Identification of Stochastic Nonlinear Models Using Optimal Estimating Functions ⋆

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 first part of the paper examines the asymptotic properties of linear prediction error method estimators, which were recently suggested for the identification of nonlinear stochastic dynamical models. It is shown that their accuracy depends not only on the shape of the unknown distribution of the data, but also on how the model is parameterized. Therefore, it is not obvious in general which linear prediction error method should be preferred. In the second part, the estimating functions approach is introduced and used to construct estimators that are asymptotically optimal with respect to a specific class of estimators. These estimators rely on partial probabilistic parametric models, and therefore neither require the computations of the likelihood function nor any marginalization integrals. The convergence and consistency of the proposed estimators are established under standard regularity and identifiability assumptions akin to those of prediction error methods. The paper is concluded by several numerical simulation examples.

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

1 Introduction

System identification is a vital topic with an essential role in scientific and technological development [37]. Driven by progressively more complex and challenging applications, the need for more accurate and reliable models has never been higher. In some situations, linear system identification may be used to obtain acceptable models, even when the underlying system is nonlinear; see [45,15,56,57,55]. However, nonlinear models are often required in order to capture the main characteristics of the system. Classical literature on nonlinear system identification considered cases with explicit correspondence between observations and innovations such that the optimal Mean-Square Error (MSE) predictor and the likelihood function can be computed analytically; see the surveys [5,28,61,46,58], the articles [33,59,51,60] and the books [49,21,6,48]. However, relaxing this assumption leads to models with computationally intractable likelihood and probability density functions.

During the last decade, several methods that may be used to deal with general stochastic nonlinear models have been introduced. For example, approximate Maximum Likelihood (ML) and Bayesian methods have been developed; see, for example, [29,50,54,69,41,66,64,20,19]. They are mainly based on sequential Monte Carlo (a.k.a particle filters/smoothers) and Markov chain Monte Carlo approximations (see [53]), and have been shown to provide impressive results on several examples and benchmark problems. However, depending on the application, they may be computationally expensive or even infeasible. Moreover, they require the specification of a full probabilistic model, and if the model is misspecified, these methods loose the optimal properties that are usually the main justification for their use.

Recently, alternative methods for estimation and approximate analysis were suggested in [65]. They are based on Taylor approximations, and therefore the analysis is valid only locally and the obtained estimators may not be consistent. Furthermore, third moments of the model's outputs are neglected in the covariance expressions, which means that the results are valid only if the data are independent over time and have symmetric distributions, severely limiting their applicability.

Estimation methods based on the Prediction Error Method (PEM) and linear predictors have been intro-

⋆ This work was supported by the Swedish Research Council via the projects 2016-06079 (NewLEADS), 2015-05285, and 2019-04956. This paper was not presented at any IFAC meeting. Corresponding author M. Abdalmoaty.

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

Accepted for publication in Automatica *18 April 2020*

duced in [2]. They rely on the first two moments of the model and are relatively easy to compute: the computations of the likelihood function are not required. The resulting estimators are, under standard regularity and identifiability assumptions, consistent and asymptotically normal. However, in general, it is not obvious which linear PEM estimator should be preferred. For instance, in the simulation example in [2, Section 8.1], the weighted output-error PEM estimator is more accurate than the Gaussian PEM estimator (denoted OE-WQPEM and OL-GPEM, respectively, there). In another example in [1], the (unweighted) output-error PEM estimator is more accurate than a PEM based on the marginal density function, unlike what may be expected. Therefore, there is a need for a systematic way of constructing estimators based on partial probabilistic knowledge – perhaps beyond the first two moments – that have covariance matrices uniformly smaller than, for example, the Gaussian Cramér-Rao Lower Bound (CRLB).

1.1 Contributions

The paper may be divided into two main parts. Firstly, we study the asymptotic behavior of the quadratic linear PEM estimators proposed in [2] (the OE-WQPEM and the OL-GPEM estimators). We show that a PEM based on a Gaussian assumption (OL-GPEM), where the mean vector and the covariance matrix of the model's outputs are jointly parameterized, is not necessarily more accurate compared to an optimally weighted output-error PEM (optimal OE-WQPEM). This comes in contrast to the claims in [65, Section IV-D]. Moreover, we show that the third and fourth moments play an important role in this behavior and establish assumptions under which one of these PEMs may be preferred. Secondly, we introduce the Estimating Functions (EFs) approach (see the seminal papers [14,22] and the books [24,32]) which can be seen as a generalization of the ML method and the PEMs. It has been developed mainly in the statistics literature with varying areas of applications such as survey sampling, stochastic processes, and biostatistics [24]. We show how it can be used to define optimal estimators among estimators based on partial probabilistic models. We propose the use of optimal linear and quadratic EFs which, in some common cases, assume closed-form expressions. This leads to estimators that are asymptotically uniformly more accurate compared to linear PEMs with quadratic criteria. The convergence of the optimal EFs is established under standard regularity assumptions, and the consistency and asymptotic normality of the corresponding estimators are given under certain identifiability assumptions.

1.2 Paper outline

The problem is formulated in Section 2. In Section 3, the ML method is briefly discussed and some properties of the score function are reviewed. The PEM is considered in Section 4: the relation to the ML method is highlighted, and the asymptotic properties of linear PEM estimators, defined via a quadratic criterion function, are examined. In Section 5, the EFs approach is introduced and then used in Section 6 to systematically construct optimal estimators. The convergence and consistency of the proposed estimators are established in Sections 7 and 8. Section 9 considers the kinship between the EF approach and the prediction error correlation method. Numerical simulation examples are given in Section 10.

1.3 Notations

Bold font is used to denote random quantities while 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-dimensional 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 [7]. For brevity, we simply use L_2^n and drop the arguments. 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 $p_{\boldsymbol{x}}$, while $\boldsymbol{x}_t \overset{\text{i.i.d}}{\sim} p_{\boldsymbol{x}}$ means that the random variables $\{\boldsymbol{x}_t\}$ are independent and identically distributed according to $p_{\boldsymbol{x}}$. The symbols $\xrightarrow{\text{a.s.}}$ and $\rightsquigarrow$ denote almost sure convergence and convergence in distribution, respectively (see [9, Chapter 4]). For a matrix M, the notation $[M]_{ij}$ denotes the ijth-entry of M. For an n-dimensional vector V, the notation $[V]_i$ denotes the ith-entry of V, $\partial_\theta V(\theta)$ is an $n \times d_\theta$ matrix. The notation $\partial_{\theta_i} V(\theta)$ denotes the derivative of $V(\theta)$ with respect to $[\theta]_i$, and $\partial_{\theta_{ij}} V(\theta)$ denotes the derivative of $\partial_{\theta_i} V(\theta)$ with respect to $[\theta]_j$. For a real-valued function $\mathcal{V}(\theta)$, the symbol $\partial_\theta^r \mathcal{V}(\theta_\circ)$ denotes $\frac{\partial^r}{\partial \theta^r}\mathcal{V}(\theta)\big|_{\theta=\theta_\circ}$ where $r \in \{1,2\}$. For a matrix valued function $M(\theta)$ we use $M^{-1}(\theta)$ and $M^\top(\theta)$ to denote $[M(\theta)]^{-1}$ and $[M(\theta)]^\top$ respectively.

2 Problem Formulation

The identification problem is formulated in a probabilistic setup where 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 a probability space $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$, $d_y \in \mathbb{N}$. The probability measure $\mathbb{P}_\theta$ is parameterized by a real vector $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$, d_θ is a finite positive integer, and an *a priori known fixed* d_u-dimensional input signal $u := \{u_t : t \in \mathbb{Z}\}$, $d_u \in \mathbb{N}$. For the consistency analysis, it will be assumed that a true parameter $\theta_\circ$ exists in the interior of Θ such that the data $\{y_1, \ldots, y_N\}$ is a subsequence of a realization y of $\boldsymbol{y} \sim \mathbb{P}_{\theta_\circ}$. Moreover, we assume that the first four moments of $\boldsymbol{y}$ exist for all $\theta \in \Theta$.

This is a general setup that includes both static and dynamical models. Nevertheless, we are interested in cases where $\mathbb{P}_\theta$ is either fully or partially determined using one of the commonly used model structures in system identification such as: stochastic nonlinear state-space models [44, Section 5.3], block-oriented models (e.g., stochastic Wiener-Hammerstein models) [21,48], basis function expansions [61], etc. Of particular interest are cases where the likelihood function is computationally intractable.

Define the data set

$$\boldsymbol{D}_N := \{(\boldsymbol{y}_k, u_k) : k = 1, \dots N\} \tag{1}$$

that contains pairs of inputs and outputs up to some time $N \in \mathbb{N}$. The problem studied in this paper is the construction of point estimators $\hat{\theta}_N$ of $\theta_\circ$ based on a given data set $\boldsymbol{D}_N$. Particular attention is given to the accuracy of the estimators. Define the random vector $\boldsymbol{Y}_N := [\boldsymbol{y}_1^\top, \dots, \boldsymbol{y}_N^\top]^\top$, by stacking the elements of the subsequence $\{\boldsymbol{y}_1, \dots, \boldsymbol{y}_N\}$ in a column vector. For brevity, we will use the notation $\boldsymbol{Y} := \boldsymbol{Y}_N$, and $\hat{\boldsymbol{\theta}}_N = \hat{\theta}(\boldsymbol{D}_N)$.

Two favored point estimation methods in system identification are the ML method and the PEM [27,62,44,52]. We start by discussing the main features of the ML method, and then in Section 4 discuss the PEM.

3 The Maximum Likelihood Method

Assume that for all $\theta \in \Theta$, the random vector $\boldsymbol{Y}$ has a density $p(\boldsymbol{Y}; \theta)$ with respect to the Lebesgue measure on $\mathbb{R}^{d_y N}$. Then the likelihood function of θ is given by $p(Y; \theta)$, seen as a function of θ. The Maximum Likelihood Estimator (MLE), when it exists, is defined as the global maximizer of the (log-)likelihood function, that is

$$\hat{\boldsymbol{\theta}}_N := \arg\max_{\theta \in \Theta} p(\boldsymbol{Y}; \theta). \tag{2}$$

Hence, the ML method requires the specification of a full probabilistic model (i.e., the form of the true distribution has to be known). If $p(Y; \theta)$ is differentiable in θ over an open neighborhood of Θ for all Y, the *score function* is defined as the $\mathbb{R}^{d_\theta}$-valued random vector

$$\boldsymbol{S}_N(\theta) := \partial_\theta \log p(\boldsymbol{Y}; \theta) = \frac{\partial_\theta p(\boldsymbol{Y}; \theta)}{p(\boldsymbol{Y}; \theta)}, \quad \theta \in \Theta,$$

and according to the definition in (2), the MLE is a root of the score function [1]; i.e., $\hat{\boldsymbol{\theta}}_N \in \underset{\theta \in \Theta}{\text{sol}} [\boldsymbol{S}_N(\theta) = 0]$. Let us further assume the following.

Assumption 1 $\boldsymbol{S}_N(\theta) \in \mathsf{L}_2^{d_\theta}$ $\forall \theta \in \Theta$.

Assumption 2 *The support of the density function $p(\boldsymbol{Y}; \theta)$ is independent of θ, and the real function $(Y, \theta) \mapsto \partial_\theta p(Y; \theta)$, where $Y \in \mathbb{R}^{d_y N}$ and $\theta \in \Theta$, is locally dominated integrable with respect to the Lebesgue measure on $\mathbb{R}^{d_y N}$ (see [40] or [8]), and therefore $\int \partial_\theta p(Y; \theta)\, \mathrm{d}Y = \partial_\theta \mathbb{E}[1; \theta] = 0$.*

Under Assumptions 1 and 2, the score function has the following well-known important properties. First, it follows immediately from Assumption 2 that $\mathbb{E}[\boldsymbol{S}_N(\theta); \theta] = 0$ $\forall \theta \in \Theta$. Second, the covariance matrix of the score function exists and is equivalently given by $-\mathbb{E}[\partial_\theta \boldsymbol{S}_N(\theta_\circ); \theta_\circ]$ (see [39, Theorem 7.5.1]). Furthermore, when certain regularity assumptions hold (e.g., as shown in [38,4]), the inverse of the matrix $\lim_{N\to\infty} \frac{1}{N} \mathbb{E}[\boldsymbol{S}_N(\theta)[\boldsymbol{S}_N(\theta)]^\top; \theta]\big|_{\theta=\theta_\circ}$, when it exists, equals the asymptotic covariance matrix of the MLE, which we denote by P^{ML}. It gives a lower bound on the asymptotic MSE of most consistent estimators except for parameters in a Lebesgue null set. In other words, the MLE is asymptotically efficient, which is usually the main justification for its use.

From the above description, it is clear that the properties of the ML method are related to the score function that plays the key role in both the analysis and the computations of the MLE; see [17,39]. Unfortunately, inference based on the score function has a couple of weaknesses. First, the computations of the score function require evaluations of the log-likelihood function and its gradient. Depending on the model, this may be a very challenging task; see [3,35,53]. Second, the performance is sensitive to the distributional assumptions; e.g., the optimal properties of the MLE are only valid if the distribution is correctly specified, otherwise the estimator may even become inconsistent; see for example [68].

4 The Prediction Error Methods

Prediction error method estimators are defined by minimizing the discrepancy between the measured outputs and predicted outputs based on the hypothesized model (see [44, Chapter 7] or [62, Chapter 7]). Let us denote the criterion function defining the PEM estimator by

$$\boldsymbol{\mathcal{V}}_N(\theta) = \frac{1}{N} \sum_{t=1}^{N} \ell(\boldsymbol{e}_t(\theta), t, \theta), \tag{3}$$

where ℓ is a user-defined non-negative real-valued function, $\boldsymbol{e}_t(\theta) := \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}(\theta)$ is the Prediction Error (PE) process, and $\hat{\boldsymbol{y}}_{t|t-1}(\theta)$ is a user-defined one-step-ahead predictor function. A PEM estimator $\hat{\boldsymbol{\theta}}_N$ is then defined as the global solution to the problem

$$\min_{\theta \in \Theta} \boldsymbol{\mathcal{V}}_N(\theta). \tag{4}$$

Under some assumptions (see [47]), it holds that

$$\sqrt{N} P_N^{-\frac{1}{2}} (\hat{\boldsymbol{\theta}}_N - \theta_\circ) \rightsquigarrow \mathcal{N}(0, I_{d_\theta}) \quad \text{as} \quad N \to \infty, \tag{5}$$

[1] The MLE may also be defined as any root of the likelihood equations $S_N(\theta) = 0$; see [10, page 499] for example.

$$\begin{aligned} P_N &:= [\partial_\theta^2 \mathcal{W}_N(\theta_\circ)]^{-1} \mathcal{U}_N(\theta_\circ)\, [\partial_\theta^2 \mathcal{W}_N(\theta_\circ)]^{-1}, \\ \mathcal{U}_N(\theta) &:= \mathbb{E}[N \partial_\theta \boldsymbol{\mathcal{V}}_N(\theta)\, \partial_\theta \boldsymbol{\mathcal{V}}_N^\top(\theta)], \\ \mathcal{W}_N(\theta) &:= \mathbb{E}[\boldsymbol{\mathcal{V}}_N(\theta)]. \end{aligned} \tag{6}$$

The factors P_N in (5) play an important role in this paper, and *we will refer to them as the normalizing factors* (regardless of how $\hat{\boldsymbol{\theta}}_N$ is defined). In the special case where $\mathcal{U}_N(\theta_\circ) \to \bar{\mathcal{U}}(\theta_\circ)$, $\mathcal{W}_N(\theta) \to \bar{\mathcal{W}}(\theta)$ uniformly over Θ as $N \to \infty$, such that $\theta_\circ$ is a unique global minimum of $\bar{\mathcal{W}}(\theta)$, $\partial_\theta^2 \bar{\mathcal{W}}(\theta_\circ) \succ 0$, $\bar{\mathcal{U}}(\theta_\circ) \succ 0$ and $\sqrt{N} \partial_\theta \mathcal{W}_N(\theta_\circ) \to 0$, it holds that $\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \theta_\circ) \rightsquigarrow \mathcal{N}(0, P^{\text{PEM}})$ where $P^{\text{PEM}} := [\partial_\theta^2 \bar{\mathcal{W}}(\theta_\circ)]^{-1} \bar{\mathcal{U}}(\theta_\circ) [\partial_\theta^2 \bar{\mathcal{W}}(\theta_\circ)]^{-1}$ is the asymptotic covariance matrix.

Observe that because the MLE is asymptotically efficient, it holds that $P^{\text{PEM}} \succeq P^{\text{ML}}$ for every $\theta_\circ$ and all choices of ℓ and $\hat{\boldsymbol{y}}_{t|t-1}$ that lead to asymptotically normal estimators. Equality is obtained, for example, when for some $c > 0$, $\partial_\theta \boldsymbol{\mathcal{V}}_N(\theta) = -c \boldsymbol{S}_N(\theta)\ \forall \theta \in \Theta$, at least asymptotically. This will be the case for example when

$$\ell(\boldsymbol{e}_t(\theta), t, \theta) := -\log(p(\boldsymbol{e}_t(\theta) | \boldsymbol{e}_1(\theta), \ldots, \boldsymbol{e}_{t-1}(\theta); \theta))$$

where $p(\boldsymbol{e}_t(\theta) | \boldsymbol{e}_1(\theta), \ldots, \boldsymbol{e}_{t-1}(\theta); \theta)$ is the conditional probability density function of $\boldsymbol{e}_t(\theta)$ when $\theta_\circ = \theta$. However, this definition is not standard in the PEM literature because here we allow ℓ to depend on previous prediction errors. Nevertheless, this still gives a well-defined problem (4) where the size of the prediction errors is measured using a data-dependent scalar function. [2] In the special case where $\hat{\boldsymbol{y}}_{t|t-1}(\theta) := \mathbb{E}[\boldsymbol{y}_t | Y_{t-1}; \theta]$ and $\{\boldsymbol{e}_t(\theta)\}$ is a sequence of independent random variables when $\theta_\circ = \theta$, the conditioning in the definition of ℓ is not required, namely $\ell(\boldsymbol{e}_t(\theta), t, \theta) := -\log(p(\boldsymbol{e}_t(\theta); \theta))$ (see [44, page 216]). The commonly used stochastic assumption in the PEM literature is that the PE process $\boldsymbol{e}(\theta)$ is standard Gaussian when $\theta_\circ = \theta$, i.e., $\boldsymbol{e}_t(\theta_\circ) \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I_{d_y})$. This leads to the standard choices $\hat{\boldsymbol{y}}_{t|t-1}(\theta) := \mathbb{E}[\boldsymbol{y}_t | Y_{t-1}; \theta]$ and $\ell(\boldsymbol{e}_t(\theta), t, \theta) := \|\boldsymbol{e}_t(\theta)\|^2$. However, all the above choices narrow the scope of the PEMs significantly and lead to the same computational problems usually faced in likelihood estimation: the computations of conditional distributions and conditional expectations are analytically intractable in general (see [13,53]). The full potential of the PEMs, in terms of balancing estimation accuracy and computational cost, becomes truly apparent when the flexibility in the choice of ℓ and $\hat{\boldsymbol{y}}_{t|t-1}$ is exploited. The following question then arises: given a parametric model, how to select *tractable* ℓ and $\hat{\boldsymbol{y}}_{t|t-1}$ such that the resulting estimator satisfies desired asymptotic properties, such as consistency and asymptotic normality?

In the recent article [2], the use of the two predictors

$$\begin{aligned} \hat{y}_t(\theta) &:= \mathbb{E}[\boldsymbol{y}_t; \theta], \text{and} \\ \hat{\boldsymbol{y}}_{t|t-1}(\theta) &:= \mathbb{E}[\boldsymbol{y}_t; \theta] + \sum_{k=1}^{t-1} \tilde{l}_{t-k}(t, \theta)(\boldsymbol{y}_k - \mathbb{E}[\boldsymbol{y}_k; \theta]), \end{aligned} \tag{7}$$

where the sequence $\{\tilde{l}_{t-k}(t, \theta)\}$ is chosen to minimize $\mathbb{E}[\|\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}(\theta)\|^2; \theta]$, were suggested. The first is a deterministic function and is referred to as the Output-Error predictor (OE-predictor). The second is linear in the past outputs and is referred to as the Optimal Linear predictor (OL-predictor). As shown in [2], these are relatively simple predictors that depend only on the first two moments of the model. By combining them with various choices for ℓ, computationally attractive PEM estimators can be defined. The simplest and most common choice for ℓ is the squared Euclidean norm; the corresponding estimators in this case are referred to as the OE-QPEM (OE Quadratic PEM) estimator and the OL-QPEM (OL Quadratic PEM) estimator, respectively. Alternative choices for ℓ are a weighted squared Euclidean norm or the negative logarithm of the marginal probability density function $\tilde{p}(\boldsymbol{e}_t(\theta); \theta)$; the latter was suggested in [1]. However, it is not clear in general which choice leads to a more accurate estimator. In the remaining part of this section we study this issue in some detail.

4.1 Optimal weights for quadratic ℓ

Define the vectors

$$\begin{aligned} \boldsymbol{E}(\theta) &:= \boldsymbol{Y} - \mu(\theta), \quad \mu(\theta) := \mathbb{E}[\boldsymbol{Y}; \theta], \\ \boldsymbol{\mathcal{E}}(\theta) &:= \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\theta) =: \left[\boldsymbol{\varepsilon}_1^\top(\theta) \ldots \boldsymbol{\varepsilon}_N^\top(\theta)\right]^\top \end{aligned} \tag{8}$$

where $\widehat{\boldsymbol{Y}}(\theta) := \left[\hat{\boldsymbol{y}}_{1|0}^\top(\theta)\ \hat{\boldsymbol{y}}_{2|1}^\top(\theta)\ \ldots\ \hat{\boldsymbol{y}}_{N|N-1}^\top(\theta)\right]^\top$ is a column vector of OL-predictors, and the matrix

$$\Sigma(\theta) := \mathbf{cov}(\boldsymbol{Y}, \boldsymbol{Y}; \theta). \tag{9}$$

Then, it holds that the vector of linear innovations

$$\boldsymbol{\mathcal{E}}(\theta) = L^{-1}(\theta) \boldsymbol{E}(\theta), \tag{10}$$

where $L(\theta)$ is a lower unitriangular matrix in the $LDL^\top$ decomposition of $\Sigma(\theta)$ (see [2, Lemma 7]).

Consider the following two cases:

(1) Let Q be a positive definite $d_y N \times d_y N$ *parameter-independent matrix*, and let $L \Lambda L^\top := Q$ be its block $LDL^\top$ decomposition, where the blocks are of size $d_y \times d_y$. Let $\ell^{(1)}(\boldsymbol{e}_t, t; \theta) := \frac{1}{2} \boldsymbol{e}_t^\top Q_t^{-1} \boldsymbol{e}_t$, in which $Q_t \succ 0$ is the tth $d_y \times d_y$ block of Λ, and consider the (filtered) PEs $\boldsymbol{e}_t(\theta) = [L^{-1} \boldsymbol{E}(\theta)]_t$.

(2) Let $\ell^{(2)}(\boldsymbol{e}_t, t; \theta) := \frac{1}{2} \boldsymbol{e}_t^\top Q_t^{-1}(\theta) \boldsymbol{e}_t + \frac{1}{2} \log\det Q_t(\theta)$, in which the PEs correspond to the linear innovation process: $\boldsymbol{e}_t(\theta) = \boldsymbol{\varepsilon}_t(\theta) := \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}(\theta)$, where $\hat{\boldsymbol{y}}_{t|t-1}(\theta)$ is the OL-predictor given by the second row of (7) and the matrix $Q_t(\theta)$ is the covariance matrix of $\boldsymbol{\varepsilon}_t(\theta)$ at time t.

[2] Allowing for these cases provides a motivation for the use of a time-dependent ℓ, even when $\boldsymbol{e}(\theta_\circ)$ is strictly stationary (see [44, page 200] for the commonly used motivation). It also makes the MLEs a strict subset of PE estimators.

Note that the log det term in the definition of $\ell^{(2)}$ is necessary for consistency: the matrices $Q_t(\theta)$ and the PEs are jointly parameterized by θ, and there are no additional parameters compared to the first case. We will refer to the estimator corresponding to $\ell^{(1)}$ as the OE-WQPEM (weighted OE-QPEM) estimator, and the estimator corresponding to $\ell^{(2)}$ as the OL-GPEM (OL Gaussian PEM) estimator because the criterion function is on the form of a Gaussian log-likelihood function.

Our goal is to find the optimal weighting matrix, denoted $Q^\star$, in the first case, and then compare the resulting estimator to the OL-GPEM estimator. We will not assume the convergence of the matrices in (6), and therefore the term "optimal" here is understood in the sense of *minimizing the normalizing factors* P_N of the errors $(\hat{\boldsymbol{\theta}}_N - \theta_\circ)$, *in the partial ordering of positive semidefinite matrices*, such that the convergence in (5) still holds. Let us denote the normalizing factors corresponding to the first case by $P_N^{(1)}(Q)$, in which the dependence on Q is made explicit, and the normalizing factors corresponding to the second case by $P_N^{(2)}$.

Case 1: Parameter-independent weights

We now consider the problem of finding the optimal weighting matrix $Q^\star \succ 0$ for the weighted OE-QPEM estimator

$$\hat{\boldsymbol{\theta}}_N := \arg\min_{\theta\in\Theta} \frac{1}{N}\frac{1}{2}\boldsymbol{E}^\top(\theta)Q^{-1}\boldsymbol{E}(\theta). \tag{11}$$

The following proposition is a straightforward extension of the linear time-invariant (LTI) case. The difference here is the dependence of the weighting matrix on u.

Proposition 3 (Optimal weights for OE-WQPEM) *Consider the OE-WQPEM estimator defined in (11). Suppose that (5) holds, define $Q^\star := \Sigma := \mathbf{cov}(Y, Y; \theta_\circ)$ and assume that $\Sigma \succ 0$. Then $P_N^{(1)}(Q) \succeq P_N^{(1)}(Q^\star)$ $\forall Q \succ 0$, and $P_N^{(1)}(Q^\star) = N\left[\partial_\theta\mu^\top(\theta_\circ)[Q^\star]^{-1}\partial_\theta\mu(\theta_\circ)\right]^{-1}$ where $\mu(\theta)$ is defined in (8).*

PROOF. From (11) and (6), it holds that

$$[\mathcal{U}_N(\theta_\circ)]_{ij} = \frac{1}{N}\partial_{\theta_i}\mu^\top(\theta_\circ)Q^{-1}\Sigma Q^{-1}\partial_{\theta_j}\mu(\theta_\circ),$$
$$\partial_{\theta_{ij}}\mathcal{W}_N(\theta_\circ) = \frac{1}{N}\partial_{\theta_i}\mu^\top(\theta_\circ)Q^{-1}\partial_{\theta_j}\mu(\theta_\circ).$$

Now observe that $P_N^{(1)}(Q)$ is on the same form as the covariance matrix of the PEM estimators in the LTI case (see [62, Section 7.5 and Appendix A7.1]). Therefore, the optimal weighting matrix is $Q^\star = \Sigma$, and

$$P_N^{(1)}(Q^\star) = N\left[\partial_\theta\mu^\top(\theta)\Sigma^{-1}\partial_\theta\mu(\theta)\right]^{-1}$$

because $\mathcal{U}_N(\theta_\circ) = [\partial_\theta^2\mathcal{W}_N(\theta_\circ)]^{-1}$ when $Q = Q^\star$.

We will refer to the estimator defined by (11) when $Q = Q^\star$ as the *optimal OE-WQPEM estimator*.

Because $Q^\star$ depends on the unknown $\theta_\circ$, the optimal OE-WQPEM estimator cannot be directly realized. A solution is to use a two-step procedure where, in the first step, $Q^\star$ is estimated using a consistent estimator; this can be achieved in a few ways depending on the assumptions. An alternative is to directly use the parameterized covariance matrix implied by the model. However, this requires adding a log det term to the criterion function to preserve consistency (see [2]), and leads to the second case where $\ell^{(2)}$ is used. We now argue that, if this option is used, the resulting estimator is not necessarily better than the optimal OE-WQPEM in the sense that it may require larger normalizing factors in (5).

Case 2: Parameter-dependent weights

The criterion function (3) when $\ell^{(2)}$ is used with the OL-predictor can be written on the Gaussian negative log-likelihood function form, and thus the OL-GPEM estimator is given by

$$\hat{\boldsymbol{\theta}}_N := \arg\min_{\theta\in\Theta} \frac{1}{2N}\left[\boldsymbol{E}^\top(\theta)Q^{-1}(\theta)\boldsymbol{E}(\theta) + \log\det Q(\theta)\right] \tag{12}$$

where $Q(\theta)$ is the covariance of $\boldsymbol{Y}$ when $\theta_\circ = \theta$. It is important to note here that $Q(\theta)$ and $\mu(\theta)$ are jointly parameterized. Elementary calculations, which we shall omit, show that

$$\begin{aligned}N\partial_{\theta_{ij}}\mathcal{W}_N(\theta_\circ) &= \partial_{\theta_i}\mu^\top(\theta)Q^{-1}(\theta_\circ)\partial_{\theta_j}\mu(\theta_\circ)\\ &\quad+\frac{1}{2}\mathbf{tr}\left(Q^{-1}(\theta_\circ)\partial_{\theta_j}Q(\theta_\circ)Q^{-1}(\theta_\circ)\partial_{\theta_i}Q(\theta_\circ)\right),\end{aligned}$$

$$\begin{aligned}N[\mathcal{U}_N(\theta_\circ)]_{ij} &= \partial_{\theta_i}\mu^\top(\theta_\circ)Q^{-1}(\theta_\circ)\partial_{\theta_j}\mu^\top(\theta_\circ)\\ &\quad-\frac{1}{4}\mathbf{tr}(Q^{-1}(\theta_\circ)\partial_{\theta_i}Q(\theta_\circ))\mathbf{tr}(Q^{-1}(\theta_\circ)\partial_{\theta_j}Q(\theta_\circ))\\ &\quad+\frac{1}{4}A_{ij}(\theta_\circ) + B_{ij}(\theta_\circ),\ \text{where}\end{aligned}$$

$$A_{ij}(\theta_\circ) := \mathbb{E}\Big[\boldsymbol{E}(\theta_\circ)\partial_{\theta_i}Q^{-1}(\theta_\circ)\boldsymbol{E}(\theta_\circ)\boldsymbol{E}^\top(\theta_\circ)\partial_{\theta_j}Q^{-1}(\theta_\circ)\boldsymbol{E}^\top(\theta_\circ)\Big],$$
$$B_{ij}(\theta_\circ) := \mathbb{E}\Big[\boldsymbol{E}(\theta_\circ)\partial_{\theta_i}Q^{-1}(\theta_\circ)\boldsymbol{E}(\theta_\circ)\boldsymbol{E}^\top(\theta_\circ)Q^{-1}(\theta_\circ)\partial_{\theta_j}\mu^\top(\theta_\circ)\Big].$$

Now observe that the evaluation of $A_{ij}(\theta_\circ)$ and $B_{ij}(\theta_\circ)$ is possible only if up to the fourth moments of the model are known, and that the expression of $P_N^{(2)}$ does not simplify in general. This is because, unlike the first case, $\mathcal{U}_N(\theta_\circ) \neq [\partial_\theta^2\mathcal{W}_N(\theta_\circ)]^{-1}$ in general. However, in the special case where the true model is Gaussian, it holds that

$$\begin{aligned}\frac{1}{4}A_{ij}(\theta_\circ)+B_{ij}(\theta_\circ) =&\\ &\frac{1}{4}\mathbf{tr}(Q^{-1}(\theta_\circ)\partial_{\theta_i}Q(\theta_\circ))\mathbf{tr}(Q^{-1}(\theta_\circ)\partial_{\theta_j}Q(\theta_\circ))\\ &+\frac{1}{2}\mathbf{tr}\left(Q^{-1}(\theta_\circ)\partial_{\theta_j}Q(\theta_\circ)Q^{-1}(\theta_\circ)\partial_{\theta_i}Q(\theta_\circ)\right),\end{aligned}$$

$$N\left[\left[P_N^{(2)}\right]^{-1}\right]_{ij} = \partial_{\theta_i}\mu^\top(\theta_\circ)Q^{-1}(\theta_\circ)\partial_{\theta_j}\mu(\theta_\circ) + \frac{1}{2}\mathbf{tr}\left(Q^{-1}(\theta_\circ)\ \partial_{\theta_j}Q(\theta_\circ)Q^{-1}(\theta_\circ)\partial_{\theta_i}Q(\theta_\circ)\right); \qquad (13)$$

see, for example, [36, Appendix 3C]. Thus,

$$P_N^{(2)} = N[\partial_\theta\mu^\top(\theta_\circ)Q^{-1}(\theta_\circ)\partial_\theta\mu(\theta_\circ) + \bar{P}_N(\theta_\circ)]^{-1}$$

where $\bar{P}_N(\theta_\circ)$ is defined as having the ijth-entry equal to the second term in (13). Because $\bar{P}_N(\theta_\circ)$ is positive semidefinite[3], it holds that

$$P_N^{(1)}(Q^\star) \succeq P_N^{(2)}, \quad \text{for every } \theta_\circ,\ N. \qquad (14)$$

4.2 Analysis for SISO models with a scalar parameter

In this part, we give conditions under which the inequality (14) holds even when the model is non-Gaussian. To simplify the analysis, we will study the problem under the following assumption.

Assumption 4 *The model is such that*
(a) $d_\theta = 1$, i.e., $\Theta \subset \mathbb{R}$,
(b) the output process $\boldsymbol{y}$ is an independent process,
(c) the convergence in (5) holds for the OL-GPEM estimator (12) and the optimal OE-WQPEM estimator, defined by (11) when $Q = Q^\star$.

Denote the central moments of $\boldsymbol{y}$ as:

$$\mu_t := \mathbb{E}[\boldsymbol{y}_t], \qquad \lambda_t := \mathbb{E}[(\boldsymbol{y}_t - \mathbb{E}[\boldsymbol{y}_t])^2],$$
$$m_t^{(3)} := \mathbb{E}[(\boldsymbol{y}_t - \mathbb{E}[\boldsymbol{y}_t])^3], \qquad m_t^{(4)} := \mathbb{E}[(\boldsymbol{y}_t - \mathbb{E}[\boldsymbol{y}_t])^4]$$

where, for brevity, we omitted the dependence on θ. Then, the expression of the normalizing factors of the optimal OE-WQPEM estimator becomes

$$P_N^{(1)}(Q^\star) = N\left(\sum_{t=1}^N A_t\right)^{-1} \qquad (15)$$

and that of the OL-GPEM estimator becomes

$$P_N^{(2)} = \left(N\sum_{t=1}^N (A_t + C_t)\right)\left(\sum_{t=1}^N (A_t + B_t)\right)^{-2} \qquad (16)$$

where

$$A_t := \frac{(\partial_\theta\mu_t)^2}{\lambda_t}, \qquad C_t := D_t + E_t - \frac{1}{2}B_t,$$
$$B_t := \frac{1}{2}\left(\frac{\partial_\theta\lambda_t}{\lambda_t}\right)^2, \quad D_t := \left(\partial_\theta\frac{1}{2\lambda_t}\right)^2 m_t^{(4)}, \qquad (17)$$
$$E_t := -\left(\partial_\theta\frac{1}{\lambda_t}\right)\frac{m_t^{(3)}\partial_\theta\mu_t}{\lambda_t}.$$

We now have the following results.

Theorem 5 *Suppose that Assumption 4 holds. Then, $P_N^{(1)}(Q^\star) \geq P_N^{(2)}$ if and only if*

$$\sum_{t=1}^N (D_t + E_t) \ \leq\ \frac{\left(\sum_{t=1}^N B_t\right)^2}{\sum_{t=1}^N A_t} + \frac{5}{2}\sum_{t=1}^N B_t \qquad (18)$$

where A_t, B_t, D_t and E_t are defined in (17). In particular, if $\boldsymbol{y}$ is fourth-order stationary, (18) reduces to

$$\kappa \leq \left[\frac{(\partial_\theta\lambda)^2}{\lambda\ (\partial_\theta\mu)^2} - \frac{4}{\lambda}\frac{\partial_\theta\mu}{\partial_\theta\lambda}m^{(3)} + 5\right], \qquad (19)$$

where $\kappa := m^{(4)}/\lambda^2$ is the kurtosis of $\boldsymbol{y}$. Moreover if, in addition, the distribution of $\boldsymbol{y}$ is symmetric around μ, the condition becomes

$$\kappa \leq \left[\frac{(\partial_\theta\lambda)^2}{\lambda\ (\partial_\theta\mu)^2} + 5\right]. \qquad (20)$$

PROOF. To simplify the notations, let $A := \sum_{t=1}^N A_t$, and similarly define B, C, D, and E. Using (15) and (16), it follows that $P_N^{(1)}(Q^\star) \geq P_N^{(2)}$ if and only if

$$\frac{A+C}{(A+B)^2} \leq \frac{1}{A} \iff C \leq \frac{B^2}{A} + 2B$$
$$\iff D + E \leq \frac{B^2}{A} + \frac{5}{2}B \qquad (21)$$

which is the same as (18). If $\boldsymbol{y}$ is fourth-order stationary, the first four moments are time-independent and,

$$C_t = \left(\frac{2\partial_\theta\mu}{\lambda\ \partial_\theta\lambda}m^{(3)} + \frac{1}{2}m^{(4)} - \frac{1}{2}\right)B_t,$$

which, when used in (21) together with the definition of κ, gives (19). Finally, (20) follows directly from (19) after recalling that $m^{(3)} = 0$ for any symmetric distribution.

Note that (20) can be satisfied for well-known heavy tailed distributions. For example, a standard Laplace (double exponential) distribution has a kurtosis equal to 6. Moreover, the Pearson type VII parametric family of distributions (see e.g., [63, Chapter 6]) includes one-parameter distributions whose kurtosis may be arbitrarily adjusted while keeping the first three moments fixed.

From the above discussion, we conclude the following:

(1) In general, the inequality (14) does not always hold; its validity does not only depend on the shape of the distribution, but also on how it is parameterized.
(2) In the very special case where $\boldsymbol{y}$ is Gaussian with jointly parameterized mean and covariance, (14) holds. In other words, the optimal OE-WQPEM estimator defined using the *true* covariance matrix is inferior to the OL-GPEM (MLE) estimator because the information provided by the parameterization of the covariance matrix is not used.

[3] Since $[\bar{P}_N(\theta_\circ)]_{ij} := \mathbf{tr}(Q^{-1}(\theta_\circ)\partial_{\theta_j}Q(\theta_\circ)Q^{-1}(\theta_\circ)\partial_{\theta_i}Q(\theta_\circ))$, straightforward calculations show that $x^\top\bar{P}_N(\theta_\circ)x = \mathbf{tr}\left(\left[\sum_i [x]_i[Q^{-1}(\theta_\circ)\partial_{\theta_i}Q(\theta_\circ)]_{ii}\right]^2\right) \geq 0.\ \forall\, x \in \mathbb{R}^{d_y}\setminus 0$.

In the remaining part of the paper, we introduce a systematic way of constructing efficient estimators of θ using the estimating functions approach. We are interested in cases where the user is unable or unwilling to compute the score function due to either the lack of a full probabilistic model or the analytic intractability of the full model. We assume that the first two or four moments of $\boldsymbol{y}$ can be computed using the given parametric model.

5 The Estimating Functions Approach

The main idea of the estimating functions approach is to replace the unknown or analytically intractable score function $\boldsymbol{S}_N(\theta)$ by a *user-defined* function $\boldsymbol{G}_N(\theta) = G_N(\boldsymbol{Y};\theta)$, known as an estimating function (EF), that is in some sense "close" to the score function and therefore keep some desirable properties of maximum likelihood estimation. When the EF satisfies a certain optimality criterion, it is referred to as a *quasi-score function* (see Definition 13).

Definition 6 (Estimating function (EF)) *A function $G_N : \mathbb{R}^{d_y N} \times \Theta \to \mathbb{R}^{d_\theta}$ is called an EF if $G_N(\cdot;\theta)$ is $\mathbb{P}_\theta$-measurable for every $\theta \in \Theta$, and $G_N(Y;\cdot)$ is a well-defined function over Θ for every $Y \in \mathbb{R}^{d_y N}$.*

A point estimator may then be derived by equating $\boldsymbol{G}_N(\theta)$ to zero and solving for θ, namely

$$\hat{\boldsymbol{\theta}}_N \in \operatorname*{sol}_{\theta\in\Theta} \left[\boldsymbol{G}_N(\theta) = 0\right].$$

That is, the estimator is given, when it exists, as a root of the EF. Note that, by the definition of EFs, $\hat{\boldsymbol{\theta}}_N$ is a well-defined estimator and that, in general, there might be multiple roots or no roots at all; this depends on the data, the model, and the EF.

To clarify the idea of the EF approach, we have the following two examples.

Example 7 (EF corresponding to a PEM) *Let $\hat{\boldsymbol{y}}_{t|t-1}(\theta)$ be the OL-predictor of $\boldsymbol{y}_t$. Then the OL-QPEM estimator is defined as*

$$\hat{\boldsymbol{\theta}}_N := \arg\min_{\theta\in\Theta} \sum_{t=1}^{N} \frac{1}{2}\|\boldsymbol{\varepsilon}_t(\theta)\|^2, \tag{22}$$

where $\boldsymbol{\varepsilon}_t(\theta) = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}(\theta)$ is the linear innovation process. Assuming that $\varepsilon_t(\theta)$ is a smooth function of θ for every t and that Θ is compact, the estimator in (22) is equivalently given as a root of the EF

$$\begin{aligned}\boldsymbol{G}_N(\theta) &= \partial_\theta \sum_{t=1}^{N} \frac{1}{2}\|\boldsymbol{\varepsilon}_t(\theta)\|^2\\ &= -\sum_{t=1}^{N} [\partial_\theta \hat{\boldsymbol{y}}_{t|t-1}(\theta)]^\top \left(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}(\theta)\right).\end{aligned}$$

Note that $\boldsymbol{G}_N$ coincides with the score function $\boldsymbol{S}_N(\theta)$ if and only if $\theta_\circ \in \Theta$ and $\boldsymbol{\varepsilon}_t(\theta) \overset{i.i.d}{\sim} \mathcal{N}(0, I_{d_y})$ when $\theta_\circ = \theta$.

The above example shows that the PEM and the ML method may be formulated as EF methods simply by using the gradient vector of the criterion function as an EF. However, the main idea here is that *EFs do not have to be associated to any optimization-based method*; they may well be directly designed as functions of the data and the parameter (see Definition 6). Hence, the main focus of the EFs approach is on the choice of families of EFs and the characterization of the "best" EF within a given family. As will become evident below, shifting the focus from selecting/weighting the criterion function in the PEMs to selecting EFs comes with a few advantages.

Example 8 (EFs do not necessarily correspond to ML or PE methods) *Assume that the linear innovation process in Example 7 is non-stationary with covariance matrices $\Lambda_t(\theta)$ that share parameters with $\hat{\boldsymbol{y}}_{t|t-1}(\theta)$. In this case, in order to improve the asymptotic properties, a weighted PEM estimator, defined as*

$$\hat{\boldsymbol{\theta}}_N := \arg\min_{\theta\in\Theta} \sum_{t=1}^{N} \boldsymbol{\varepsilon}_t(\theta)^\top \Lambda_t(\theta)^{-1} \boldsymbol{\varepsilon}_t(\theta),$$

may be proposed. However, it can be easily shown that such an estimator cannot be consistent. Adding the term $\log\det\Lambda_t(\theta)$ to the summand of the criterion function leads to the OL-GPEM estimator. However, as shown in the previous section, it is not necessarily better than the optimal OL-WQPEM estimator, and it is not clear which one should be preferred. Moreover, the optimal OL-WQPEM estimator cannot be directly realized.

On the other hand, as shown in the coming sections, the estimator $\hat{\boldsymbol{\theta}}_N$ obtained as a root of the EF

$$\boldsymbol{G}_N(\theta) = -2\sum_{t=1}^{N} [\partial_\theta \hat{\boldsymbol{y}}_{t|t-1}(\theta)]^\top\ \Lambda_t^{-1}(\theta)(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}(\theta)),$$

under standard regularity and identifiability assumptions, is consistent and asymptotically normal, namely

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

with the least normalizing factors P_N when only the knowledge of the first two moments of the model is used (see Section 6.1). Notice that this EF may not correspond to any PE criterion function or any likelihood function, even when it is the derivative of some function over Θ. Hence, using the same partial probabilistic model, the EF approach can provide a larger class of consistent and asymptotically normal estimators compared to PEMs.

5.1 Optimal Estimating Functions

A desired property of EFs is *unbiasedness*, which permits a type of likelihood inference.

Definition 9 (Unbiased EF) *An estimating function $\boldsymbol{G}_N$ is called unbiased if $\mathbb{E}[\boldsymbol{G}_N(\theta);\theta] = 0\ \forall\theta \in \Theta$ in which the expectation is with respect to $\mathbb{P}_\theta$.*

Estimators defined as roots of unbiased EFs are invariant under smooth bijections, but they are not necessarily unbiased. The unbiasedness of an EF together with some regularity and identifiability assumptions implies the consistency of the associated estimator (see Sections 7 and 8).

The theory of optimal EFs is used to construct asymptotically optimal estimators within a given class of estimators. It is developed for regular families of EFs according to the following definition.

Definition 10 (Regular EFs) *An EF $\boldsymbol{G}_N$ is called regular if it is unbiased, $\boldsymbol{G}_N(\theta) \in \mathsf{L}_2^{d_\theta}\ \forall\theta \in \Theta$, there exists $\bar{N} \in \mathbb{N}$ such that it is differentiable in θ for all Y over an open neighborhood of Θ, and the $d_\theta \times d_\theta$ matrices $\mathbb{E}[\partial_\theta \boldsymbol{G}_N(\theta);\theta]$ and $\mathbb{E}[\boldsymbol{G}_N(\theta)\boldsymbol{G}_N^\top(\theta);\theta]$ are positive definite for all $\theta \in \Theta$, when $N > \bar{N}$. A regular family $\mathcal{G}_N$ of EFs is a set of regular EFs.*

Now notice that for any given EF $\tilde{\boldsymbol{G}}_N$ it is always possible to construct a new EF of the form

$$\boldsymbol{G}_N(\theta) = M(\theta)\tilde{\boldsymbol{G}}_N(\theta), \qquad \forall\theta \in \Theta,$$

in which $M(\theta)$ is a bounded $d_\theta \times d_\theta$ positive definite matrix that is independent of $\boldsymbol{Y}$ (i.e., M is a constant in $\mathsf{L}_2^{d_\theta}$). In this case, $\boldsymbol{G}_N(\theta)$ and $\tilde{\boldsymbol{G}}_N(\theta)$ have exactly the same roots and therefore are called equivalent. However, the covariance matrices of two equivalent EFs may differ. In order to relate the properties of the EFs to the estimators they define, a standardization is required. We will adapt the following standardization, similar to [32].

Definition 11 (Standardized EF) *The standardized EF corresponding to a regular EF $\boldsymbol{G}_N$ is defined as*

$$\bar{\boldsymbol{G}}_N(\theta) := -\mathbb{E}[\partial_\theta \boldsymbol{G}_N^\top(\theta);\theta]\mathbb{E}[\boldsymbol{G}_N(\theta)\boldsymbol{G}_N^\top(\theta);\theta]^{-1}\ \boldsymbol{G}_N(\theta).$$

It follows directly from the definition that the covariance matrices of equivalent standardized EFs are always equal and are given by the Godambe information matrix.

Definition 12 (Godambe information matrix) *The Godambe information matrix of a regular EF $\boldsymbol{G}_N$ is defined as $\mathbb{E}[\bar{\boldsymbol{G}}_N(\theta)\bar{\boldsymbol{G}}_N^\top(\theta);\theta]$, or equivalently as*

$$\mathbb{E}[\partial_\theta \boldsymbol{G}_N^\top(\theta);\theta]\mathbb{E}[\boldsymbol{G}_N(\theta)\boldsymbol{G}_N^\top(\theta);\theta]^{-1}\mathbb{E}[\partial_\theta \boldsymbol{G}_N;\theta].$$

The importance of the Godambe information matrix stems from its relation to the asymptotic accuracy of the resulting estimator. It can be seen as a generalization of the Fisher information matrix – they are in fact equal if $\boldsymbol{G}_N(\theta)$ is equivalent to the score function. To further clarify this, suppose that $\theta_\circ \in \Theta$ and assume that $\hat{\boldsymbol{\theta}}_N$ is a root of $\bar{\boldsymbol{G}}_N(\theta)$ that converges almost surely to $\theta_\circ$. Then, under appropriate conditions (see Theorem 31), as $N \to \infty$ it holds that

$$\sqrt{N}P_N^{-\frac{1}{2}}(\hat{\boldsymbol{\theta}}_N - \theta_\circ) \rightsquigarrow \mathcal{N}(0, I_{d_\theta})$$

where $P_N = \mathbb{E}[N^{-1}\bar{\boldsymbol{G}}_N(\theta)\bar{\boldsymbol{G}}_N^\top(\theta);\theta]^{-1}$ is the inverse of the normalized Godambe information matrix.

Therefore, an *optimal EF* in a regular family $\mathcal{G}_N$ of EFs, *leading to optimal estimators among those defined using EFs in $\mathcal{G}_N$*, may be defined as the EF in $\mathcal{G}_N$ *with the largest Godambe information matrix* in the partial ordering of positive semidefinite matrices. This optimality notion is known as Godambe (Godambe-Durbin) optimality (see [14,22,23,25]) and can be seen as a generalization of the optimality criterion of unbiased estimators in the Gauss-Markov theorem ([40, Theorem 4.12]) to unbiased EFs. Notice that it is a *finite sample criterion* on the EF which is *tied to the asymptotic optimality of the resulting estimator.*

Definition 13 (Optimal EF) *Let $\mathcal{G}_N$ be a regular family of EFs. An EF $\boldsymbol{G}_N^\star$ is called a quasi-score function in $\mathcal{G}_N$, if $\boldsymbol{G}_N^\star \in \mathcal{G}_N$ and the matrix*

$$\mathbb{E}[\bar{\boldsymbol{G}}_N^\star(\theta)[\bar{\boldsymbol{G}}_N^\star(\theta)]^\top;\theta] - \mathbb{E}[\bar{\boldsymbol{G}}_N(\theta)\bar{\boldsymbol{G}}_N^\top(\theta);\theta] \succeq 0$$

for all $\theta \in \Theta$ and all $\boldsymbol{G}_N \in \mathcal{G}_N$.

We will refer to any EF satisfying the above definition as the quasi-score function in $\mathcal{G}_N$. An estimator based on the quasi-score function in $\mathcal{G}_N$ may now be defined.

Definition 14 (The $\mathcal{G}_N$-optimal estimator) *The $\mathcal{G}_N$-optimal estimator is defined, when it exists, as*

$$\hat{\boldsymbol{\theta}}_N \in \underset{\theta\in\Theta}{\operatorname{sol}}\left[\boldsymbol{G}_N^\star(\theta) = 0\right], \tag{23}$$

where $\boldsymbol{G}_N^\star$ is the quasi-score function in $\mathcal{G}_N$.

It has been shown in [25] (under a mild regularity condition) that every quasi-score function $\boldsymbol{G}_N^\star$ in a regular family $\mathcal{G}_N$ of EFs is such that

$$\begin{aligned}&\mathbb{E}[(\boldsymbol{S}_N(\theta) - \bar{\boldsymbol{G}}_N(\theta))(\boldsymbol{S}_N(\theta) - \bar{\boldsymbol{G}}_N(\theta))^\top]\\ &\qquad - \mathbb{E}[(\boldsymbol{S}_N(\theta) - \bar{\boldsymbol{G}}_N^\star(\theta))(\boldsymbol{S}_N(\theta) - \bar{\boldsymbol{G}}_N^\star(\theta))^\top] \succeq 0\end{aligned}$$

for all $\boldsymbol{G}_N \in \mathcal{G}_N$. Hence, the quasi-score function in $\mathcal{G}_N$ is the element in $\mathcal{G}_N$ *minimizing the dispersion distance to the score function.* If $\mathcal{G}_N$ is large enough to contain the score function, the quasi-score function is then equivalent to the score function; namely $\bar{\boldsymbol{G}}_N^\star(\theta) = \boldsymbol{S}_N(\theta)$. From the above, a geometric interpretation of the Godambe optimality notion can be given in the Hilbert space $\mathsf{L}_2^{d_\theta}$: for every θ, the quasi-score function in $\mathcal{G}_N$ is the orthogonal projection of the score function onto $\mathcal{G}_N$ (see [34] and [32]).

The next result, due to Heyde in [31], gives a very useful characterization of quasi-score functions.

Lemma 15 (Characterization of optimal EFs) *Suppose that the regular family $\mathcal{G}_N$ of EFs is closed under addition. Then $\boldsymbol{G}_N^\star$ is a quasi-score function in $\mathcal{G}_N$ if and only if $\boldsymbol{G}_N^\star \in \mathcal{G}_N$ and*

$$\begin{aligned}&\mathbb{E}[\partial_\theta \boldsymbol{G}_N(\theta);\theta]^{-1}\mathbb{E}[\boldsymbol{G}_N(\theta)[\boldsymbol{G}_N^\star(\theta)]^\top;\theta]\\ &\quad= \mathbb{E}[\partial_\theta \boldsymbol{G}_N^\star(\theta);\theta]^{-1}\mathbb{E}[\boldsymbol{G}_N^\star(\theta)[\boldsymbol{G}_N^\star(\theta)]^\top;\theta]\end{aligned}$$

for all $\boldsymbol{G}_N \in \mathcal{G}_N$, or equivalently

$$\mathbb{E}[\partial_\theta \boldsymbol{G}_N(\theta);\theta]^{-1}\mathbb{E}[\boldsymbol{G}_N(\theta)[\boldsymbol{G}_N^\star(\theta)]^\top;\theta] \qquad (24)$$

is a constant matrix for all $\boldsymbol{G}_N \in \mathcal{G}_N$.

PROOF. See [31, Theorem 1] or [32, Theorem 2.1].

Lemma 15 provides an easy way of examining the optimality of EFs. However, finding a quasi-score function in a general regular family $\mathcal{G}_N$ of EFs is not an easy task. Fortunately, restricting $\mathcal{G}_N$ to EFs with a specific form usually makes it possible to compute the quasi-score function. In the following section, we consider regular families of EFs that are linear and quadratic in the data (or equivalently in the linear innovations). We then make use of Lemma 15 to find the quasi-score functions. Specific forms of these two families were considered by several authors in the statistics literature, in particular when the EFs have independent differences; see, for example, [11,18,12,25,26,67].

6 Linear and Quadratic EFs

We start by considering a family of linear EFs.

6.1 Linear regular EFs

Define the linear regular family of EFs as

$$\begin{aligned}\mathcal{G}_N^L := \Big\{ &\boldsymbol{G}_N : \boldsymbol{G}_N \text{ regular}, \boldsymbol{G}_N(\theta) := \sum_{t=1}^N a_t(\theta)\boldsymbol{e}_t(\theta),\\ &\boldsymbol{e}_t(\theta) := \boldsymbol{y}_t - \mu_t(\theta),\ a_t(\theta) \in \mathbb{R}^{d_\theta \times d_y}\ \forall t, \theta \Big\}.\end{aligned}$$

The EFs in $\mathcal{G}_N^L$ are *linear in the outputs* $\boldsymbol{y}_t$, $\mu_t(\theta) := \mathbb{E}[\boldsymbol{y}_t;\theta]$ is the OE-predictor, and for all θ and t, $a_t(\theta)$ is a $d_\theta \times d_y$ deterministic matrix with properties ensuring that $\boldsymbol{G}_N$ is regular. Different EFs in $\mathcal{G}_N^L$ are obtained by varying $a_t(\theta)$ as functions of θ. Note that these EFs arise naturally in PEMs. For instance, if there exists a unique root of the EF defined by taking $a_t(\theta) = \partial_\theta \mu_t^\top(\theta)$, then the obtained estimator is in fact the OE-QPEM estimator. However, such an EF is not a quasi-score function in $\mathcal{G}_N^L$ as we will see. It is of interest to note that every EF in $\mathcal{G}_N^L$ can be written in the form

$$\tilde{\boldsymbol{G}}_N(\theta) = \sum_{t=1}^N \tilde{a}_t(\theta)(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}(\theta)) = \sum_{t=1}^N \underbrace{\tilde{a}_t(\theta)\boldsymbol{\varepsilon}_t(\theta)}_{=:\boldsymbol{g}_t(\theta)} \qquad (25)$$

where $\boldsymbol{\varepsilon}(\theta)$ is the linear innovation process. This form becomes convenient when analyzing the asymptotic properties of the EFs and the corresponding estimators. We formalize this remark in the following proposition.

Proposition 16 *An EF of the form (25), where $\hat{\boldsymbol{y}}_{t|t-1}(\theta)$ is the OL-predictor of $\boldsymbol{y}_t$, and $\tilde{a}_t(\theta) \in \mathbb{R}^{d_\theta \times d_y}$ for all $\theta \in \Theta$ such that the EF is regular, belongs to $\mathcal{G}_N^L$. Furthermore, varying $\tilde{a}_t(\theta)$ over $\mathbb{R}^{d_\theta \times d_y}$-valued functions of θ, such that the EF is regular, spans $\mathcal{G}_N^L$.*

PROOF. Use (10) in (25) to see that

$$\begin{aligned}\tilde{\boldsymbol{G}}_N(\theta) &= [\tilde{a}_1(\theta)\ldots\tilde{a}_N(\theta)]L_N^{-1}(\theta)(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y};\theta])\\ &= \sum_{t=1}^N a_t(\theta)(\boldsymbol{y}_t - \mu_t(\theta))\end{aligned}$$

where $[a_1(\theta)\ldots a_N(\theta)] = [\tilde{a}_1(\theta)\ldots\tilde{a}_N(\theta)]L_N^{-1}(\theta)$, and $L_N^{-1}(\theta)$ is nonsingular.

Thus, in $\mathcal{G}_N^L$, using the OE-predictor is equivalent to using the OL-predictor, because the resulting functions are related by a non-singular linear transformation.

The quasi-score function in $\mathcal{G}_N^L$

The following theorem shows that the quasi-score function in $\mathcal{G}_N^L$ is determined by the first and second moments of $\boldsymbol{y}$.

Theorem 17 (The quasi-score function in $\mathcal{G}_N^L$) *Let $\boldsymbol{G}_N^\star$ denote the quasi-score function in $\mathcal{G}_N^L$. Then*

$$\boldsymbol{G}_N^\star(\theta) = -\partial_\theta \mu^\top(\theta)\Sigma^{-1}(\theta)\boldsymbol{E}(\theta) \qquad (26)$$

where $\mu(\theta)$, $\boldsymbol{E}(\theta)$ and $\Sigma(\theta)$ are defined in (8) and (9).

PROOF. Let $G_N \in \mathcal{G}_N^L$ and define the $d_\theta \times d_y N$ matrix $A(\theta) := [a_1(\theta)\ldots a_N(\theta)]$ so that $G_N(\theta) = A(\theta)\boldsymbol{E}(\theta)$. Now observe that, because $A(\theta)$ is data-independent and $\mathbb{E}[\boldsymbol{E}(\theta);\theta] = 0$ by definition, the linearity of the expectation operator and the product rule shows that $\mathbb{E}[\partial_\theta \boldsymbol{G}_N(\theta);\theta] = -A(\theta)\partial_\theta \mu(\theta)$. Moreover, it holds that

$$\begin{aligned}\mathbb{E}[\boldsymbol{G}_N(\theta)[\boldsymbol{G}_N^\star(\theta)]^\top;\theta] &= \mathbb{E}[A(\theta)\boldsymbol{E}(\theta)\boldsymbol{E}^\top(\theta)[A^\star(\theta)]^\top]\\ &= A(\theta)\Sigma(\theta)[A^\star(\theta)]^\top.\end{aligned}$$

After noting that $\mathcal{G}_N^L$ is closed under addition, the necessary and sufficient condition (24) of Lemma 15 implies that $[A^\star(\theta)]^\top = -\Sigma^{-1}(\theta)\partial_\theta \mu(\theta)$ and hence (26).

The $\mathcal{G}_N^L$-optimal estimator is then defined as a root of the quasi-score function in (26). Notice that, while only the first moment of $\boldsymbol{y}$ is required to define an EF in $\mathcal{G}_N^L$, the second moments of $\boldsymbol{y}$ are required to compute the quasi-score function. When the EFs are quadratic in the data, higher moments will be required.

6.2 Quadratic regular EFs

Define the family of quadratic regular EFs as

$$\begin{aligned}\mathcal{G}_N^Q := \Bigg\{ & \boldsymbol{G}_N : \boldsymbol{G}_N \text{ regular}, \boldsymbol{G}_N(\theta) := \sum_{t=1}^{N} \boldsymbol{g}_t(\theta), \\ & \boldsymbol{g}_t(\theta) := a_t(\theta)\boldsymbol{e}_t(\theta) + \\ & \sum_{s=1}^{t}\sum_{i=1}^{d_y}\sum_{j=1}^{i} b_{t_i s_j}(\theta)\left([\boldsymbol{e}_t(\theta)]_i\,[\boldsymbol{e}_s(\theta)]_j - [\sigma_{ts}(\theta)]_{ij}\right), \quad (27) \\ & \boldsymbol{e}_t(\theta) := \boldsymbol{y}_t - \mathbb{E}[\boldsymbol{y}_t(\theta)],\ a_t(\theta) \in \mathbb{R}^{d_\theta \times d_y} \\ & b_{t_i s_j}(\theta) \in \mathbb{R}^{d_\theta}\quad \forall t, \theta \Bigg\},\end{aligned}$$

where we used the notation $[\boldsymbol{e}_t(\theta)]_i$ to denote the ith-entry of $\boldsymbol{e}_t(\theta)$, and $\sigma_{ts}(\theta)$ to denote the covariance matrix $\mathbf{cov}(\boldsymbol{y}_t, \boldsymbol{y}_s; \theta)$. Notice that these EFs arise naturally in PEMs. For example, the derivative of the OL-GPEM criterion function (12) belongs to $\mathcal{G}_N^Q$; however, as we will see, it is not necessarily a quasi-score function in $\mathcal{G}_N^Q$.

Every $\boldsymbol{G}_N \in \mathcal{G}_N^Q$ can be written in vector form as $\boldsymbol{G}_N(\theta) = \tilde{A}(\theta)\boldsymbol{E}(\theta) + \tilde{B}(\theta)\tilde{\boldsymbol{V}}(\theta)$ where $\tilde{A}(\theta)$ is a $d_\theta \times d_y N$ matrix, $\tilde{B}(\theta)$ is a $d_\theta \times d_y N(d_y N + 1)/2$ matrix, and $\tilde{\boldsymbol{V}}(\theta) := \mathbf{vech}[\boldsymbol{E}(\theta)\boldsymbol{E}^\top(\theta) - \Sigma(\theta)]$ is a random vector of dimension $d_y N(d_y N+1)/2$, where $\mathbf{vech}[\cdot]$ is the half-vectorization operator[4]. Moreover, it is possible to write every EF in $\mathcal{G}_N^Q$ in terms of the linear innovations, just as with $\mathcal{G}_N^L$. The following proposition is similar to Proposition 16.

Proposition 18 *Every EF $\boldsymbol{G}_N \in \mathcal{G}_N^Q$ may be expressed in terms of the linear innovations as follows*

$$\boldsymbol{G}_N(\theta) = M(\theta)\boldsymbol{H}(\theta), \tag{28}$$

where $M(\theta) := [A(\theta)\ \ B(\theta)]$ in which $A(\theta)$ is a $d_\theta \times d_y N$ matrix, $B(\theta)$ is a $d_\theta \times d_y N(d_y N+1)/2$ matrix,

$$\begin{aligned}\boldsymbol{H}(\theta) &:= [\boldsymbol{\mathcal{E}}^\top(\theta)\ \ \boldsymbol{V}^\top(\theta)]^\top, \\ \boldsymbol{V}(\theta) &:= \mathbf{vech}[\boldsymbol{\mathcal{E}}(\theta)\boldsymbol{\mathcal{E}}^\top(\theta) - \Lambda(\theta)],\end{aligned} \tag{29}$$

in which $\Lambda(\theta) := \mathbf{cov}(\boldsymbol{\mathcal{E}}(\theta), \boldsymbol{\mathcal{E}}(\theta); \theta)$.

PROOF. Use (29) and (10) in (28) to see that

$$\begin{aligned}\boldsymbol{G}_N(\theta) = {} & A(\theta)L^{-1}(\theta)\boldsymbol{E}(\theta) + \\ & B(\theta)\mathbf{vech}[L^{-1}(\theta)(\boldsymbol{E}(\theta)\boldsymbol{E}^\top(\theta) - \Sigma(\theta))L^{-\top}(\theta)].\end{aligned}$$

Using the properties of the $\mathbf{vech}$ operator, it holds that

$$\begin{aligned}\mathbf{vech}[L^{-1}(\theta)(\boldsymbol{E}(\theta)\boldsymbol{E}^\top(\theta) - \Sigma(\theta))L^{-\top}(\theta)] = {} & \\ C(\theta)\mathbf{vech}[(\boldsymbol{E}(\theta)\boldsymbol{E}^\top(\theta) - \Sigma(\theta))] &\end{aligned}$$

in which $C(\theta) \succ 0\ \forall\theta$ (because $L(\theta)$ is; see [30, Section 16.4 d, pages 361 and 362]). The proof is then completed by taking $\tilde{A}(\theta) = A(\theta)L^{-1}(\theta)$ and $\tilde{B}(\theta) = B(\theta)C(\theta)$.

Notice that (28) can be expanded as follows

$$\begin{aligned}\boldsymbol{G}_N(\theta) = \sum_{t=1}^{N} & a_t(\theta)\boldsymbol{\varepsilon}_t + b_t(\theta)(\boldsymbol{\varepsilon}_t(\theta)\boldsymbol{\varepsilon}_t^\top(\theta) - \Lambda_t(\theta)) \\ & + \sum_{s<t} c_{t,s}(\theta)\boldsymbol{\varepsilon}_t(\theta)\boldsymbol{\varepsilon}_s(\theta)^\top\end{aligned} \tag{30}$$

where cross-products of linear innovations appear.

The quasi-score function in $\mathcal{G}_N^Q$

We now use the necessary and sufficient conditions of Lemma 15 to find the quasi-score function in $\mathcal{G}_N^Q$. The following theorem shows that it is determined by the first four moments of $\boldsymbol{y}$.

Theorem 19 (The quasi-score function in $\mathcal{G}_N^Q$) *Let $\boldsymbol{G}_N^\star$ denote the quasi-score function in $\mathcal{G}_N^Q$. Then*

$$\boldsymbol{G}_N^\star(\theta) = \mathbb{E}[\partial_\theta \boldsymbol{H}^\top(\theta);\theta]\mathbb{E}[\boldsymbol{H}(\theta)\boldsymbol{H}^\top(\theta);\theta]^{-1}\boldsymbol{H}(\theta) \tag{31}$$

where $\boldsymbol{H}(\theta)$ is defined in (29).

PROOF. The proof is similar to that of Theorem 17. Because $\mathbb{E}[\partial_\theta \boldsymbol{G}_N(\theta);\theta] = M(\theta)\mathbb{E}[\partial_\theta \boldsymbol{H}(\theta);\theta]$ and $\mathbb{E}[\boldsymbol{G}_N(\theta)[\boldsymbol{G}_N^\star(\theta)]^\top;\theta] = M(\theta)\mathbb{E}[\boldsymbol{H}(\theta)\boldsymbol{H}^\top(\theta);\theta][M^\star(\theta)]^\top$, the necessary and sufficient condition of Lemma 15 is satisfied when $M^\star(\theta) = \mathbb{E}[\partial_\theta \boldsymbol{H}^\top(\theta);\theta]\mathbb{E}[\boldsymbol{H}(\theta)\boldsymbol{H}^\top(\theta);\theta]^{-1}$, which proves (31).

Remark 20 *If $\boldsymbol{y}$ is a Gaussian process, the score function $\boldsymbol{S}_N \in \mathcal{G}_N^Q$ is equivalent to the quasi-score function in $\mathcal{G}_N^Q$. In this case, the linear innovation process $\boldsymbol{\varepsilon}$ is an independent process and the quasi-score function in $\mathcal{G}_N^Q$ given in terms of $\boldsymbol{\varepsilon}$ does not contain any cross-products; i.e, the factors $c_{t,s}$ in (30) are identically zero. In general however, these factors will be different from zero.*

An estimator based on the quasi-score function in (31) may be defined; however, when N is large, such an estimator would require computing and inverting matrices of relatively large dimensions. It is of interest in this case to observe that if the distribution of $\boldsymbol{\mathcal{E}}(\theta)$ is symmetric around its mean value, $\mathbb{E}[\boldsymbol{\mathcal{E}}(\theta)\boldsymbol{V}^\top(\theta);\theta] = 0$ for every θ, and therefore the quasi-score function reduces to

$$\boldsymbol{G}_N^\star(\theta) = \tilde{\boldsymbol{G}}_N^\star(\theta) + \mathbb{E}[\partial_\theta \boldsymbol{V}(\theta)]^\top\mathbb{E}[\boldsymbol{V}(\theta)\boldsymbol{V}^\top(\theta);\theta]^{-1}\boldsymbol{V}(\theta) \tag{32}$$

where $\tilde{\boldsymbol{G}}^\star(\theta)$ is the quasi-score function in $\mathcal{G}_N^L$. To simplify the computations, one may be tempted to use (32) even when the third moments are not zero. However, if the symmetry assumption is not maintained, the combination (32) will not be optimal in $\mathcal{G}_N^Q$; what is more,

[4] The half-vectorization of a symmetric matrix M of size n is the column vector of size $n(n+1)/2$ obtained by vectorizing the lower triangular part of M.

there will not be any guarantee that the Godambe information matrix of the EF in (32) will be larger than that of the quasi-score function in $\mathcal{G}_N^L$. This is due to the fact that, if the two terms on the right-hand side of (32) are correlated, the contribution due to the second term to the Godambe information matrix of the quasi-score function in $\mathcal{G}_N^L$ may be negative. [5] This issue is related to the general question of how to combine two (or more) EFs that may not necessarily belong to nested subspaces. For example, it may be of interest to use EFs that have different forms or that are constructed based on two different experiments (thus, different data sets). A solution is obtained by orthogonalization, as we show in the following theorem.

Theorem 21 (Optimal combinations of EFs) *Let $\bar{\boldsymbol{G}}_N^\star$ be the standardized quasi-score function in a regular family $\mathcal{G}_N$ of EFs and consider a regular EF $\tilde{\boldsymbol{G}}_N$. Define the regular family $\mathcal{G}_N^C$ of EFs as the set of all combination EFs $\boldsymbol{J}_N(\theta) := M_1(\theta)\bar{\boldsymbol{G}}_N^\star(\theta) + M_2(\theta)\tilde{\boldsymbol{G}}_N(\theta)$, where $M_1(\theta)$ and $M_2(\theta)$ are $d_\theta \times d_\theta$ data-independent matrices such that $\boldsymbol{J}_N$ is regular. Let $\boldsymbol{J}_N^\star$ denote the quasi-score function in $\mathcal{G}_N^C$. Then*

$$\begin{aligned}\boldsymbol{J}_N^\star(\theta) := \bar{\boldsymbol{G}}_N^\star(\theta) - \\ \mathbb{E}[\partial_\theta \boldsymbol{H}^\top(\theta);\theta]\mathbb{E}[\boldsymbol{H}(\theta)\boldsymbol{H}^\top(\theta);\theta]^{-1}\boldsymbol{H}(\theta),\end{aligned} \tag{33}$$

where

$$\begin{aligned}\boldsymbol{H}(\theta) := \tilde{\boldsymbol{G}}_N(\theta) - \\ \mathbb{E}[\tilde{\boldsymbol{G}}_N(\theta)[\bar{\boldsymbol{G}}_N^\star(\theta)]^\top;\theta]\mathbb{E}[\bar{\boldsymbol{G}}_N^\star(\theta)[\bar{\boldsymbol{G}}_N^\star(\theta)]^\top;\theta]^{-1}\bar{\boldsymbol{G}}_N^\star(\theta),\end{aligned}$$

and $\mathbb{E}[\bar{\boldsymbol{J}}_N^\star(\theta)[\bar{\boldsymbol{J}}_N^\star(\theta)]^\top;\theta] - \mathbb{E}[\bar{\boldsymbol{G}}_N^\star(\theta)[\bar{\boldsymbol{G}}_N^\star(\theta)]^\top;\theta] \succeq 0$; In other words, the Godambe information matrix of the combination in (33) is at least as large as that of $\boldsymbol{G}_N^\star$.

PROOF. The proof follows that of Theorems 17 and 19. Define the matrices $M(\theta) := [M_1(\theta)\ \ M_2(\theta)]$ and $\boldsymbol{K}(\theta) := [[\bar{\boldsymbol{G}}_N^\star(\theta)]^\top\ \ \boldsymbol{H}^\top(\theta)]^\top$, then observe that $\mathbb{E}[\boldsymbol{J}_N(\theta)[\boldsymbol{J}_N^\star(\theta)]^\top;\theta] = M(\theta)\mathbb{E}[\boldsymbol{K}(\theta)\boldsymbol{K}^\top(\theta);\theta][M^\star(\theta)]^\top$, $\mathbb{E}[\partial_\theta \boldsymbol{J}_N(\theta);\theta] = M(\theta)\mathbb{E}[\partial_\theta \boldsymbol{K}(\theta);\theta]$ where

$$\mathbb{E}[\partial_\theta \boldsymbol{K}(\theta);\theta] = \Big[\mathbb{E}[\partial_\theta \bar{\boldsymbol{G}}_N^\star(\theta);\theta]^\top\ \ \mathbb{E}[\partial_\theta \boldsymbol{H}(\theta);\theta]^\top\Big]^\top$$

$$\mathbb{E}[\boldsymbol{K}(\theta)\boldsymbol{K}^\top(\theta);\theta] = \begin{bmatrix}\mathbb{E}[\bar{\boldsymbol{G}}_N^\star(\theta)[\bar{\boldsymbol{G}}_N^\star(\theta)]^\top;\theta] & 0 \\ 0 & \mathbb{E}[\boldsymbol{H}(\theta)\boldsymbol{H}^\top(\theta);\theta]\end{bmatrix}.$$

The last equality follows because, by definition, $\boldsymbol{H}_N$ is orthogonal to $\bar{\boldsymbol{G}}_N^\star$. Since, by construction, $\mathcal{G}_N^C$ is closed under addition and $\mathbb{E}[\partial_\theta \bar{\boldsymbol{G}}_N^\star(\theta);\theta]^\top = -\mathbb{E}[\bar{\boldsymbol{G}}_N^\star(\theta)[\bar{\boldsymbol{G}}_N^\star(\theta)]^\top;\theta]$, Lemma 15 implies that

$$\begin{aligned}M^\star(\theta) &= \mathbb{E}[\partial_\theta \boldsymbol{K}^\top(\theta);\theta]\mathbb{E}[\boldsymbol{K}(\theta)\boldsymbol{K}^\top(\theta);\theta]^{-1} \\ &= \Big[-1\ \ \mathbb{E}[\partial_\theta \boldsymbol{H}^\top(\theta);\theta]\mathbb{E}[\boldsymbol{H}(\theta)\boldsymbol{H}^\top(\theta);\theta]^{-1}\Big]\end{aligned}$$

which proves (33). The last part is established by noticing that $\bar{\boldsymbol{G}}_N^\star \in \mathcal{G}_N^C$.

The above discussion shows that, when considering a quadratic EF, third and fourth order cross-moments have to be used to correctly weight the contribution from the linear and nonlinear parts. The use of marginal moments when the data are correlated is not sufficient to guarantee improvement over the quasi-score function in $\mathcal{G}_N^L$. Nevertheless, as we now show, not all the cross-moments are required. The idea is simply to use only the diagonal blocks of $\boldsymbol{E}(\theta)\boldsymbol{E}^\top(\theta)$ to define the quadratic part of the EF. Let

$$\begin{aligned}\boldsymbol{V}_d(\theta) &:= [\boldsymbol{v}_1^\top(\theta)\ldots\boldsymbol{v}_N^\top(\theta)]^\top \\ \boldsymbol{v}_t(\theta) &:= \mathbf{vech}[\boldsymbol{e}_t(\theta)\boldsymbol{e}_t^\top(\theta) - \sigma_{tt}(\theta)],\end{aligned} \tag{34}$$

where we used the same notations introduced in (27), and define the family of regular EFs

$$\begin{aligned}\mathcal{G}_N^q := \big\{&\boldsymbol{G}_N : \boldsymbol{G}_N \text{ regular},\ \boldsymbol{G}_N(\theta) := A(\theta)\boldsymbol{E}(\theta) + B(\theta)\boldsymbol{V}_d(\theta), \\ &A(\theta), B(\theta) \in \mathbb{R}^{d_\theta \times d_y N}\ \forall\theta\big\}.\end{aligned}$$

Theorem 22 (The quasi-score function in $\mathcal{G}_N^q$) *Let $\boldsymbol{G}_N^\star$ denote the quasi-score function in $\mathcal{G}_N^q$. Then*

$$\boldsymbol{G}_N^\star(\theta) = \mathbb{E}[\partial_\theta \boldsymbol{H}_d^\top(\theta);\theta]\mathbb{E}[\boldsymbol{H}_d(\theta)\boldsymbol{H}_d^\top(\theta);\theta]^{-1}\boldsymbol{H}_d(\theta) \tag{35}$$

where $\boldsymbol{H}_d(\theta) := [\boldsymbol{E}^\top(\theta)\ \boldsymbol{V}_d^\top(\theta)]^\top$, and $\boldsymbol{V}_d(\theta)$ as in (34).

PROOF. The proof is identical to that of Theorem 19.

An equivalent EF to that in (35) can be obtained using Theorem 21: let $\boldsymbol{G}_N^\star$ be the quasi-score function in $\mathcal{G}_N^L$, let $\tilde{\boldsymbol{G}}_N(\theta) := \mathbb{E}[\partial_\theta \boldsymbol{V}_d(\theta);\theta]^\top\mathbb{E}[\boldsymbol{V}_d(\theta)\boldsymbol{V}_d^\top(\theta);\theta]^{-1}\boldsymbol{V}_d(\theta)$. Then, the EF in (33) is a quasi-score function in $\mathcal{G}_N^q$. Notice that EFs in $\mathcal{G}_N^q$ may be written in terms of the linear innovations as

$$\boldsymbol{G}_N(\theta) = \sum_{t=1}^{N} \underbrace{a_t(\theta)\boldsymbol{\varepsilon}_t + b_t(\theta)(\boldsymbol{\varepsilon}_t(\theta)\boldsymbol{\varepsilon}_t^\top(\theta) - \Lambda_t(\theta))}_{=:\boldsymbol{g}_t(\theta)} \tag{36}$$

for some sequence of vectors $\{a_t(\theta)\}$ and $\{b_t(\theta)\}$, and that $\mathcal{G}_N^L \subset \mathcal{G}_N^q$ (see (25)). Hence the quasi-score function in $\mathcal{G}_N^q$ is guaranteed to have a Godambe information matrix at least as large as that of the quasi-score function in $\mathcal{G}_N^L$. We now have the following important remark.

Remark 23 *Standardized EFs satisfy the relation $\mathbb{E}[\bar{\boldsymbol{G}}_N(\theta)[\bar{\boldsymbol{G}}_N(\theta)]^\top;\theta] = -\mathbb{E}[\partial_\theta \bar{\boldsymbol{G}}_N(\theta);\theta]$ for all $\forall\theta \in \Theta$ (see Definition 11), which is a property of score functions. Moreover, the standardized quasi-score functions in $\mathcal{G}_N^L$ and $\mathcal{G}_N^q$ satisfy the relation $\bar{\boldsymbol{G}}_N^\star(\theta) = -\boldsymbol{G}_N^\star(\theta)$ and the Godambe information matrix in this case is given by $\mathbb{E}[\bar{\boldsymbol{G}}_N^\star(\theta)[\bar{\boldsymbol{G}}_N^\star(\theta)]^\top;\theta] = \mathbb{E}[\boldsymbol{G}_N^\star(\theta)[\boldsymbol{G}_N^\star(\theta)]^\top;\theta]$.*

In the next two sections, we discuss the convergence and asymptotic properties of the $\mathcal{G}_N^L$-optimal and $\mathcal{G}_N^q$-optimal estimators. The analysis is performed under analogous assumptions as those used for the analysis of PEM estimators; see [43,47] and [2]. A similar analysis may be done for the $\mathcal{G}_N^Q$-optimal estimator; however,

[5] The obvious case is when $\boldsymbol{S}_N \in \mathcal{G}_N^L$; see Section 10.3.

due to the presence of the cross-products $\boldsymbol{\varepsilon}_t(\theta)\boldsymbol{\varepsilon}_s^\top(\theta)$ in the quasi-score function (see (30), Remark 20), a more stringent set of assumptions will be required. A thorough study of this case is postponed to a future contribution.

7 Convergence and Consistency

In what follows, we will only be concerned with the standardized quasi-score functions and thus, for brevity, we will drop the overbar and star superscript from the notations. The convergence of the estimators will be established under the following assumptions, similar to PEMs.

Assumption 24

(1) The data $\boldsymbol{y}$ is exponentially forgetting of order $r = 4$ (see [42,43] for a definition).
(2) Θ is compact, and the parameterization of $\mu(\theta)$ and $\Sigma(\theta)$ is two times continuously differentiable over an open neighborhood of Θ.
(3) The function $(t, \theta) \mapsto \mathbb{E}[\boldsymbol{y}_t; \theta]$ and its derivative with respect to θ are uniformly bounded in t, θ.
(4) The matrices $\Lambda_t(\theta)$ are uniformly bounded in t, θ.
(5) The entries of the matrix $L^{-1}(\theta)$ given by the $LDL^\top$ decomposition of $\Sigma(\theta)$ are uniformly exponentially decaying in t, θ.

Assumption 25 *In addition to Assumption 24, the third and fourth moments of the linear innovations $\boldsymbol{\varepsilon}_t(\theta)$ are uniformly bounded in t, θ.*

The first condition in Assumption 24 concerns the data. It is satisfied whenever the underlying system is (exponentially) stable. It is particularly easy to check for block-oriented models where all dynamical blocks are LTI; this includes stochastic Wiener models which are widely used in practice. The fifth condition in Assumption 24 is equivalent to the noise model invertibility assumption in the LTI case (see [44, Section 3.2]). For general (quasi-)stationary models, it is satisfied whenever the model's output has a strictly positive rational spectrum. In the case of non-stationary models, it needs to be checked for the used inputs; note that $L^{-1}(\theta)$ depends on u. Finally, Assumption 25 concerns the model and the used inputs. It is required only for the convergence of the $\mathcal{G}_N^q$-optimal estimator. The basic convergence result is given in the following lemma where $\mathbb{E}_\circ$ denotes the expectation with respect to the true distribution of the data.

Lemma 26 (Uniform convergence) *Suppose that Assumption 24 holds. Then as $N \to \infty$,*

$$\sup_{\theta\in\Theta} \|\boldsymbol{G}_N(\theta) - \mathbb{E}_\circ[\boldsymbol{G}_N(\theta)]\| \xrightarrow{a.s.} 0, \qquad (37)$$

and

$$\sup_{\theta\in\Theta} \|\partial_\theta\boldsymbol{G}_N(\theta) - \mathbb{E}_\circ[\partial_\theta\boldsymbol{G}_N(\theta)]\| \xrightarrow{a.s.} 0, \qquad (38)$$

where $\boldsymbol{G}_N$ is the quasi-score function in $\mathcal{G}_N^L$. In addition, the sequence $\{\mathbb{E}_\circ[\boldsymbol{G}_N(\theta)]\}$ is uniformly equicontinuous over Θ. Moreover, if Assumption 25 holds, the conclusions hold true for the quasi-score function in $\mathcal{G}_N^q$.

PROOF. The proof is constructed componentwise, i.e., for $[\partial_\theta\boldsymbol{G}_N(\theta)]_{ij}$, $i,j = 1,\dots,d_\theta$, and is analogous to that proof of [43, Lemma 3.1]: to establish (37) and (38), replace the function $l(t,\theta,\boldsymbol{\varepsilon})$ there by $[\boldsymbol{g}_t(\theta)]_{ij}$ and $[\partial_\theta\boldsymbol{g}_t(\theta)]_{ij}$ respectively, where $\boldsymbol{g}_t(\theta)$ is on the form in (25) for the $\mathcal{G}_N^L$-optimal estimator and on the form in (36) for the $\mathcal{G}_N^q$-optimal estimator. Observe that here, the used EFs are linear and quadratic in the (linear) prediction errors, and thus when Assumptions 24 and 25 hold, the conditions required by [43, Lemma 3.1] are satisfied.

Before proving the convergence of the above defined estimators, we need to establish their existence. In order to establish the existence of at least one sequence of roots, we will rely on the following assumption.

Assumption 27 *For all sufficiently large N the set $\mathcal{Z}_N := \Big\{\theta \in \Theta : \mathbb{E}_\circ[\boldsymbol{G}_N(\theta)] = 0\Big\}$ is non-empty, and for every $\theta^* \in \mathcal{Z}_N$ it holds that $\mathbb{E}_\circ[\partial_\theta\boldsymbol{G}_N(\theta^*)] \succ 0$ so that there exists $r > 0$ such that θ^* is a unique root of $\mathbb{E}_\circ[\boldsymbol{G}_N(\theta)]$ on $B_r(\theta^*) := \{\theta \in \Theta : \|\theta - \theta^*\| \leq r\}$.*

Assumption 27 requires that the expected value of $\boldsymbol{G}_N$, with respect to the true distribution of the data, has at least one root when N is sufficiently large. For EFs in $\mathcal{G}_N^L$ it means that

$$\exists\theta^* \in \Theta : \mathbb{E}_\circ[\boldsymbol{Y}] = \mu(\theta^*) \quad \text{for all large } N. \qquad (39)$$

For EFs in $\mathcal{G}_N^Q$ it means that, in addition to (39),

$$\mathbb{E}_\circ[\boldsymbol{Y}\boldsymbol{Y}^\top] = \Sigma(\theta^*) + \mu(\theta^*)\mu(\theta^*)^\top.$$

This automatically holds once $\theta_\circ \in \Theta$ and μ and Σ are correctly specified. However, in general when $\theta_\circ \notin \Theta$, the condition is not easy to analyze and the existence of a solution is not generally guaranteed; in these instances a direct check of the condition for the used model is necessary. In case several roots exist, the assumption requires that they are isolated.

The following lemma establishes the existence of the $\mathcal{G}_N^L$-optimal and $\mathcal{G}_N^q$-optimal estimators subject to Assumption (27); it is a modified version of [73, Theorem 1].

Lemma 28 (Existence and convergence of roots) *Suppose that Assumption 27 holds. Then, for all sufficiently large N, $\theta^* \in \mathcal{Z}_N$, there exists $\hat{\boldsymbol{\theta}}_N \in B_r(\theta^*)$ for some $r > 0$ such that $\boldsymbol{G}_N(\hat{\boldsymbol{\theta}}_N) = 0$ almost surely. Furthermore, $\hat{\boldsymbol{\theta}}_N \xrightarrow{a.s.} \theta^*$ as $N \to \infty$.*

PROOF. Note that, for sufficiently large N,

$$\begin{aligned}\|\partial_\theta\boldsymbol{G}_N(\theta) - &\mathbb{E}_\circ[\partial_\theta\boldsymbol{G}_N(\theta^*)]\| \\ &\leq \|\partial_\theta\boldsymbol{G}_N(\theta) - \mathbb{E}_\circ[\partial_\theta\boldsymbol{G}_N(\theta)]\| \\ &\quad + \|\mathbb{E}_\circ[\partial_\theta\boldsymbol{G}_N(\theta)] - \mathbb{E}_\circ[\partial_\theta\boldsymbol{G}_N(\theta^*)]\| \\ &< c_{\theta^*}\end{aligned}$$

for all $\theta^* \in B_s(\theta^*)$, where $0 < s \leq r$ and c_{θ^*} is a constant dependent on θ^*. The last inequality is a consequence of the assumptions and Lemma 26: the first term converges to zero and the second is equicontinuous in θ. Now, observe that Lemma 26 also implies that, $\forall\theta^* \in \mathcal{Z}_N$ for all N sufficiently large, $\boldsymbol{G}_N(\theta^*) \xrightarrow{\text{a.s.}} 0$ as $N \to \infty$. Thus, by [73, Lemma 2], and taking $c_{\theta^*} = \frac{1}{2}\left\|\mathbb{E}_\circ\left[[\partial_\theta \boldsymbol{G}_N(\theta^*)]^{-1}\right]\right\|^{-1}$ as required there, there exists $\hat{\boldsymbol{\theta}}_N \in B_r(\theta^*)$ which is almost surely a root of $\boldsymbol{G}_N(\theta)$ for all sufficiently large N. The second part of the theorem holds because, by Assumption 27, θ^* is a unique root of $\mathbb{E}_\circ[\boldsymbol{G}_N(\theta)]$ on $B_r(\theta^*)$. Let $\bar{\theta}$ be any limit point of $\hat{\boldsymbol{\theta}}_N$. Then all subsequences $\{\hat{\boldsymbol{\theta}}_{N_i}\}$ converge to $\bar{\theta}$, and it holds that $\boldsymbol{G}_{N_i}(\hat{\boldsymbol{\theta}}_{N_i}) - \mathbb{E}_\circ[\boldsymbol{G}_{N_i}(\bar{\theta})] \xrightarrow{\text{a.s.}} 0$. Thus $\mathbb{E}_\circ[\boldsymbol{G}_{N_i}(\bar{\theta})] = 0$ for N_i sufficiently large. Hence, the assumption that θ^* is a unique root of $\mathbb{E}_\circ[\boldsymbol{G}_N(\theta)]$ over $B_s(\theta^*)$ for sufficiently large N implies that $\hat{\boldsymbol{\theta}}_N \xrightarrow{\text{a.s.}} \theta^*$ as $N \to \infty$.

Consistency of the estimators can now be established whenever $\theta_0 \in \Theta$ and u is such that $\theta_\circ$ is identifiable. We adapt the following definition from [2].

Definition 29 (Identifiable parameterization) *For a given input signal 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 $\tilde{N} > N$, it holds that $\mu(\theta) = \mu(\tilde{\theta}) \Leftrightarrow \theta = \tilde{\theta}$.*
- *a second-order identifiable parameterization if there exists $\tilde{N} \in \mathbb{N}$ such that $\forall \theta, \tilde{\theta} \in \Theta$, and $\tilde{N} > N$, it holds that $\mu(\theta) = \mu(\tilde{\theta})$, $\Sigma(\theta) = \Sigma(\tilde{\theta}) \Leftrightarrow \theta = \tilde{\theta}$.*

Notice that the above notion of identifiability depends both on the used model structure and the experimental conditions. When $\theta_\circ \in \Theta$, and due to the unbiasedness of the EFs, it holds that $\mathbb{E}_\circ[\boldsymbol{G}_N(\theta_\circ)] = \mathbb{E}[\boldsymbol{G}_N(\theta_\circ); \theta_\circ] = 0$ for every N, and therefore $\theta_\circ \in \mathcal{Z}_N$ for every N. It is straightforward to see that if u and Θ are such that the parameterization is identifiable, then $\theta_\circ$ is the only element in $\mathcal{Z}_N$ when N is sufficiently large as the following theorem asserts.

Theorem 30 (Consistency) *Suppose that Assumptions 24 and 27 hold. Let the input u be such that Θ constitutes a first-order identifiable parameterization. Then, if $\theta_\circ \in \Theta$ the $\mathcal{G}_N^L$-optimal estimator is strongly consistent, i.e., $\hat{\boldsymbol{\theta}}_N \xrightarrow{a.s.} \theta_\circ$ as $N \to \infty$.*

Suppose that, in addition, Assumption 25 holds and u is such that Θ constitutes a second-order identifiable parameterization. Then the $\mathcal{G}_N^q$-optimal estimator is strongly consistent, i.e., $\hat{\boldsymbol{\theta}}_N \xrightarrow{a.s.} \theta_\circ$ as $N \to \infty$.

PROOF. Define $\mathcal{S}_\epsilon := \{\theta \in \Theta : \|\theta - \theta_\circ\| > \epsilon > 0\}$, let $\boldsymbol{G}_N$ denote the quasi-score function either in $\mathcal{G}_N^L$ or $\mathcal{G}_N^q$. Then, by the properties of $\{\mathbb{E}[\boldsymbol{G}_N(\theta); \theta_\circ]\}$ and the identifiability assumption, it holds that for every $\epsilon > 0$ and sufficiently large N, there exists $\delta_\epsilon > 0$ such that

$$\min_{\theta \in \mathcal{S}_\epsilon} \|\mathbb{E}[\boldsymbol{G}_N(\theta); \theta_\circ]\| > \delta_\epsilon. \tag{40}$$

Consequently,

$$\begin{aligned}
&\min_{\theta \in \mathcal{S}_\epsilon} \|\boldsymbol{G}_N(\theta)\| - \|\boldsymbol{G}_N(\theta_\circ)\| \\
&\geq \min_{\theta \in \mathcal{S}_\epsilon} \{\|\boldsymbol{G}_N(\theta)\| - \|\mathbb{E}_\circ[\boldsymbol{G}_N(\theta)]\|\} \\
&\quad + \min_{\theta \in \mathcal{S}_\epsilon} \|\mathbb{E}_\circ[\boldsymbol{G}_N(\theta)]\| - \|\mathbb{E}_\circ[\boldsymbol{G}_N(\theta_\circ)]\| \\
&\qquad + \{\|\mathbb{E}_\circ[\boldsymbol{G}_N(\theta_\circ)]\| - \|\boldsymbol{G}_N(\theta_\circ)\|\} \\
&\geq \min_{\theta \in \mathcal{S}_\epsilon} \{\|\boldsymbol{G}_N(\theta)\| - \|\mathbb{E}_\circ[\boldsymbol{G}_N(\theta)]\|\} \\
&\quad + \min_{\theta \in \mathcal{S}_\epsilon} \|\mathbb{E}_\circ[\boldsymbol{G}_N(\theta)]\| + \{\|\mathbb{E}_\circ[\boldsymbol{G}_N(\theta_\circ)]\| - \|\boldsymbol{G}_N(\theta_\circ)\|\} \\
&\geq \min_{\theta \in \mathcal{S}_\epsilon} \|\mathbb{E}_\circ[\boldsymbol{G}_N(\theta)]\| - 2 \sup_{\theta \in \Theta} \|\boldsymbol{G}_N(\theta) - \mathbb{E}_\circ[\boldsymbol{G}_N(\theta)]\| \\
&> 0 \text{ a.s. for sufficiently large } N.
\end{aligned} \tag{41}$$

The last inequality is a consequence of Lemma 26 together with (40), and the third inequality holds because, by the reverse triangle inequality, we have

$$\begin{aligned}
&\min_{\theta \in \mathcal{S}_\epsilon} \{\|\boldsymbol{G}_N(\theta)\| - \mathbb{E}_\circ[\|\boldsymbol{G}_N(\theta)\|]\} + \{\|\mathbb{E}_\circ[\boldsymbol{G}_N(\theta_\circ)]\| - \|\boldsymbol{G}_N(\theta_\circ)\|\} \\
&\quad \leq \min_{\theta \in \mathcal{S}_\epsilon} \|\boldsymbol{G}_N(\theta) - \mathbb{E}_\circ[\boldsymbol{G}_N(\theta)]\| + \|\boldsymbol{G}_N(\theta_\circ) - \mathbb{E}_\circ[\boldsymbol{G}_N(\theta_\circ)]\| \\
&\quad \leq 2 \sup_{\theta \in \Theta} \|\boldsymbol{G}_N(\theta) - \mathbb{E}_\circ[\boldsymbol{G}_N(\theta)]\|.
\end{aligned}$$

Let $\hat{\boldsymbol{\theta}}_N$ denote either the $\mathcal{G}_N^L$-optimal or the $\mathcal{G}_N^q$-optimal estimator – their existence is guaranteed by Lemma 28. Then, by definition, $\|\boldsymbol{G}_N(\hat{\boldsymbol{\theta}}_N)\| - \|\boldsymbol{G}_N(\theta_\circ)\| = -\|\boldsymbol{G}_N(\theta_\circ)\| \leq 0$, where $\boldsymbol{G}_N$ is the quasi-score function defining $\hat{\boldsymbol{\theta}}_N$, which in the light of (41) implies that $\hat{\boldsymbol{\theta}}_N \notin \mathcal{S}_\epsilon$ a.s. for sufficiently large N. Because ϵ is arbitrary, it holds that $\hat{\boldsymbol{\theta}}_N \xrightarrow{\text{a.s.}} \theta_\circ$ as $N \to \infty$.

8 Asymptotic Normality

Based on the consistency results of the $\mathcal{G}_N^L$-optimal and $\mathcal{G}_N^q$-optimal estimators, their asymptotic normality can be shown with no additional assumptions on the model. This advantage is due to the definition of $\hat{\boldsymbol{\theta}}_N$ as roots and the use of Assumption 27. However, a slight strengthening of the assumption on the data is required, as shown in the following theorem.

Theorem 31 (Asymptotic normality) *Assume that the hypotheses of Theorem 30 hold. Furthermore, assume that the data is exponentially forgetting of order $r > 4$ (see [47, page 32]). Introduce the (normalizing) matrices $P_N := \mathbb{E}[N^{-1}\boldsymbol{G}_N(\theta_\circ)\boldsymbol{G}_N^\top(\theta_\circ); \theta_\circ]^{-1}$, $N \in \mathbb{N}$. Then*

$$\sqrt{N} P_N^{-\frac{1}{2}}(\hat{\boldsymbol{\theta}}_N - \theta_\circ) \rightsquigarrow \mathcal{N}(0, I_{d_\theta}) \quad \text{as} \quad N \to \infty,$$

where I_{d_θ} denotes the identity matrix of size d_θ, $\hat{\boldsymbol{\theta}}_N$ denotes the $\mathcal{G}_N^L$-optimal and $\mathcal{G}_N^q$-optimal estimators, and $\boldsymbol{G}_N$ denotes the corresponding quasi-score functions. Moreover, if the data and the model are such that

$$\mathbb{E}[N^{-1}\boldsymbol{G}_N(\theta)\boldsymbol{G}_N^\top(\theta); \theta_\circ]^{-1} \to Q(\theta) \succ 0 \text{ as } N \to \infty \tag{42}$$

then $\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \theta_\circ) \rightsquigarrow \mathcal{N}(0, P)$, where $P := Q(\theta_\circ)$.

PROOF. The proof is analogous to that of Theorem 1 in [47] when the derivative and the Hessian of the criterion function there are replaced by $\boldsymbol{G}(\theta)$ and $\partial_\theta \boldsymbol{G}_N(\theta)$.

Observe that here, by Assumption 10 and the assumption that $\theta_\circ \in \Theta$, the sequence of normalizing factors $\{P_N\}$ is well-defined and every P_N is positive definite. Moreover, due to the properties of quasi-score functions, it holds that $-\mathbb{E}[\partial_\theta \boldsymbol{G}_N(\theta);\theta] = \mathbb{E}[\boldsymbol{G}_N(\theta)\boldsymbol{G}_N^\top(\theta);\theta]$ for all $\theta \in \Theta,\ N \in \mathbb{N}$ (see Remark 23), which leads to the given (reduced) expression of the normalizing factors (compare to those of PEM estimators (6)). The second part of the theorem is an implication of the first part and the assumption in (42) by a direct application of Slutsky's theorem. The assumption in (42) will be satisfied, for example, when $\boldsymbol{y}$ is asymptotically stationary, quasi-stationary or periodic.

We end this section by the following important concluding result.

Theorem 32 *Suppose that the hypotheses of Theorem 30 and Theorem 31 are satisfied for the $\mathcal{G}_N^L$-optimal and $\mathcal{G}_N^q$-optimal estimators. Denote the corresponding normalizing factors as P_N^L and P_N^q respectively. Then, it holds that $P_N^L \succeq P_N^q,\ N \in \mathbb{N}$. Moreover, if (42) holds for the quasi-score function in $\mathcal{G}_N^L$ and $\mathcal{G}_N^q$, then $P^L \succeq P^q$ with obvious notations.*

PROOF. This is a consequence of the definition of quasi-score functions, the fact that $\mathcal{G}_N^L \subset \mathcal{G}_N^q,\ \forall N \in \mathbb{N}$, and the definitions of P_N and P in Theorem 31.

In the next section we briefly discuss the relation between the PE correlation approach and the EF approach.

9 Links to the correlation approach

The idea of defining point estimators as solutions to algebraic equations is not new to the system identification community. For instance, the well-known instrumental-variable method was introduced in the system identification literature in the 1960s; see e.g. the articles [70–72], and the books [16,62]. Such methods are representative of the PE correlation approach as described by Ljung in [44, Section 7.5]. In the correlation approach, one defines a point estimator $\hat{\boldsymbol{\theta}}_N := \underset{\theta\in\Theta}{\operatorname{sol}}[f_N(\theta, \boldsymbol{D}_N) = 0]$, where

$$f_N(\theta, \boldsymbol{D}_N) := \frac{1}{N}\sum_{t=1}^{N} \zeta(t,\theta)\alpha(\tilde{\boldsymbol{e}}_t(\theta)), \qquad (43)$$

in which $\zeta(t,\theta) = \zeta(t, \boldsymbol{D}_{t-1}, \theta)$ is a sequence of correlation vectors constructed from past data, and α is a transformation of a linearly filtered PE process $\tilde{\boldsymbol{e}}(\theta)$. The formal links to the more general EF approach are obvious. Observe that the $\mathcal{G}_N^L$-optimal estimator can be seen in a correlation approach where $\boldsymbol{e} = \boldsymbol{\varepsilon}$ (the linear innovation process), $\alpha(\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}$, and $\zeta(t,\theta) := \mathbb{E}[\partial_\theta \boldsymbol{\varepsilon}_t(\theta);\theta]^\top E[\boldsymbol{\varepsilon}_t(\theta)\boldsymbol{\varepsilon}_t^\top(\theta);\theta]^{-1}$. However, estimators defined as roots of EFs in $\mathcal{G}_N^q$ or $\mathcal{G}_N^Q$ do not necessarily fit in the PE correlation approach; compare (30) and (36) to (43) and notice that the function $\alpha(\cdot)$ in (43) is static, and that $\zeta(t,\theta)$ is a $d_\theta \times d_y$ matrix derived from D_{t-1}.

The theory of optimal EFs provides a criteria for optimal weighting of the sum defining f_N. It also unifies the PE minimization and the PE correlation methods under one framework with a systematic way of constructing efficient estimators based on partial model specifications.

10 Examples

In this section, we consider several examples where we compute the asymptotic covariance matrices of the $\mathcal{G}_N^L$-optimal and $\mathcal{G}_N^q$-optimal estimators in addition to different PEM estimators. The examples show that the linear and quadratic quasi-score functions may be used to systematically construct optimal estimators when only the knowledge of the first two and four moments is used, respectively. While the examples are given for nonlinear models, it should be clear that the methods also apply to the linear case when the full distribution is not specified.

We start by providing an outline of the main steps of the proposed estimation methods.

10.1 Methods outline

For simplicity, we assume $d_y = 1$, and for any vector E we define E^2 as a vector with entries $[E^2]_i := [E]_i^2$.

(1) Use the model to write the outputs in a vector form $\boldsymbol{Y}(\theta) = \mathcal{M}(U, \boldsymbol{Z}; \theta)$ where U is a vector of known inputs and $\boldsymbol{Z}$ is a vector of disturbances and noises whose moments are parameterized by θ.

(2) Compute the following vectors and matrices, where all expectations are with respect to $\boldsymbol{Z}$:
  - $\mu(\theta) = \mathbb{E}[\boldsymbol{Y}(\theta);\theta]$, and its gradient $\partial_\theta\mu(\theta)$,
  - $\Sigma(\theta) = \mathbb{E}[\boldsymbol{E}(\theta)\boldsymbol{E}^\top(\theta);\theta]$, where $\boldsymbol{E}(\theta) = \boldsymbol{Y}(\theta) - \mu(\theta)$,
  - $\sigma(\theta)$ as $[\sigma(\theta)]_i = [\Sigma(\theta)]_{ii}$, and its gradient $\partial_\theta\sigma(\theta)$,
  - $T(\theta) = \mathbb{E}[\boldsymbol{E}^2(\theta)\boldsymbol{E}^\top(\theta);\theta]$, $P(\theta) = \mathbb{E}[\boldsymbol{E}^2(\theta)[\boldsymbol{E}^2(\theta)]^\top;\theta]$.

(3) Use the data to realize the $\mathcal{G}_N^L$-optimal and $\mathcal{G}_N^q$-optimal estimators by solving the two systems of equations $\partial_\theta\mu^\top(\theta)\Sigma^{-1}(\theta)(Y - \mu(\theta)) = 0$, and

$$\begin{bmatrix}\partial_\theta\mu^\top(\theta) & \partial_\theta\sigma^\top(\theta)\end{bmatrix}\begin{bmatrix}\Sigma(\theta) & T^\top(\theta)\\ T(\theta) & P(\theta)\end{bmatrix}^{-1}\begin{bmatrix}Y - \mu(\theta)\\ (Y-\mu(\theta))^2 - \sigma(\theta)\end{bmatrix} = 0,$$

respectively, where Y is the vector of observed outputs.

10.2 A Gaussian model

Suppose that

$$\begin{aligned}\boldsymbol{y}_t &= \theta_\circ u_t^2 + \sqrt{2}\theta_\circ u_t^2 \boldsymbol{w}_t,\\ \boldsymbol{w}_t &\overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1), \quad t = 1,\ldots,N\end{aligned} \qquad (44)$$

in which $\theta_\circ = 1$ and the input signal u is a known realization of an independent standard Gaussian process. In this case, conditioned on u_t, $\boldsymbol{y}_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(\theta_\circ u_t^2, 2{\theta_\circ}^2 u_t^4)$. Let $\boldsymbol{e}_t(\theta) = \boldsymbol{y}_t - \theta u_t^2$ and consider the three estimators obtained using a criterion function where

$$(1)\quad \ell^{(1)}(\boldsymbol{e}_t(\theta)) = \frac{1}{2}\boldsymbol{e}_t(\theta)^2, \qquad \text{(OE-QPEM)}$$

$$(2)\quad \ell^{(2)}(\boldsymbol{e}_t(\theta), t) = \frac{1}{2}\frac{\boldsymbol{e}_t(\theta)^2}{2{\theta_\circ}^2 u_t^4}, \qquad \text{(optimal OE-WQPEM)}$$

$$(3)\quad \ell^{(3)}(\boldsymbol{e}_t(\theta), t, \theta) = \frac{1}{2}\frac{\boldsymbol{e}_t(\theta)^2}{2\theta^2 u_t^4} + \log(\theta). \qquad \text{(OL-GPEM)}$$

In addition, consider the $\mathcal{G}_N^L$-optimal and $\mathcal{G}_N^q$-optimal estimators. Observe that we used the true parameter in the definition of $\ell^{(2)}$, so that we are using the optimal weights. Also note that the OL-GPEM estimator coincides with the MLE in this example.

By Theorem 17, the quasi-score function in $\mathcal{G}_N^L$ is

$$\boldsymbol{G}_N(\theta) = \sum_{t=1}^{N} \frac{-1}{2\theta^2 u_t^2}(\boldsymbol{y}_t - \theta u_t^2),$$

therefore $\mathbb{E}[\boldsymbol{G}_N(\theta)\boldsymbol{G}_N^\top(\theta); \theta] = \frac{N}{2\theta^2}$. By Theorem 31, the asymptotic variance of the $\mathcal{G}_N^L$-optimal estimator is $\mathbb{E}[N^{-1}\boldsymbol{G}_N(\theta_\circ)\boldsymbol{G}_N^\top(\theta_\circ); \theta_\circ]^{-1} = 2$. On the other hand, by Theorem 19, the quasi-score function in $\mathcal{G}_N^Q$ is

$$\boldsymbol{G}_N(\theta) = \sum_{t=1}^{N} \frac{-1}{2\theta^2 u_t^2}(\boldsymbol{y}_t - \theta u_t^2) + \frac{-1}{2\theta^3 u_t^4}(\boldsymbol{y}_t - \theta u_t^2)^2 + \frac{1}{\theta}.$$

This EF coincides with the quasi-score function in $\mathcal{G}_N^q$ and is equivalent to the score function because $\boldsymbol{y}$ is an independent Gaussian process (see Remark 20). In this case, $\mathbb{E}[\boldsymbol{G}_N(\theta)\boldsymbol{G}_N^\top(\theta); \theta] = \frac{5N}{2\theta}$ and therefore, by Theorem 31, the asymptotic variance of the $\mathcal{G}_N^q$-optimal estimator is $\mathbb{E}[N^{-1}\boldsymbol{G}_N(\theta_\circ)\boldsymbol{G}_N^\top(\theta_\circ); \theta_\circ]^{-1} = 0.4$, which is equal to CRLB for the model in (44).

Table 1 shows the asymptotic variance of the five estimators. The OE-QPEM estimator has the largest asymptotic variance. The OL-GPEM estimator is efficient and coincides with the MLE with an asymptotic CRLB equal to 0.4. The OE-WQPEM estimator, which uses the knowledge of the exact variances of $\boldsymbol{y}_t$ for all t but ignores the joint parameterization of the mean and the variance, has an asymptotic variance of 2; five times larger than the CRLB. The $\mathcal{G}_N^L$-optimal estimator has the same asymptotic variance as the OE-WQPEM estimator. The $\mathcal{G}_N^q$-optimal estimator is efficient and achieves the asymptotic CRLB. The analytic results are confirmed by running Monte Carlo simulations in which we used 10^4 observations and 10^4 Monte Carlo realizations.

If the assumption that $\boldsymbol{y}$ is Gaussian is not maintained, the results will be much different, as we now show.

Table 1
The asymptotic variance for five estimators of the Gaussian model in Section 10.2 (Notice that the optimal OE-WQPEM requires the knowledge of $\theta_\circ$ in order to define the weighting.)

Estimator	**Analytic asy. var.**	**MC approx**
OE-QPEM	$\frac{70}{3}\theta_\circ^2 \approx 23.333$	23.185
optimal OE-WQPEM	$2\theta_\circ^2 = 2$	2.019
OL-GPEM (≡ MLE)	$0.4\theta_\circ^2 =$ *0.4 (CRLB)*	*0.4022*
$\mathcal{G}_N^L$-optimal	2	2.019
$\mathcal{G}_N^q$*-optimal (≡ MLE)*	*0.4 (CRLB)*	*0.4022*

10.3 A Gamma model

Suppose that

$$\begin{aligned} \boldsymbol{y}_t &= \theta_\circ u_t^2 \boldsymbol{w}_t^2, \\ \boldsymbol{w}_t &\overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1), \qquad t = 1, \ldots, N. \end{aligned} \tag{45}$$

In this case, the outputs have a Gamma distribution; namely $\boldsymbol{y}_t \overset{\text{i.i.d}}{\sim} \Gamma(\alpha, \beta_t(\theta_\circ)), t = 1, \ldots, N$, where $\alpha = \frac{1}{2}$, $\beta_t(\theta_\circ) = 2\theta_\circ u_t^2$, $\theta_\circ = 1$. Notice that the first two moments of the Gamma model in (45) coincide with the Gaussian model in Section 10.2; namely $\mu_t = \theta u_t^2$, $\lambda_t = 2\theta^2 u_t^4$. However, the third and fourth central moments here are $m_t^{(3)} = 8\theta^3 u_t^6$, $m_t^{(4)} = 60\theta^4 u_t^8$ where we used the notations introduced after Assumption 4.

Consider the same five estimators used in Section 10.2 and observe that because the definitions of the OE-QPEM estimator, the optimal OE-WQPEM estimator, and the $\mathcal{G}_N^L$-optimal estimator depend only on the first and second moments of $\boldsymbol{y}$, their asymptotic variances are exactly the same as with the Gaussian case. However, the quasi-score function in $\mathcal{G}_N^q$ is now given by

$$\boldsymbol{G}_N(\theta) = \sum_{t=1}^{N} \frac{-1}{2\theta^2 u_t^2}(\boldsymbol{y}_t - \theta u_t^2),$$

which coincides with the quasi-score function in $\mathcal{G}_N^L$ and the score function (the multipliers of the quadratic terms in (30) and (36) are zero). The asymptotic variance of the $\mathcal{G}_N^q$-optimal estimator is $\mathbb{E}[N^{-1}\boldsymbol{G}_N(\theta_\circ)\boldsymbol{G}_N^\top(\theta_\circ); \theta_\circ]^{-1} = 2$, which is also the asymptotic CRLB for the model in (45).

Table 2 summarises the results for the five estimators. Here, the optimal OL-WQPEM and the $\mathcal{G}_N^L$-optimal estimators are both efficient. However, the weighting of the optimal OE-WQPEM estimator requires the knowledge of $\theta_\circ$ unlike the $\mathcal{G}_N^L$-optimal estimator. From these results, it is clear that under departure from normality, a pseudo-score function based on a Gaussian log-likelihood has no optimality properties (e.g., within the class of unbiased quadratic EFs), and therefore comes with no support or motivation. On the other hand, using a quasi-score function within a well-defined space, say the space of unbiased quadratic functions $\mathcal{G}_N^q$, leads to optimal estimators (within the given class) regardless of the true distribution once the conditions discussed in Sections 7 and 8 are satisfied. To further clarify the

Table 2
The asymptotic variance for five estimators of the Gamma model in Section 10.3

Estimator	Analytic asy. var.	MC approx
OE-QPEM	$\frac{70}{3}\theta_\circ^2 \approx 23.333$	23.282
optimal OE-WQPEM (≡ MLE)	$2\theta_\circ^2 = 2$ *(CRLB)*	*1.983*
OL-GPEM	$\frac{74}{25}\theta_\circ^2 = 2.96$	2.948
$\mathcal{G}_N^L$*-optimal (≡ MLE)*	*2 (CRLB)*	*1.983*
$\mathcal{G}_N^q$*-optimal (≡ MLE)*	*2 (CRLB)*	*1.983*

above conclusions, we consider next an example where the score function is analytically intractable.

10.4 A bi-modal model

Suppose that

$$\begin{aligned}\boldsymbol{y}_t &= \theta_\circ(u_t + \boldsymbol{w}_t)^2,\\ &= \theta_\circ u_t^2 + \theta_\circ \boldsymbol{w}_t^2 + 2\theta_\circ u_t \boldsymbol{w}_t, \quad t = 1, \dots, N,\end{aligned}$$

where $\boldsymbol{w}_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$, and once more, $\theta_\circ = 1$ and the inputs are known realizations of $\boldsymbol{u}_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$. Observe that in this case, the log-likelihood of θ is analytically intractable, and that the distribution of $\boldsymbol{y}$ may be bi-modal. However, the mean and the variance of $\boldsymbol{y}$ are given by the closed-form expressions $\mu_t = \theta(u_t^2+1)$, $\lambda_t = \theta^2(4u_t^2+2)$. The third and fourth centered moments are also available in closed-form and are given by $m_t^{(3)} = \theta^3(24u_t^2+8)$, $m_t^{(4)} = \theta^4(48u_t^4 + 240u_t^2 + 60)$. We now consider the five estimators used in the above two examples. Straightforward computations show that the quasi-score function in $\mathcal{G}_N^L$ is

$$\boldsymbol{G}_N(\theta) = \sum_{t=1}^{N} \frac{-(u_t^2+1)}{\theta^2(4u_t^2+2)}(\boldsymbol{y}_t - \theta(u_t^2+1)),$$

and the quasi-score function in $\mathcal{G}_N^q$ is

$$\boldsymbol{G}_N(\theta) = \sum_{t=1}^{N} a_t^\star(\theta) \begin{bmatrix} (\boldsymbol{y}_t - \theta(u_t^2+1)) \\ (\boldsymbol{y}_t - \theta(u_t^2+1))^2 - \theta^2(4u_t^2+2) \end{bmatrix},$$

$$a_t^\star(\theta) = \begin{bmatrix} \frac{-(32u_t^6+64u_t^4+120u_t^2+24)}{\theta^2(128u_t^6+384u_t^4+288u_t^2+48)} \\ \frac{-8u_t^4}{\theta^3(128u_t^6+384u_t^4+288u_t^2+48)} \end{bmatrix}^\top.$$

Except for the OE-QPEM estimator, the analytical expressions of the asymptotic variance for all the estimators involve limits of ratios of polynomials of the input, and therefore are analytically intractable; they are approximated here using Monte Carlo simulations. The values are given in Table 3. Once more, we see that the asymptotic variance of the optimal OE-WQPEM estimator is smaller than the OL-GPEM estimator. As expected, see Theorem 32, the $\mathcal{G}_N^q$-optimal estimator has the least asymptotic variance among all the estimators, followed by the $\mathcal{G}_N^L$-optimal estimator.

In the following example, we consider the identification of a relatively challenging model whose output is dependent over time.

Table 3
Asymptotic variances for the five estimators in Section 10.4

Estimator	Analytic asy. var.	MC approx
OE-QPEM	$\frac{25}{9}\theta_\circ^2 \approx 2.777$	2.774
optimal OE-WQPEM	N/A	1.475
OL-GPEM	N/A	2.061
$\mathcal{G}_N^L$-optimal	N/A	1.475
$\mathcal{G}_N^q$*-optimal*	N/A	*1.452*

Table 4
MSE for the six estimators in Section 10.5, approximated using 1000 Monte Carlo simulations when $\boldsymbol{N = 500}$.

Estimator	$\widehat{\mathbf{MSE}(\hat{\boldsymbol{\alpha}})}$	$\widehat{\mathbf{MSE}(\hat{\boldsymbol{\lambda}})}$	$\widehat{\mathbf{MSE}(\hat{\boldsymbol{\theta}})}$
OE-QPEM	6.033×10^{-4}	2.709×10^{-2}	2.769×10^{-2}
OL-GPEM	6.818×10^{-4}	1.282×10^{-2}	1.350×10^{-2}
OE-WQPEM	3.555×10^{-4}	1.605×10^{-2}	1.641×10^{-2}
optimal OE-WQPEM	3.568×10^{-4}	1.606×10^{-2}	1.641×10^{-2}
$\mathcal{G}_N^L$-optimal	3.547×10^{-4}	1.611×10^{-2}	1.647×10^{-2}
$\mathcal{G}_N^q$*-optimal*	3.149×10^{-4}	0.720×10^{-2}	0.751×10^{-2}

10.5 First-order stochastic Wiener model

Suppose that

$$\boldsymbol{y}_t = (G(\mathrm{q};\alpha_\circ)u_t + \boldsymbol{w}_t)^2 + H(\mathrm{q})\boldsymbol{v}_t, \quad t = 1, \dots, N,$$

where $G(\mathrm{q};\alpha_\circ) = \frac{1}{1+\alpha_\circ\,\mathrm{q}^{-1}}$, $H(\mathrm{q}) = \frac{1}{1-0.8\,\mathrm{q}^{-1}}$, $\boldsymbol{w}_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,\lambda_\circ)$ which are independent of $\boldsymbol{v}_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 0.1)$, and $\theta_\circ := [\alpha_\circ\ \lambda_\circ]^\top = [-0.7\ 1]^\top$. Let the inputs be *a priori* known realizations of $\boldsymbol{u}_t \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$. In this case, the likelihood function of θ is analytically intractable and the outputs $\boldsymbol{y}_t$ are dependent over time, which makes the problem challenging. However, it is still possible to compute the first four moments of $\boldsymbol{y}$ in closed-form; thus, one may use any of the linear PEM estimators or the EFs estimators proposed in this paper. Tables 4 and 5 show the MSE, approximated using 1000 Monte Carlo realizations, of six estimators of $\theta_\circ$ when $N = 500$ and 1000 respectively. The estimator OE-WQPEM refers to the weighted OE-QPEM estimator when the weighting matrix is given by $\Sigma(\hat{\theta}_N)$ where $\hat{\theta}_N$ is the OL-GPEM estimate obtained using the same data set. The results show that the OE-QPEM estimator has the largest MSE. The optimal OE-WQPEM estimator (that uses the true parameters to compute the optimal weighting) has an MSE (approximately) equal to that of the $\mathcal{G}_N^L$-optimal estimator and the OE-WQPEM estimator. The $\mathcal{G}_N^q$-optimal estimator has the smallest MSE among all the estimators.

11 Conclusions

In this contribution, we showed that the accuracy of linear PEMs based on a Gaussian assumption depends, not only on the shape of the unknown distribution of the data, but also on the parameterization (see Section 4.1 and Theorem 5). Thus, it is not obvious in general which linear PEM estimator should be preferred. On the other hand, the EFs approach provides a systematic way

Table 5
MSE for the six estimators in Section 10.5, approximated using 1000 Monte Carlo simulations when $\boldsymbol{N = 1000}$.

Estimator	$\widehat{\mathbf{MSE}(\hat{\boldsymbol{\alpha}})}$	$\widehat{\mathbf{MSE}(\hat{\boldsymbol{\lambda}})}$	$\widehat{\mathbf{MSE}(\hat{\boldsymbol{\theta}})}$
OE-QPEM	2.733×10^{-4}	12.72×10^{-3}	12.99×10^{-3}
OL-GPEM	3.038×10^{-4}	7.636×10^{-3}	7.940×10^{-3}
OE-WQPEM	1.643×10^{-4}	8.280×10^{-3}	8.444×10^{-3}
optimal OE-WQPEM	1.642×10^{-4}	8.256×10^{-3}	8.420×10^{-3}
$\mathcal{G}_N^L$-optimal	1.643×10^{-4}	8.266×10^{-3}	8.430×10^{-3}
$\mathcal{G}_N^q$-optimal	1.459×10^{-4}	3.700×10^{-3}	3.846×10^{-3}

of constructing optimal consistent estimators, within a given class, for many challenging models. It generalizes the ML method and the PEMs, and provides an optimality criterion that may be used under model misspecification. We defined the $\mathcal{G}_N^L$-optimal and $\mathcal{G}_N^q$-optimal estimators based on quasi-score functions that are linear and quadratic in the PEs, respectively. Their convergence and consistency were established under standard assumptions akin to those of PEMs. As shown by the theoretical analysis and the simulation examples, unless the model is Gaussian, the $\mathcal{G}_N^q$-optimal estimator is uniformly asymptotically more accurate than linear PEMs when a quadratic criterion is used.

References

[1] M. Abdalmoaty and H. Hjalmarsson. Consistent estimators of stochastic MIMO Wiener models based on suboptimal predictors. In *57th IEEE Conference on Decision and Control (CDC)*, pages 3842–3847, Dec 2018.

[2] M. Abdalmoaty and H. Hjalmarsson. Linear prediction error methods for stochastic nonlinear models. *Automatica*, 105:49–63, 2019.

[3] C. Andrieu, A. Doucet, S. S. Singh, and V. B. Tadic. Particle methods for change detection, system identification, and control. *Proceedings of the IEEE*, 92(3):423–438, 2004.

[4] R. R. Bahadur. On Fisher's bound for asymptotic variances. *Ann. of Math. Statis.*, 35(4):1545–1552, 1964.

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

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

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

[8] G. Casella and R. L. Berger. *Statistical Inference.* Duxbury Thomson Learning, 2008.

[9] K. L. Chung. *A Course in Probability Theory.* Academic Press, 2001.

[10] H. Cramér. *Mathematical Methods of Statistics.* Princeton University Press, 1946.

[11] M. Crowder. On consistency and inconsistency of estimating equations. *Econometric Theory*, 2(3):305–330, 1986.

[12] M. Crowder. On linear and quadratic estimating functions. *Biometrika*, 74(3):591–597, 1987.

[13] 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.

[14] J. Durbin. Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 22(1):139–153, 1960.

[15] M. Enqvist. *Linear Models of Nonlinear Systems.* Dissertation No. 985, Linköping University, Sweden, 2005.

[16] P. Eykhoff. *System Identification. Parameter and State Estimation.* John Wiley, 1974.

[17] T. S. Ferguson. *A Course in Large Sample Theory.* Chapman &Hall/CRC, 1996.

[18] D. Firth. On the efficiency of quasi-likelihood estimation. *Biometrika*, 74(2):233–245, 1987.

[19] G. Giordano, S. Gros, and J. Sjöberg. An improved method for Wiener-Hammerstein system identification based on the fractional approach. *Automatica*, 94:349 – 360, 2018.

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

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

[22] V. P. Godambe. An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, 31(4):1208–1211, 12 1960.

[23] V. P. Godambe. The foundations of finite sample estimation in stochastic processes. *Biometrika*, 72(2):419–428, 1985.

[24] V. P. Godambe. *Estimating Functions.* Oxford science publications. Clarendon Press, 1991.

[25] V. P. Godambe and C. C. Heyde. Quasi-likelihood and optimal estimation, correspondent paper. *International Statistical Review*, 55(3):231–244, 1987.

[26] V. P. Godambe and M. E. Thompson. An extension of quasi-likelihood estimation. *Journal of Statistical Planning and Inference*, 22(2):137 – 152, 1989.

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

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

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

[30] D. A. Harville. *Matrix Algebra From a Statistician's Perspective.* Springer New York, 2006.

[31] C. C. Heyde. Fixed sample and asymptotic optimality for classes of estimating functions. *Contemporary Mathematics*, 80:241–247, 1988.

[32] C. C. Heyde. *Quasi-Likelihood And Its Application: A General Approach to Optimal Parameter Estimation.* Springer New York, 1997.

[33] 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.

[34] A. M. Kagan. Fisher information contained in a finite-dimensional linear space, and a correctly posed version of the method of moments. *Problems Inform. Transmission*, 12:4:98–115, 1976. English translation from Russian; Probl. Peredachi Inf., 12:2 (1976), 2042.

[35] N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin. On particle methods for parameter estimation in state-space models. *Statist. Sci.*, 30(3):328–351, 2015.

[36] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation theory.* PTR Prentice-Hall, 1993.

[37] F. Lamnabhi-Lagarrigue, A. Annaswamy, S. Engell, A. Isaksson, P. Khargonekar, R. M. Murray, H. Nijmeijer, T. Samad, D. Tilbury, and P. Van den Hof. Systems & control for the future of humanity, research agenda: Current and future roles, impact and grand challenges. *Annual Reviews in Control*, 43(Supplement C):1 – 64, 2017.

[38] L. Le Cam. On some asymptotic properties of maximum likelihood estimates and related results. *University of California Publications in Statistics*, 1:277–330, 1953.

[39] E. L. Lehmann. *Elements of Large-Sample Theory.* Springer, 1999.

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

[41] 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.

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

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

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

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

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

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

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

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

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

[51] 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.

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

[53] T. B. Schön, F. Lindsten, J. Dahlin, J. Wågberg, C. A. Naesseth, A. Svensson, and L. Dai. Sequential Monte Carlo methods for system identification. *IFAC-PapersOnLine*, 48(28):775 – 786, 2015.

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

[55] J. Schoukens and L. Ljung. Nonlinear system identification: A user-oriented road map. *IEEE Control Systems Magazine*, 39(6):28–99, Dec 2019.

[56] 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.

[57] 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.

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

[59] 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.

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

[61] 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.

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

[63] A. Stuart and K. Ord. *Kendall's Advanced Theory of Statistics: Volume 1: Distribution Theory.* Wiley, 2009.

[64] 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.

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

[66] 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.

[67] R. W.M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3):439–447, 1974.

[68] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

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

[70] K. Wong and E. Polak. Identification of linear discrete time systems using the instrumental variable method. *IEEE Transactions on Automatic Control*, 12(6):707–718, 1967.

[71] P. Young. An instrumental variable method for real-time identification of a noisy process. *Automatica*, 6(2):271 – 287, 1970.

[72] P. Young. Some observations on instrumental variable methods of time-series analysis. *International Journal of Control*, 23(5):593–612, 1976.

[73] K.-H. Yuan and R. I. Jennrich. Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, 65(2):245 – 260, 1998.