Consistent Estimators of Stochastic MIMO Wiener Models based on Suboptimal Predictors

Mohamed Rasheed Abdalmoaty and Håkan Hjalmarsson

***Abstract*— We consider a parameter estimation problem in a general class of stochastic multiple-inputs multiple-outputs Wiener models, where the likelihood function is, in general, analytically intractable. When the output signal is a scalar independent stochastic process, the likelihood function of the parameters is given by a product of scalar integrals. In this case, numerical integration may be efficiently used to approximately solve the maximum likelihood problem. Otherwise, the likelihood function is given by a challenging multidimensional integral. In this contribution, we argue that by ignoring the dynamical character of the stochastic disturbances, a computationally attractive estimator based on a suboptimal predictor can be constructed by evaluating *scalar integrals* regardless of the number of outputs. Under some conditions, the convergence of the resulting estimators can be established and consistency is achieved under certain identifiability hypothesis. We highlight the relationship between the resulting estimators and a recently proposed prediction error method estimator. We also remark that the method can be used for a wider class of stochastic nonlinear models. The performance of the method is demonstrated by a numerical simulation example using a 2-inputs 2-outputs model with 9 parameters.**

I. INTRODUCTION

There has been a growing interest in the estimation problem of stochastic Wiener models during the last decade; see for example [5], [23], [22], [21], [2]. A Wiener model is defined by a Linear-Time Invariant (LTI) model followed by a static nonlinear function at its output as shown in Figure 1. When a disturbance $\boldsymbol{w}$ is present before the

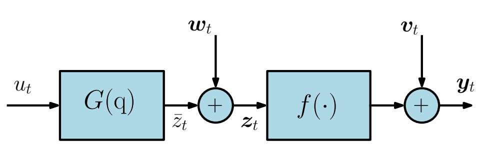

Fig. 1: A stochastic Wiener model: G represents a linear-time invariant model, and f is a static nonlinear function. The signals $\boldsymbol{w}$ and $\boldsymbol{v}$ represent an unobserved stochastic disturbance and measurement noise respectively.

nonlinearity, the model is called "stochastic". This is a reasonable model for stochastic linear systems when the measurement sensors are nonlinear, as well as for several complicated nonlinear systems; see for example [7], [8], and [25]. It is also well known that the class of multiple-inputs multiple-outputs (MIMO) Wiener models is rich enough to approximate, arbitrarily well, a large class of nonlinear systems [3]. It represents one of the simplest instances of the block-oriented models [4], [14] which is used to model nonlinear systems by interconnecting LTI models and static (memoryless) nonlinear functions in various forms.

Historically, most of the research literature on Wiener models and similar block-oriented models considered cases where the only source of uncertainty is at the model's output; i.e., when $\boldsymbol{w} = 0$. Under this assumption, the likelihood function of the parameters and the optimal Mean-Square Error (MSE) one-step ahead predictor may be computed analytically, and therefore the Maximum Likelihood (ML) problem or a Prediction Error Method (PEM) problem can be formulated and solved using standard numerical optimization algorithms. The focus has been on problems such as model structure selection and parameter initialization. Several specialized methods of structure selection have been developed in both time and frequency domain; see for example the recent survey [17]. It has also been recognized that linear approximations may be used to initialize the estimation problem [18], [15], [20], [19], [13], [16].

If the assumption $\boldsymbol{w} = 0$ is not maintained, the resulting estimators are not guaranteed to be consistent, as was first shown in [6]. A main difficulty associated with the presence of an unobserved disturbance before the nonlinearity is the intractability of the likelihood function, as well as the optimal one-step predictor. An expression for the likelihood function in this case was derived in [5] where an implementation of a quasi-Newton algorithm based on numerical integration was proposed for ML estimation in cases where $\boldsymbol{w}$ is independent and all signals are scalars. It was also observed in the simulation example of [5] that the resulting estimator is consistent even if $\boldsymbol{w}$ is colored, while [24] concluded the opposite. Notice that if the outputs are independent, the optimal one-step predictor is given by the unconditional expectation which, depending on the model, may not assume a closed-form expression. To obtain a PEM estimator, an approximation based on the unscented transform has been suggested in [22].

When the disturbance $\boldsymbol{w}$ is colored and the model has multiple outputs, the likelihood function and the optimal predictor are given in terms of high-dimensional marginalization integrals that are challenging to compute. A solution to the ML problem in this case, based on the expectation-maximization algorithm and sequential Monte Carlo (particle) smoothers, has been given in [23], [24]. Although

This work was supported by the Swedish Research Council through the project System identification: Unleashing the algorithms, contract 2015-05285 and the research environment NewLEADS - New Directions in Learning Dynamical Systems, contract 2016-06079.

Mohamed Rasheed Abdalmoaty and Håkan Hjalmarsson are with the Automatic Control Department, School of Electrical Engineering and Computer Science, KTH Royal institute of Technology, Stockholm, Sweden (e-mail: {mohamed.abdalmoaty, hakan.hjalmarsson}@ee.kth.se).

asymptotically optimal, the convergence of the expectation-maximization algorithm may be slow in some cases. To reduce the computations, consistent but suboptimal two-stage estimators based on a best linear approximation have been suggested in [21]. More recently, computationally attractive consistent estimators in a PEM framework based on suboptimal linear predictors have been proposed in [1].

In this contribution, we argue that consistent and asymptotically normal estimators of stochastic MIMO models may be obtained using suboptimal predictors in the expression of the likelihood function. It is shown that the resulting expression coincides with the true marginal likelihood function which leads to a problem involving only scalar integrals, that may be solved efficiently using numerical integration, regardless of the sample size or the number of outputs. Furthermore, we highlight the relationship between the resulting estimator and a PEM estimator based on the OE-type predictor suggested in [1]; we show that the resulting estimator is equivalent to a PEM estimator based on the OE-type predictor and a specific choice for the cost function. While the consistency and asymptotic normality of the former may be established under a weaker identifiability hypothesis (identifiability via marginal likelihood functions), its asymptotic covariance is not necessarily smaller than the OE-PEM estimator as shown by the simulation example. A full explanation of this observation will be considered in a future contribution.

The outline of the paper is as follows. In section II, we fix the notations and review the construction of the likelihood function when $\boldsymbol{w}$ is independent. Section 2 contains our main arguments, and Section IV is devoted to a numerical simulation example. The paper is concluded in Section V with a remark on wider classes of models.

II. The independent case

Suppose that the underlying system is given by the relations

$$\begin{aligned} \bar{z}_t &= G(\mathrm{q};\theta^\circ)u_t \\ \boldsymbol{z}_t &= \bar{z}_t + \boldsymbol{w}_t \\ \boldsymbol{y}_t &= f(\boldsymbol{z}_t;\theta^\circ) + \boldsymbol{v}_t, \quad t = 1,\dots,N. \end{aligned}$$

as shown in Figure 1, in which $G(\mathrm{q};\theta^\circ)$ is a $d_y \times d_u$ transfer operator matrix, q denotes the forward-shift operator on doubly infinite sequences, $\boldsymbol{y}_t \in \mathbb{R}^{d_y}$ and $u_t \in \mathbb{R}^{d_u}$ for some finite positive integers d_y and d_u. We will denote the k^{th} output at time t by $\boldsymbol{y}_t(k)$. Observe that we are using a bold font to denote random variables, and a regular font to denote realizations thereof. The input sequence $u = \{u_t\}$ is a priori known, and the measurement noise $\boldsymbol{v}$ and the disturbance $\boldsymbol{w}$ are $\mathbb{R}^{d_y}$-valued independent and mutually independent stochastic processes with known probability density functions:

$$\boldsymbol{v}_t(k) \sim p_v(\boldsymbol{v}_t(k);\theta^\circ), \quad \text{and} \quad \boldsymbol{w}_t(k) \sim p_w(\boldsymbol{w}_t(k);\theta^\circ),$$

independent over $t \in \mathbb{Z}$, and $k \in \{1,\dots,d_y\}$. The true parameter $\theta^\circ \in \Theta \subset \mathbb{R}^d$ for some finite positive integer d. Let the data be given as a column vector of stacked outputs

$$\boldsymbol{Y} := \begin{bmatrix} \boldsymbol{y}_1^\top & \boldsymbol{y}_2^\top & \dots & \boldsymbol{y}_N^\top \end{bmatrix}^\top,$$

for some finite positive integer N, and denote a given observation by Y. The ML estimation method requires the evaluation of the likelihood function $\theta \mapsto p(Y;\theta)$; the ML estimator is defined, when exists, as the random variable

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

In the class of stochastic Wiener models, the likelihood function is analytically intractable in general. However, when the process disturbance and the measurement noise are independent, as above, with known distributions, the likelihood function can be approximated either by using numerical integration methods or by running Monte Carlo simulations on the model. To see this, observe that, due to the independence assumption, the Probability Density Function (PDF) of $\boldsymbol{Y}$ factorizes as

$$p(\boldsymbol{Y};\theta) = \prod_{t=1}^{N}\prod_{k=1}^{d_y} p(\boldsymbol{y}_t(k);\theta). \tag{1}$$

For the analysis and the numerical implementation, it is more convenient to work with the negative log-likelihood function

$$-\log(p(Y;\theta)) = -\sum_{t=1}^{N}\sum_{k=1}^{d_y} \log(p(y_t(k);\theta)) \tag{2}$$

where products turn into sums, and $\log(\cdot)$ is the logarithmic function. A ML estimate is then given as a global minimizer of the negative log-likelihood.

An expression for the likelihood function of θ at $y_t(k)$ can be easily obtained by conditioning on the unobserved random variable $\boldsymbol{w}_t(k)$: if $\boldsymbol{w}_t(k)$ is given (hence $\boldsymbol{z}_t(k)$ is known), the random variable $\boldsymbol{y}_t(k)$ has a known density function, and therefore

$$\begin{aligned} p(y_t(k);\theta) &= \int_{\mathbb{R}} p_v(y_t(k)|w;\theta)p_w(w)\,\mathrm{d}w \\ &= \int_{\mathbb{R}} p_v\left(y_t(k) - f\left([G(\mathrm{q};\theta)u_t]_k + w;\theta\right)\right)p_w(w)\,\mathrm{d}w \\ &= \mathbb{E}_{\boldsymbol{w}}[p_v\left(y_t(k) - f\left([G(\mathrm{q};\theta)u_t]_k + \boldsymbol{w};\theta\right)\right)]. \end{aligned}$$

These are *scalar* marginalization integrals that can be efficiently and accurately computed, for example using the Simpson's rule with M subintervals [5] for some large positive integer M. The last expression also suggests that it can be approximated using Monte Carlo simulations [2];

$$\widehat{p(y_t(k);\theta)} = \frac{1}{M}\sum_{m=1}^{M} p_v\left(y_t(k) - f\left([G(\mathrm{q},\theta)u_t]_k + w_t^{(m)}(\theta);\theta\right)\right)$$

where $\{\boldsymbol{w}_t^m(\theta)\} \overset{\text{iid}}{\sim} p_w(\boldsymbol{w}_t;\theta)$. An approximate ML estimate is then given by

$$\hat{\theta}_{N,M} := \arg\min_{\theta\in\Theta} \; -\sum_{t=1}^{N}\sum_{k=1}^{d_y}\log\left(\widehat{p(y_t(k);\theta)}\right).$$

In either case, when the number of subinterval or Monte Carlo samples $M \to \infty$ fast enough, $N \to \infty$ at a sufficiently slower rate[1], and under standard identifiability and regularity conditions

$$\hat{\boldsymbol{\theta}}_{N,M} \xrightarrow{\text{a.s.}} \theta^\circ$$

by a uniform law of large numbers, and the estimator coincides with the MLE. Therefore, when $N, M \to \infty$ the estimator is consistent in all the cases where the MLE is consistent, and inherits all the asymptotic properties of the MLE, like statistical efficiency.

III. The colored case

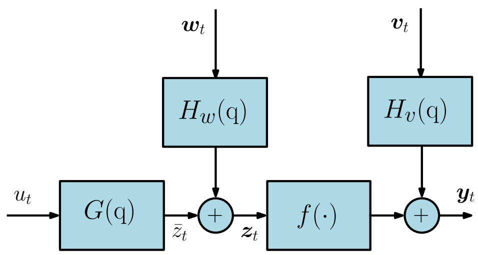

Fig. 2: Stochastic Wiener model with colored process noise

Now suppose that disturbances and measurement noise are dependent, over time and space (i.e, t and k), such that the system is given by the relations

$$\begin{aligned}\bar{z}_t &= G(\mathrm{q};\theta^\circ)u_t\\ \boldsymbol{z}_t &= \bar{z}_t + H_w(\mathrm{q};\theta^\circ)\boldsymbol{w}_t\\ \boldsymbol{y}_t &= f(z_t;\theta^\circ) + H_v(\mathrm{q};\theta^\circ)\boldsymbol{v}_t.\end{aligned} \tag{3}$$

In this case the outputs $\boldsymbol{y}_t$ are dependent; and the computations of the exact likelihood function is not an easy task. It is given by a multidimensional marginalization integral over $\mathbb{R}^{d_y N}$, where the value $d_y N$ for common applications is in the order of thousands. Let us define the column vector

$$\boldsymbol{Z} := \begin{bmatrix}\boldsymbol{z}_1^\top & \boldsymbol{z}_2^\top & \dots & \boldsymbol{z}_N^\top\end{bmatrix}^\top.$$

Then,

$$\begin{aligned}\log p(\boldsymbol{Y};\theta) &= \int_{\mathbb{R}^{d_y N}} p(\boldsymbol{Y}, Z;\theta)\,\mathrm{d}Z,\\ &= \int_{\mathbb{R}^{d_y N}} p(\boldsymbol{Y}|Z;\theta)p(Z;\theta)\,\mathrm{d}Z\end{aligned}$$

in which the integrands can be evaluated using the optimal MSE one-step ahead predictors and the known PDFs of $\boldsymbol{w}$ and $\boldsymbol{v}$. Because $\boldsymbol{z}$ is the output of a linear model, the optimal predictor is computed in a straightforward manner by inverting the disturbance model H_w; the prediction error process is defined as

$$\begin{aligned}\boldsymbol{e}_t^z(\theta) =& \boldsymbol{z}_t - \hat{\boldsymbol{z}}_{t|t-1}(\theta)\\ =& H_w^{-1}(\mathrm{q},\theta)(\boldsymbol{z}_t - G(\mathrm{q},\theta)u_t),\end{aligned} \tag{4}$$

which coincides with $\boldsymbol{w}$ at $\theta^\circ = \theta$ and therefore has the same distribution. Similarly, an optimal one-step ahead predictor of the output may be obtained by conditioning on z_t and inverting the noise model H_v; the prediction error process is then defined as

$$\begin{aligned}\boldsymbol{e}_t^y(\theta|z_t) =& \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}(\theta|z_t)\\ =& H_v^{-1}(\mathrm{q},\theta)(\boldsymbol{y}_t - f(z_t;\theta)),\end{aligned} \tag{5}$$

[1] an exact minimizer is required only asymptotically.

which coincides with $\boldsymbol{v}$ at $\theta^\circ = \theta$ and therefore has the same distribution. The joint likelihood function is thus given by the multidimensional integral

$$p(Y;\theta) = \int_{\mathbb{R}^{d_y N}} \prod_{t=1}^{N} p_v(\boldsymbol{e}_t^y(\theta|z_t)\,;\theta)\,p_w(\boldsymbol{e}_t^z(\theta)\,;\theta)\,\mathrm{d}Z. \tag{6}$$

Notice that it is not possible to interchange the integration and the product operations; in this case, to obtain the ML estimate, one has to deal with a multidimensional integration problem over a high-dimensional space.

Remark 1: The ML estimate defined as the maximizer of (6) is equivalently given by solving a PEM problem defined using a *complete probabilistic model*

$$\begin{aligned}\text{Predictor:}\quad & \hat{y}_{t|t-1}(\theta) := \mathbb{E}\left[\boldsymbol{y}_t|y_1,\dots,y_{t-1};\theta\right],\\ \text{PE process:}\quad & \boldsymbol{e}_t(\theta) = y_t - \hat{y}_{t|t-1}(\theta),\\ \text{Cost function:}\quad & -\sum_{t=1}^{N}\log p_{\boldsymbol{e}_t}(\boldsymbol{e}_t(\theta)\,;\theta)\end{aligned}$$

where $p_{\boldsymbol{e}_t}(\boldsymbol{e};\theta)$ is the PDF associated with $\boldsymbol{e}_t$. See page 216 and the first paragraph of page 217 in [11]. Note that even in cases where $f(x) = x$, and depending on the densities $p_{\boldsymbol{v}}$ and $p_{\boldsymbol{w}}$, the above predictor and associated densities $p_{\boldsymbol{\varepsilon}_t}$ are available analytically only in very few cases, and that it is generally not possible to find the form of $p_{\boldsymbol{\varepsilon}_t}$.

Fortunately, prediction error methods can be defined in many ways, and the used predictor and cost function are user choices; see [11, Section 3.3]. Recently, suboptimal predictors which are linear in the data have been suggested in [1] to bypass the computation of the likelihood function (6) in particular for cases where $\boldsymbol{w}$ is colored. These predictors lead to consistent and asymptotically normal PEM estimators and, in several relevant cases, may be given in terms of closed-form expressions. A simple predictor can be defined as the (unconditional) mean of the outputs, i.e.,

$$\hat{y}_{t|t-1}(\theta) = \mathbb{E}[\boldsymbol{y}_t;\theta],$$

which is independent of the data. In the linear case, this predictor coincides with the predictor obtained using an output-error model, and therefore it was referred to as the OE-type predictor; it averages the output $\boldsymbol{y}_t$ over all possible values of $\boldsymbol{w}_t$ under its marginal distribution. An associated PEM estimator, referred to as OE-PEM, is given by

$$\hat{\boldsymbol{\theta}}_N := \arg\min_{\theta\in\Theta} \sum_{t=1}^{N}\sum_{k=1}^{d_y}\|\boldsymbol{y}_t(k) - \mathbb{E}[\boldsymbol{y}_t(k);\theta]\|_2^2. \tag{7}$$

In a straightforward manner, the above idea may be applied to define predictor of $\boldsymbol{z}$ as well as a conditional predictor of $\boldsymbol{y}$ in (4) and (5) to get

$$\begin{aligned}\boldsymbol{\varepsilon}_t^z(\theta) =& \boldsymbol{z}_t - \hat{\boldsymbol{z}}_{t|t-1}(\theta),\\ =& \boldsymbol{z}_t - G(\mathrm{q},\theta)u_t,\end{aligned}$$

$$\begin{aligned}\boldsymbol{\varepsilon}_t^y(\theta|z_t) =& \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}(\theta|z_t),\\ =& \boldsymbol{y}_t - f(z_t;\theta),\end{aligned}$$

It is important now to observe that whenever p_v and p_w are members of location-scale families, the distribution of the prediction error processes $\varepsilon_t^z(\theta)$ and $\varepsilon_t^y(\theta|z_t)$ will have the same form as p_v and p_w but not the same parameters. This holds because, when $\theta^\circ = \theta$, the two processes coincide with $H_w(\mathrm{q})\boldsymbol{w}_t$ and $H_v(\mathrm{q})\boldsymbol{v}_t$ respectively.

The above arguments lead us to the following expression

$$\tilde{p}(Y;\theta) = \int_{\mathbb{R}^{d_y N}} \prod_{t=1}^{N} p_v(\varepsilon_t^y(\theta|z_t)\,;\theta)\, p_w(\varepsilon_t^z(\theta)\,;\theta)\,\mathrm{d}Z. \quad (8)$$

Since the t^{th} factor of the integrand is now a function of only z_t, it is possible to interchange the integration and product operations; by taking the logarithm of both sides in (8) we obtain

$$\log \tilde{p}(Y;\theta) = \sum_{t=1}^{N}\sum_{k=1}^{d_y} \log \tilde{p}(y_t(k);\theta), \quad (9)$$

a sum of $d_y N$ terms involving *only scalar integrals*; the value $\tilde{p}(y_t(k);\theta)$ is given by

$$\int p_{v(k)}(y_t(k) - f(z,\theta);\theta)\;\; p_{w(k)}(z - [G(\mathrm{q};\theta)u_t]_k;\theta)\,\mathrm{d}z.$$

Thus, we are able to use either Monte Carlo simulations or deterministic numerical integration, as shown in the previous section, to evaluate the scalar integrals and approximately solve the problem.

Relation to the OE-PEM estimator

It is straightforward to see, from the above arguments, that the quantities $\tilde{p}(y_t(k);\theta)$ are in fact the true marginal likelihood function of θ; that is, for any given feasible θ the PDFs defined by $\tilde{p}(\boldsymbol{y}_t(k);\theta)$ determine the true distributions of $\boldsymbol{y}_t(k)$ at $\theta^\circ = \theta$. Therefore, we refer to the maximizer of (9) as the marginal ML estimator.

Now observe that under the assumption that $\boldsymbol{y}$ possesses a finite mean value for all t, we may always write

$$\boldsymbol{y}_t = \mathbb{E}[\boldsymbol{y}_t;\theta] + \boldsymbol{\zeta}_t(u,\theta), \quad t \in \mathbb{Z} \quad (10)$$

in which $\boldsymbol{\zeta}_t(u,\theta)$ is an $\mathbb{R}^{d_y}$-valued random variable with zero mean and some PDF $p_{\boldsymbol{\zeta}_t}$. The process $\boldsymbol{\zeta}$ is in fact the prediction error process defined by the OE-type predictor. Let the PDF of the k^{th} entry of $\boldsymbol{\zeta}_t$ be denoted as $p_{\boldsymbol{\zeta}_t(k)}(\boldsymbol{\zeta}_t(k);\theta)$ and observe that the dependence on u and θ is not explicit in the notation. Then, the *true* marginal likelihood functions

$$\begin{aligned}\tilde{p}(y_t(k);\theta) &= p_{\boldsymbol{\zeta}_t(k)}(\zeta_t(k);\theta)\\ &= p_{\boldsymbol{\zeta}_t(k)}(\boldsymbol{y}_t(k) - \mathbb{E}[\boldsymbol{y}_t(k);\theta];\theta)\end{aligned} \quad (11)$$

Therefore the estimator defined as the maximizer of (9) is equivalently given by solving a PEM problem defined using the following *partial probabilistic model*

$$\begin{aligned}\text{OE-type predictor:}\quad & \hat{y}_{t|t-1}(\theta) := \mathbb{E}\left[\boldsymbol{y}_t;\theta\right],\\ \text{PE process:}\quad & \boldsymbol{\zeta}_t = y_t - \hat{y}_{t|t-1}(\theta),\\ \text{Cost function:}\quad & -\sum_{t=1}^{N}\sum_{k=1}^{d_y}\log\;\; p_{\boldsymbol{\zeta}_t(k)}(\boldsymbol{\zeta}_t(k);\theta)\end{aligned}$$

In other words, we obtain an OE-type PEM estimator similar to (7)—except that instead of using time- and parameter-independent Euclidean norm, the cost function here is defined using the true marginal negative log-likelihoods, $\log p_{\boldsymbol{\zeta}_t(k)}$, which are input- and parameter-dependent.

We now state our main result.

Conjecture 3.1 (consistency and asymptotic normality): Under the identifiability hypothesis

$$\tilde{p}(y_t(k);\theta) = \tilde{p}(y_t(k);\theta^\circ)\;\; \forall k, y_t \iff \theta = \theta^\circ \quad \forall \theta^\circ \in \Theta$$

and some regularity conditions, the marginal ML estimator defined as the maximum of the quantity in (9) is almost surely consistent for the parameter in (3). Furthermore, the estimator is $o(1/\sqrt{N})$ a.s. and asymptotically normal with an asymptotic covariance matrix

$$P(\theta^\circ) = [\nabla_\theta^2 Q(\theta^\circ)]^{-1} R(\theta^\circ) [\nabla_\theta^2 Q(\theta^\circ)]^{-1}$$

in which

$$\begin{aligned}Q(\theta) &= \lim_{N\to\infty}\frac{1}{N}\mathbb{E}[\log\tilde{p}(\boldsymbol{Y}_N;\theta)]\\ R(\theta^\circ) &= \mathbb{E}\left[\lim_{N\to\infty}\frac{1}{N}\nabla_\theta \log\tilde{p}(\boldsymbol{Y};\theta^\circ)[\nabla_\theta\log\tilde{p}(\boldsymbol{Y};\theta^\circ)]^\top\right]\end{aligned}$$

where the expectation is with respect to the true underlying probability measure, provided that the limits exist.

Proof outline: The main idea here is that the used objective function is a valid log-likelihood function; it is in fact given by the true marginal log-likelihood functions. Note that the expectation of the gradient vector

$$\nabla_\theta \log\tilde{p}(\boldsymbol{Y};\theta) = \sum_{t=1}^{N}\sum_{k=1}^{d_y}\nabla_\theta\log\tilde{p}(\boldsymbol{y}_t(k);\theta), \quad (12)$$

when evaluated at θ°, is zero for all $\theta^\circ \in \Theta$. To see this, observe that $\tilde{p}(\boldsymbol{y}_t(k);\theta)$ is the true PDF of $\boldsymbol{y}_t(k)$; therefore, under some regularity conditions that allow for interchanging the differentiation and integration operations, it holds that

$$\begin{aligned}\mathbb{E}[\nabla_\theta\log\tilde{p}(\boldsymbol{y}_t(k);\theta)] &= \mathbb{E}\left[\frac{1}{\tilde{p}(\boldsymbol{y}_t(k);\theta)}\nabla_\theta\tilde{p}(\boldsymbol{y}_t(k);\theta)\right]\\ &= \int \nabla_\theta\tilde{p}(\tilde{y};\theta)\,\mathrm{d}\tilde{y}\\ &= \nabla_\theta\int\tilde{p}(\tilde{y};\theta)\,\mathrm{d}\tilde{y} = 0 \;\;\forall\theta\in\Theta\end{aligned}$$

where $\mathbb{E}$ is with respect to $\tilde{p}(\boldsymbol{y}_t(k);\theta)$.

Now observe that in case we established that (12) behaves like its expectation as $N \to \infty$, and assuming that Θ constitutes an identifiable parameterization via the marginals, it follows that the estimator is $\sqrt{N}$-consistent and asymptotically normal. For this, it is sufficient, for example, to prove a dominance condition:

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\tilde{p}(\boldsymbol{y}_t;\theta)\right] < \infty,$$

which may be achieved using a set of regularity conditions such as continuity and sufficient smoothness of $\theta \mapsto \tilde{p}(Y;\theta)$ for every Y, compactness of Θ, uniform exponential stability of the predictors, and exponential stability of the data (see [10], [12] for example). ∎

IV. NUMERICAL EXAMPLE

In this section, we demonstrate the consistency and asymptotic normality of the marginal ML estimator on the stochastic Wiener model depicted in Figure 3. The model has two

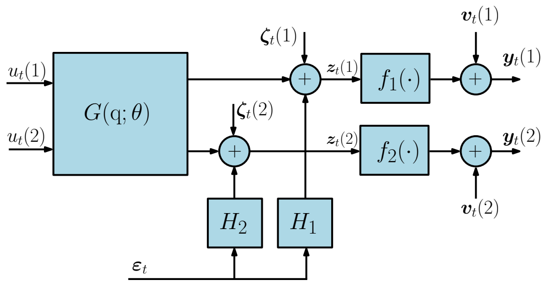

Fig. 3: 2-inputs 2-outputs stochastic Wiener model with colored process noise

inputs and two outputs related by the equations

$$\begin{aligned}
\boldsymbol{y}_t(1) &= f_1(\boldsymbol{z}_t(1);\theta) + \boldsymbol{v}_t(1),\\
\boldsymbol{z}_t(1) &= G_{11}(\mathrm{q};\theta)u_t(1) + G_{12}(\mathrm{q};\theta)u_t(2) + \boldsymbol{w}_t(1),\\
\boldsymbol{w}_t(1) &= H_1(\mathrm{q})\varepsilon_t + \boldsymbol{\zeta}_t(1),\\
\boldsymbol{y}_t(2) &= f_2(\boldsymbol{z}_t(2);\theta) + \boldsymbol{v}_t(2),\\
\boldsymbol{z}_t(2) &= G_{21}(\mathrm{q};\theta)u_t(1) + G_{22}(\mathrm{q};\theta)u_t(2) + \boldsymbol{w}_t(2),\\
\boldsymbol{w}_t(2) &= H_2(\mathrm{q})\varepsilon_t + \boldsymbol{\zeta}_t(2), \quad t = 1,2,3,\dots
\end{aligned}$$

in which the transfer operators

$$\begin{aligned}
G_{ij}(\mathrm{q},\theta) &= \frac{b_{ij}}{1 + a_{ij}\mathrm{q}^{-1}}, \quad i,j \in \{1,2\},\\
H_1(\mathrm{q}) &= \frac{0.7}{1 - 0.2\mathrm{q}^{-1} + 0.8\mathrm{q}^{-2}},\\
H_2(\mathrm{q}) &= \frac{0.75 - 1.2\mathrm{q}^{-1}}{1 - 0.2\mathrm{q}^{-1} + 0.5\mathrm{q}^{-2}},
\end{aligned}$$

and the nonlinear functions, representing a saturation at the output, are given by

$$f_1(x;\theta) = f_2(x;\theta) = \frac{L}{1 + \exp(-kx)} - \frac{L}{2}.$$

The inputs $u(1)$ and $u(2)$ are known independent realizations of standard Gaussian processes. The noise processes $\boldsymbol{v}(1)$, $\boldsymbol{v}(2)$, and the disturbance $\boldsymbol{\varepsilon}$ are independent and mutually independent stationary Gaussian processes with zero mean and variance 0.1. The disturbances $\boldsymbol{\zeta}(1)$ and $\boldsymbol{\zeta}(2)$ are independent and mutually independent stationary Gaussian process with zero mean and variances 4.9 and 3.6 respectively; they are independent of $\boldsymbol{v}(1)$, $\boldsymbol{v}(2)$ and $\boldsymbol{\varepsilon}$. Therefore, the process disturbances $\boldsymbol{w}_t(1)$ and $\boldsymbol{w}_t(2)$ are zero mean stationary processes with variances $\lambda_{w(1)} = 5.04$ and $\lambda_{w(2)} = 3.84$ (rounded to the second decimal digit).

For this model, the process disturbances are clearly colored and dependent; thus, the outputs of the model are colored and dependent as well. The parameters of the models are constrained to ensure stability: $|a_{ij}| < 1,\ i,j \in \{1,2\}$ and the value k in the nonlinearity is assumed known and is fixed to 1 so that the parameterization is identifiable. The true parameter is fixed to

$$\begin{aligned}
\theta^\circ &= \begin{bmatrix} a_{11} & a_{12} & a_{21} & a_{22} & b_{11} & b_{12} & b_{21} & b_{22} & L \end{bmatrix}\\
&= \begin{bmatrix} 0.5 & 0.1 & -0.8 & 0.7 & 4.9 & 2.1 & -1.2 & 3 & 2 \end{bmatrix}.
\end{aligned}$$

These choices lead to a difficult problem where the disturbances saturate outputs which would otherwise be in the (almost) linear part of the nonlinearity.

We conducted a Monte Carlo simulation using 1000 data sets, corresponding to different realizations of the disturbances and noise, for increasing values of N ranging from 200 to 1600. The parameter vector θ was estimated using three different methods: the marginal ML studied in this paper, the OE-PEM method proposed in [1], and a PEM ignoring $\boldsymbol{w}$ completely. The problems were initialized at θ° to avoid possible local solutions. The results are reported in Table I and Figures 4 and 5.

TABLE I: The mean value and the standard deviations of the three estimators approximated based on 1000 Monte Carlo simulation when $N = 1200$.

True	Marginal ML	OE-PEM	PEM ignoring $\boldsymbol{w}$
$a_{11} = 0.5$	0.499 ± 0.015	0.499 ± 0.016	0.500 ± 0.017
$a_{12} = 0.1$	0.097 ± 0.053	0.098 ± 0.057	0.103 ± 0.057
$a_{21} = -0.8$	-0.798 ± 0.016	-0.798 ± 0.018	-0.799 ± 0.020
$a_{22} = 0.7$	0.700 ± 0.011	0.700 ± 0.012	0.699 ± 0.012
$b_{11} = 4.9$	4.943 ± 0.373	4.913 ± 0.286	2.833 ± 0.186
$b_{12} = 2.1$	2.114 ± 0.194	2.100 ± 0.164	1.214 ± 0.108
$b_{21} = -1.2$	-1.205 ± 0.109	-1.201 ± 0.098	-0.748 ± 0.068
$b_{22} = 3$	3.001 ± 0.209	2.994 ± 0.172	1.879 ± 0.115
$L = 2$	1.999 ± 0.022	1.991 ± 0.026	2.033 ± 0.030
$\lambda_{w(1)} = 5.04$	5.081 ± 0.772	5.032 ± 0.103	not estimated
$\lambda_{w(2)} = 3.84$	3.822 ± 0.508	3.840 ± 0.043	not estimated

The simulation results indicate the consistency and asymptotic normality of the marginal ML estimator; while the OE-PEM estimator performs better on the current example.

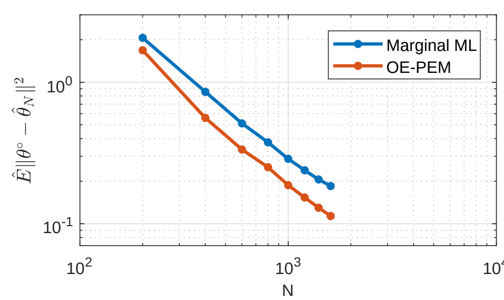

Fig. 4: The MSE of the marginal ML and the OE-PEM estimators defined in (9) and (7) for different values of N.

V. CONCLUDING REMARKS

We pointed out that a consistent estimator for stochastic Weiner models can, under some regularity and identifiability conditions, be constructed as the maximizer of a product of *scalar integrals* that may be solved efficiently using numerical integration techniques. Therefore, the estimator enjoys a computational advantage compared to the ML estimator, which requires the evaluation of a multidimensional integral over a high-dimensional domain. The price to pay is a loss of statistical asymptotic efficiency. However, as is well known (see [9]), efficient estimators are not unique; and efficiency can be retrieved by a single step of a Newton-Raphson scheme. This requires the approximation of the exact score function, for example using *only a single run* of any particle smoothing algorithms.

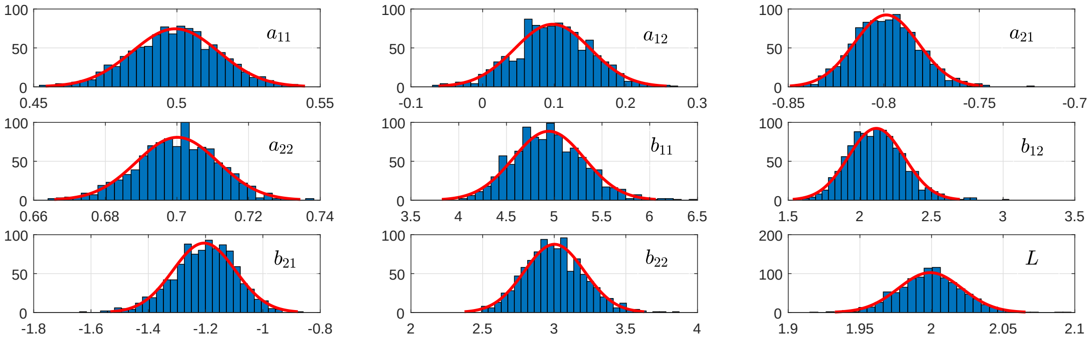

Fig. 5: Unnormalized histograms of 1000 independent realizations of the marginal ML estimator with fitted Gaussian PDFs when $N = 1200$; the results indicate that the estimator is asymptotically normal about the true parameter.

It is of interest to observe that the ideas discussed here are not specific to stochastic Wiener models. In fact, they may be used to construct consistent estimators for fairly general stochastic nonlinear models. Consider for example a nonlinear state-space model

$$\begin{aligned} \boldsymbol{x}_{t+1} &= h(\boldsymbol{x}_t, u_t; \theta) + \boldsymbol{w}_t \\ \boldsymbol{y}_t &= g(\boldsymbol{x}_t, u_t; \theta) + H(\mathrm{q}; \theta)\boldsymbol{v}_t \end{aligned}$$

in which $\boldsymbol{w}_t \sim p_{w_t}(\boldsymbol{w}_t; \theta)$, and $\boldsymbol{v}_t \sim p_{v_t}(\boldsymbol{v}_t; \theta)$. Then, a consistent estimator of θ may be obtained by maximizing

$$\sum_{t=1}^{N} \log \int p_{v_t}((\boldsymbol{y}_t - g(\tilde{x}, u_t; \theta); \theta) p_{x_t}(\tilde{x} - \mathbb{E}[\boldsymbol{x}_t; \theta]; \theta) \, \mathrm{d}\tilde{x}.$$

A special case is when

$$\begin{aligned} x_{t+1} &= h(x_t, u_t; \theta) \\ \boldsymbol{z}_t &= x_t + H_w(\mathrm{q}; \theta)\boldsymbol{w}_t \\ \boldsymbol{y}_t &= g(\boldsymbol{z}_t, u_t; \theta) + H_v(\mathrm{q}; \theta)\boldsymbol{v}_t. \end{aligned}$$

This case covers Hammerstein-Wiener models when the model h is static and reduces to (3) if h is static and linear.

One advantage of the proposed estimator in comparison to those in [1] is a relaxed identifiability hypothesis: instead of requiring identifiability via the first (two) moment(s), it only requires identifiability via all marginal moments. However, their asymptotic properties (covariance) is not necessarily better than the OE-PEM estimator.

References

[1] M. R. Abdalmoaty. *Learning Stochastic Nonlinear Dynamical Systems Using Non-stationary Linear Predictors*. Licentiate thesis, KTH Royal Institute of Technology, Automatic Control Department, 2017.

[2] M. R. Abdalmoaty and H. Hjalmarsson. A simulated maximum likelihood method for estimation of stochastic Wiener systems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 3060–3065, Dec 2016.

[3] S. Boyd and L. Chua. Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems*, 32(11):1150–1161, Nov 1985.

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

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

[6] A. Hagenblad, L. Ljung, and A. Wills. Maximum likelihood identification of Wiener models. *IFAC Proceedings Volumes*, 41(2):2714 – 2719, 2008. 17th IFAC World Congress.

[7] W. I. Hunter and M. J. Korenberg. The identification of nonlinear biological systems: Wiener and Hammerstein cascade models. *Biol. Cybern.*, 55(2-3):135–144, Nov. 1986.

[8] A. Kalafatis, N. Arifin, L. Wang, and W. Cluett. A new approach to the identification of pH processes based on the Wiener model. *Chemical Engineering Science*, 50(23):3693 – 3701, 1995.

[9] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, 2011.

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

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

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

[13] A. Marconato, J. Sjöberg, J. Suykens, and J. Schoukens. Separate initialization of dynamics and nonlinearities in nonlinear state-space models. In *2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, pages 2104–2108, 2012.

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

[15] 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.

[16] M. Schoukens. *Identification of parallel block-oriented models starting from the best linear approximation*. PhD thesis, Vrije Universiteit Brussel, 2015.

[17] M. Schoukens and K. Tiels. Identification of nonlinear block-oriented systems starting from linear approximations: A survey. *CoRR*, abs/1607.01217, 2016.

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

[19] J. Sjöberg, L. Lauwers, and J. Schoukens. Identification of Wiener-Hammerstein models: Two algorithms based on the best split of a linear model applied to the SYSID'09 benchmark problem. *Control Engineering Practice*, 20(11):1119 – 1125, 2012. Special Section: Wiener-Hammerstein System Identification Benchmark.

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

[21] B. Wahlberg, J. Welsh, and Lennart. Identification of stochastic Wiener systems using indirect inference. *IFAC-PapersOnLine*, 48(28):620 – 625, 2015. 17th IFAC Symposium on System Identification SYSID 2015 Beijing, China, 1921 October 2015.

[22] 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.

[23] A. Wills, T. B. Schön, L. Ljung, and B. Ninness. Blind identification of Wiener models. *IFAC Proceedings Volumes*, 44(1):5597 – 5602, 2011. 18th IFAC World Congress.

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

[25] Y. Zhu. Distillation column identification for control using Wiener model. In *American Control Conference, 1999. Proceedings of the 1999*, volume 5, pages 3462–3466 vol.5, 1999.