naesseth, F. Lindsten, and T. B. Schön. High-dimensional filtering using nested sequential Monte Carlo. *arXiv preprint arXiv:1612.09162*, 2016.

[47] O. Nelles. *Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models*. Springer, 2001.

[48] B. Ninness, A. Wills, and A. Mills. UNIT: A freely available system identification toolbox. *Control Engineering Practice*, 21(5):631 – 644, 2013.

[49] B. Ninness, A. Wills, and T. B. Schön. Estimation of general nonlinear state-space systems. In *49th IEEE Conference on Decision and Control, Atlanta, Georgia, USA*, pages 1–6, 2010.

[50] J. Olsson and J. Westerborn. Efficient particle-based online smoothing in general hidden markov models: the paris algorithm. *Bernoulli*, 23(3):1951–1996, 2017.

[51] J. Olsson and J. Westerborn. Particle-based, online estimation of tangent filters with application to parameter estimation in nonlinear state-space models. *arXiv preprint arXiv:1712.08466*, 2017.

[52] A. Padilla, H. Garnier, and M. Gilson. Version 7.0 of the CONTSID toolbox. *IFAC-PapersOnLine*, 48(28):757 – 762, 2015.

[53] J. Paduart, L. Lauwers, J. Swevers, K. Smolders, J. Schoukens, and R. Pintelon. Identification of nonlinear systems using polynomial nonlinear state space models. *Automatica*, 46(4):647 – 656, 2010.

[54] R. Pintelon and J. Schoukens. *System Identification: A Frequency Domain Approach*. Wiley, 2nd edition, 2012.

[55] B. T. Polyak. A new method of stochastic approximation type. *Automation and Remote Control*, 51(7):937–946, 1990.

[56] T. B. Schön, A. Wills, and B. Ninness. System identification of nonlinear state-space models. *Automatica*, 47(1):39 – 49, 2011.

[57] J. Schoukens, A. Marconato, R. Pintelon, Y. Rolain, M. Schoukens, K. Tiels, L. Vanbeylen, G. Vandersteen, and A. Van Mulders. System identification in a real world. In *13th IEEE International Workshop on Advanced Motion Control*, pages 1–9, 2014.

[58] J. Schoukens, M. Vaes, and R. Pintelon. Linear system identification in a nonlinear setting: Nonparametric analysis of the nonlinear distortions and their impact on the best linear approximation. *IEEE Control Systems*, 36(3):38–69, 2016.

[59] M. Schoukens and JP. Noël. Three benchmarks addressing open challenges in nonlinear system identification. *IFAC-PapersOnLine*, 50(1):446 – 451, 2017.

[60] M. Schoukens and K. Tiels. Identification of block-oriented nonlinear systems starting from linear approximations: A survey. *Automatica*, 85:272 – 292, 2017.

[61] J. Sjöberg. On estimation of nonlinear black-box models: how to obtain a good initialization. In *Proceedings of the 1997 IEEE Signal Processing Society Workshop*, pages 72–81, 1997.

[62] J. Sjöberg and J. Schoukens. Initializing Wiener-Hammerstein models based on partitioning of the best linear approximation. *Automatica*, 48(2):353 – 359, 2012.

[63] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, PY. Glorennec, H. Hjalmarsson, and A. Juditsky. Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, 31(12):1691 – 1724, 1995.

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

[65] A. Svensson, T. B. Schön, and F. Lindsten. Learning of state-space models with highly informative observations: A tempered sequential Monte Carlo solution. *Mechanical Systems and Signal Processing*, 104:915 – 928, 2018.

[66] B. Wahlberg and L. Ljung. Algorithms and performance analysis for stochastic Wiener system identification. *IEEE Control Systems Letters*, 2(3):471–476, July 2018.

[67] B. Wahlberg, J. Welsh, and L. Ljung. Identification of Wiener systems with process noise is a nonlinear errors-in-variables problem. In *53rd IEEE Conference on Decision and Control*, pages 3328–3333, Dec 2014.

[68] G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.

[69] A. Wigren, L. Murray, and F. Lindsten. Improving the particle filter in high dimensions using conjugate artificial process noise. *IFAC-PapersOnLine*, 51(15):670 – 675, 2018.

[70] A. Wills, T. B. Schön, L. Ljung, and B. Ninness. Identification of Hammerstein-Wiener models. *Automatica*, 49(1):70–81, January 2013.

[71] H. Wold. *A Study in the Analysis of Stationary Time Series*. Almqvist & Wiksells boktrycheri-a.-b., 1938.

[72] N. Young. *An Introduction to Hilbert Space*. Cambridge University Press, 1988.

[73] ZD. Yuan and L. Ljung. Unprejudiced optimal open loop input design for identification of transfer functions. *Automatica*, 21(6):697 – 708, 1985.

A Wold's decomposition

A relevant result that gives insights regarding the structure of second-order stochastic processes is Wold's decomposition introduced in [71] and its extension in [12].

Theorem 28 (Extension of Wold's decomposition to non-stationary stochastic processes) *For every given n-dimensional stochastic process $\boldsymbol{y}$ with finite second moments and mean function m_t, there is a uniquely determined decomposition $\boldsymbol{y}_t - m_t = \boldsymbol{y}_t^r + \boldsymbol{y}_t^d \quad t \in \mathbb{Z}$ with the following properties:*

(a) the processes $\boldsymbol{y}^r$ and $\boldsymbol{y}^d$ are orthogonal and $\boldsymbol{y}_t^r, \boldsymbol{y}_t^d \in \mathcal{H}_t \subset \mathsf{L}_2^n \ \ \forall t \in \mathbb{Z}$,

(b) the process $\boldsymbol{y}^d$ is linearly deterministic, i.e., $\boldsymbol{y}_t^d \in \mathcal{H}_{t-s} \subset \mathsf{L}_2^n \ \ \forall t, s \in \mathbb{N}$,

(c) the process $\boldsymbol{y}^r$ is purely non-deterministic and

$$\boldsymbol{y}_t^r = \sum_{k=0}^{\infty} h_k(t)\boldsymbol{\varepsilon}_{t-k}, \quad t \in \mathbb{Z} \tag{A.1}$$

in which $\boldsymbol{\varepsilon}_t \in \mathsf{L}_2^n$ is the innovation in $\boldsymbol{y}_t$ with a covariance matrix $\Lambda_t := \mathbb{E}[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^\top]$ satisfying

$$\|\Lambda_t\| < \infty \ \ \forall t \in \mathbb{Z}, \quad \textit{and} \quad \mathbb{E}[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j^\top] = 0 \ \ \forall k \neq j \in \mathbb{Z},$$

where $\|\cdot\|$ is the squared Frobenius norm, such that

$$\sum_{k=0}^{\infty} h_k(t)\Lambda_{t-k}h_k^\top(t) \succeq 0, \quad \sum_{k=0}^{\infty} \|h_k(t)\Lambda_{t-k}h_k^\top(t)\| < \infty$$
$$\forall t \in \mathbb{Z} \quad \textit{where}$$
$$h_n(t)\Lambda_{t-n} = \mathbb{E}[\boldsymbol{y}_t \boldsymbol{\varepsilon}_{t-s}^\top] = \mathbb{E}[\boldsymbol{y}_t^r \boldsymbol{\varepsilon}_{t-s}^\top] \quad \forall s \in \mathbb{N}_0, \ \textit{and}$$
$$h_0(t)\Lambda_t = \Lambda_t = \Lambda_t h_0^\top(t), \quad \forall t \in \mathbb{Z}. \tag{A.2}$$

Furthermore, if the covariance matrix of $\boldsymbol{\varepsilon}_{t-s} \ \forall s \in \mathbb{N}, \ \forall t \in \mathbb{Z}$ is full rank, the sequence $\{h_k(t) : k \in \mathbb{N}_0, \ t \in \mathbb{Z}\}$ is uniquely determined and $h_0(t) = I \ \forall t \in \mathbb{Z}$.

PROOF. The proof is due to Harold Cramér in [12].

The symbol $\mathcal{H}_t$ denotes the span of $\{\boldsymbol{y}_s : s \leq t\}$. When the linear deterministic part $\boldsymbol{y}^d$ is zero, the process $\boldsymbol{y}$ is known as a purely non-deterministic process. The last part of Theorem 28 states that the second-order properties of a purely non-deterministic full rank process $\boldsymbol{y} \subset \mathsf{L}_2^n$ correspond to the pair of sequences $(\{h_k(t) : k \in \mathbb{N}_0, \ t \in \mathbb{Z}\}, \{\Lambda_t : t \in \mathbb{Z}\})$. Once the second is given or normalized, the first is determined uniquely (see the third row of (A.2)). In general, the process $\boldsymbol{y}_t^r$ is given as the output of a time-varying filter, whose impulse response sequence is $\{h_k(t)\Sigma_t : k \in \mathbb{N}_0, \ t \in \mathbb{Z}\}$, due to a white noise input $\tilde{\varepsilon}_t = \Sigma_t^{-1}\varepsilon_t$ where Σ_t is any sequence of positive definite square matrices such that $\{h_k(t)\Sigma_t\}$ is *square summable*. The following example shows the uniqueness of Wold's decomposition using a moving-average process.

Example 29 *Consider the second-order discrete-time stationary stochastic process given by*

$$\boldsymbol{y}_t = \boldsymbol{e}_t - 2\boldsymbol{e}_{t-1}, \quad t \in \mathbb{Z}, \tag{A.3}$$

where $\boldsymbol{e}_t$ is white noise with unit variance. Observe that we may write $0.5\boldsymbol{y}_t = 0.5\boldsymbol{e}_t - \boldsymbol{e}_{t-1} = (0.5\mathrm{q} - 1)\boldsymbol{e}_{t-1}, t \in \mathbb{Z}$ in which q is the forward-shift operator (see [5]). Assuming zero initial conditions, the solution is given by $\boldsymbol{e}_{t-1} = -0.5\sum_{k=0}^{\infty} 0.5^k \boldsymbol{y}_{t+k}$, showing that $\boldsymbol{e}_{t-1} \notin \mathcal{H}_{t-1}$. Therefore, (A.3) is not Wold's decomposition of $\boldsymbol{y}$. To get Wold's decomposition, we need to write $\boldsymbol{y}_t$ in terms of the innovations $\{\boldsymbol{\varepsilon}_s\}_{s \leq t}$. Notice that, by redefining the white process in (A.3), $\boldsymbol{\varepsilon}_t - 0.5\boldsymbol{\varepsilon}_{t-1} = \boldsymbol{y}_t \ \Leftrightarrow \ \boldsymbol{\varepsilon}_t = \sum_{k=0}^{\infty} 0.5^k \boldsymbol{y}_{t-k} = \boldsymbol{e}_t - 3\sum_{k=1}^{\infty} 0.5^k \boldsymbol{e}_{t-k}$ and it follows that $\boldsymbol{\varepsilon}_t$ is the innovation process and Wold's decomposition of $\boldsymbol{y}$ is $\boldsymbol{y}_t = \boldsymbol{\varepsilon}_t - 0.5\boldsymbol{\varepsilon}_{t-1}$, with $\mathbf{var}(\boldsymbol{\varepsilon}_t) = 4 \ \forall t \in \mathbb{Z}$. It is obvious that Wold's decomposition is an incomplete representation that captures only the second-order properties of the process (compare to Problem 3T.4 in [40]).