Automatica 49 (2013) 70–81

Contents lists available at SciVerse ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Identification of Hammerstein–Wiener models[☆]

Adrian Wills[a,1], Thomas B. Schön[b], Lennart Ljung[b], Brett Ninness[a]

[a] School of Electrical Engineering and Computer Science, University of Newcastle, Callaghan, NSW 2308, Australia

[b] Division of Automatic Control, Linköping University, SE-581 83 Linköping, Sweden

ARTICLE INFO

Article history:
Received 12 September 2011
Received in revised form
18 April 2012
Accepted 31 July 2012
Available online 25 October 2012

Keywords:
System identification
Hammerstein
Wiener
Block-oriented models
Nonlinear models
Dynamic systems
Monte Carlo method
Smoothing
Expectation maximization algorithm
Particle methods
Maximum Likelihood

ABSTRACT

This paper develops and illustrates a new maximum-likelihood based method for the identification of Hammerstein–Wiener model structures. A central aspect is that a very general situation is considered wherein multivariable data, non-invertible Hammerstein and Wiener nonlinearities, and colored stochastic disturbances both before and after the Wiener nonlinearity are all catered for. The method developed here addresses the blind Wiener estimation problem as a special case.

1. Introduction

A useful and general class of nonlinear dynamical models are so-called *block-oriented models* that consist of configurations of *linear dynamic* blocks and *nonlinear memoryless* blocks. The simplest examples in this class are cascaded systems with the nonlinear block either preceding (*Hammerstein model*) or following (*Wiener model*) the linear block. The Hammerstein model was apparently first discussed in Narendra and Gallman (1966), while the Wiener model has its roots in Wiener's interest in a nonlinear system using Volterra expansions (Schetzen, 1980; Wiener, 1942).

The model where a nonlinear block both precedes and follows a linear dynamic system is called a *Hammerstein–Wiener model*. This is illustrated diagrammatically in Fig. 1. More recently, generalisations based on feedback variants have been studied, such as the work (Hsu, Vincent, & Poolla, 2006; Schoukens, Nemeth, Crama, Rolain, & Pintelon, 2003).

The literature on how to estimate the Hammerstein–Wiener model (and the Hammerstein or Wiener only special cases) is extensive indeed, as evidenced by the selection (Bai, 2002a; Billings & Fakhouri, 1982; Giri & Bai, 2010; Kalafatis, Wang, & Cluett, 1997; Raich, Zhou, & Viberg, 2005; Westwick & Verhaegen, 1996; Wigren, 1993) and their bibliographies. In relation to this, it is important to emphasize that the work here is distinguished from these and other previous contributions in the following ways.

First, the models here are fully multivariable in that all signals passing between all linear and nonlinear blocks in Fig. 1 may not only be multivariable, but may be of differing dimensions.

Second, the memoryless nonlinear blocks f_H and f_W illustrated in Fig. 1 may be of very general form. For example, they need not be invertible as is often required in pre-existing literature.

Finally, the models considered here allow for a stochastic disturbance *before* the final Wiener nonlinearity f_W. This is illustrated as the signal v_t in Fig. 1. This point is significant, since in the absence of v_t, the model is essentially an output-error one for which standard estimation methods are well

[☆] This work was supported by: the Australian Research Council through their Discovery Project Program; and the project Calibrating Nonlinear Dynamical Models (Contract number: 621-2010-5876) and CADICS, a Linneaus Center funded by the Swedish Research Council and also funded by the Swedish Research Council. The material in this paper was partially presented at the 18th IFAC World Congress, August 28–September 2, 2011, Milan, Italy. This paper was recommended for publication in revised form by Associate Editor (System Parameter Estimation) Er-Wei Bai under the direction of Editor Torsten Söderström.

E-mail addresses: Adrian.Wills@newcastle.edu.au (A. Wills), schon@isy.liu.se (T.B. Schön), ljung@isy.liu.se (L. Ljung), Brett.Ninness@newcastle.edu.au (B. Ninness).

[1] Tel.: +61 2 49216028; fax: +61 2 49216993.

0005-1098/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2012.09.018

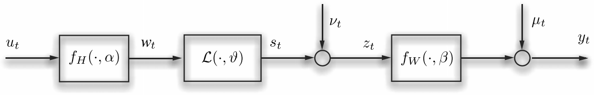

Fig. 1. The general Hammerstein–Wiener model structure, which consists of sandwiching a linear time invariant system $\mathcal{L}$ between memoryless nonlinearities f_H and f_W.

established. However, the presence of ν_t significantly complicates the estimation problem due to the difficulty of computing the influence of f_W on it.

Furthermore, in this paper ν_t may be a linearly correlated (colored) process, as may the stochastic disturbance μ_t shown if Fig. 1. Importantly, by allowing ν_t to be colored it may capture noise entering "internally" to the linear component $\mathcal{L}$, which can be necessary for accurate modeling (Ljung & Wills, 2010).

It was established in Hagenblad, Ljung, and Wills (2008) that ignoring ν_t when it is present so that a simple output error solution can be employed typically gives a biased estimate, and it was then shown how a maximum likelihood method in case ν_t is white can be used to obtain unbiased estimates. That treatment was extended in Wills and Ljung (2010) for the scalar signal case to a practical maximum likelihood method for ν_t of general color.

This paper also adopts a maximum likelihood approach, and employs two main tools. The problem of computing the effect of the Wiener nonlinearity on the noise ν_t will be addressed by using particle filtering and smoothing techniques. This allows the formulation of the appropriate likelihood, and in order to compute as estimate a local maximizer, the second main tool is adopted. Namely, the expectation maximization (EM) algorithm.

Finally, it is important to note that the exogenous input u_t may be absent in the model shown in Fig. 1 so that since ν_t may be colored, the techniques developed here provide a solution to the blind Wiener estimation problem, which has also attracted significant interest (Abed-Meraim, Qiu, & Hua, 1997; Bai, 2002b; Vanbeylen, Pintelon, & Schoukens, 2009; Wills, Schön, Ljung, & Ninness, 2011).

2. Problem formulation and model structure

This paper addresses the problem of using N-point data measurements of input $U_N \triangleq \{u_1, \ldots, u_N\}$ and output $Y_N \triangleq \{y_1, \ldots, y_N\}$ to estimate a coefficient vector θ that parametrizes a block nonlinear structure modeling these observations.

The particular model structure considered is illustrated in Fig. 1, and may be expressed as

$$y_t = f_W(z_t, \beta) + \mu_t, \tag{1}$$

$$z_t = \mathcal{L}(w_t, \vartheta) + \nu_t, \tag{2}$$

$$w_t = f_H(u_t, \alpha). \tag{3}$$

Here, $f_H(\cdot, \alpha)$ and $f_W(\cdot, \beta)$ are memoryless nonlinearities that are respectively parametrized by vectors $\alpha \in \mathbf{R}^{n_\alpha}$ and $\beta \in \mathbf{R}^{n_\beta}$, while $\mathcal{L}(\cdot, \vartheta)$ is a linear time-invariant system parametrized by $\vartheta \in \mathbf{R}^{n_\vartheta}$. The terms μ_t and ν_t are zero mean stationary stochastic processes modeling measurement and modeling errors.

This represents a Hammerstein–Wiener model structure. It is particularly general in that it allows for a correlated noise term ν_t preceding the Wiener nonlinearity $f_W(\cdot, \beta)$. Furthermore, all signals may be multivariable with

$$u_t \in \mathbf{R}^{n_u}, \qquad w_t \in \mathbf{R}^{n_w}, \qquad z_t \in \mathbf{R}^{n_z}, \qquad y_t \in \mathbf{R}^{n_y} \tag{4}$$

and the dimensions of μ_t, ν_t being conformal to those of y_t and z_t.

The linear dynamics $\mathcal{L}(\cdot, \vartheta)$ are modeled by the state space structure

$$\begin{bmatrix} x_{t+1} \\ s_t \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x_t \\ w_t \end{bmatrix} \tag{5}$$

with $\vartheta \in \mathbf{R}^{n_\vartheta}$ denoting a vector containing the non-constrained elements of the system matrices A, B, C, D.

Likewise, the correlation structure of the stationary processes ν_t and μ_t are also modeled via a state space structure

$$\xi_{t+1} = A_\xi \xi_t + \upsilon_t, \tag{6a}$$

$$\nu_t = C_\nu \xi_t, \tag{6b}$$

$$\mu_t = C_\mu \xi_t + e_t, \tag{6c}$$

where

$$\begin{bmatrix} \upsilon_t \\ e_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \right). \tag{7}$$

In what follows, these noise models are "fully" parametrized in that no elements in the matrices A_ξ, C_ν and C_μ specifying them are constrained. The matrices Q and R are also "fully" parametrized, but assumed to be symmetric and positive definite. We denote by $\lambda \in \mathbf{R}^{n_\lambda}$ a vector containing them.

Finally, the Hammerstein $f_H(\cdot, \alpha)$ and Wiener $f_W(\cdot, \beta)$ memoryless nonlinearities may be quite general. They need not be invertible, but it is required that the derivatives

$$\frac{\partial}{\partial \alpha} f_H(\cdot, \alpha), \qquad \frac{\partial}{\partial \beta} f_W(\cdot, \beta) \tag{8}$$

with respect to their parameter vectors α and β exist. This is satisfied by many common examples, such as deadzone, saturation, polynomial and piecewise linear descriptions.

The combined linear, nonlinear, and noise descriptions comprising the model structure (1)–(3) are therefore parametrized by the vector

$$\theta = [\vartheta^T, \lambda^T, \alpha^T, \beta^T]^T. \tag{9}$$

3. Maximum likelihood estimation

This paper examines the formation of an estimate $\widehat{\theta}$ of θ via the maximum likelihood (ML) approach

$$\widehat{\theta} = \arg\max_{\theta} L_\theta(Y_N), \qquad L_\theta(Y_N) \triangleq \log p_\theta(Y_N). \tag{10}$$

Here $p_\theta(Y_N)$ denotes the joint density of the measurements Y_N and via subscript makes explicit that according to the model (1)–(3) it will depend upon θ, and likewise for $L_\theta(Y_N)$.

Via Bayes' rule, the log-likelihood can be expressed as

$$L_\theta(Y_N) = \sum_{t=1}^{N} \log p_\theta(y_t \mid Y_{t-1}), \qquad p_\theta(y_1 \mid Y_0) \triangleq p_\theta(y_1). \tag{11}$$

This provides a means for evaluating the criterion $L_\theta(Y_N)$ if the prediction density $p_\theta(y_t \mid Y_{t-1})$ can be computed.

If all stochastic components appeared additively *after* the Wiener nonlinearity f_W, then the prediction density could be straightforwardly obtained via a Kalman filter.

However, our model structure is more general in that it allows for the noise term ν_t preceding the Wiener nonlinearity, with the penalty that evaluating $p_\theta(y_t \mid Y_{t-1})$ is then a serious challenge.

Recently developed sequential importance sampling or "particle filter" methods (Djurić & Godsill, 2002; Doucet & Johansen, 2011) offer a potential solution for computing (approximately) the required prediction density. Unfortunately, the resulting approximations of $p_\theta(y_t \mid Y_{t-1})$ are not differentiable (or even necessarily continuous) with respect to θ. Therefore, computing the maximizer $\widehat{\theta}$ is complicated, since standard gradient-based search techniques cannot be used.

To address this difficulty, the work here employs the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; McLachlan & Krishnan, 2008) to compute the maximizer $\widehat{\theta}$, since this technique avoids the need to compute $L_\theta(Y_N)$ or its derivatives. Sequential importance sampling methods are still employed, but critically this is by way of using particle *smoothers* as opposed to particle filters.

The reader is referred to the comprehensive monograph on the topic (McLachlan & Krishnan, 2008), and the previous works (Gibson & Ninness, 2005; Schön, Wills, & Ninness, 2011; Wills, Ninness, & Gibson, 2009) for an introduction and explanation of the EM algorithm. Central to the approach is the employment of so-called "incomplete data" X and a given fixed value $\theta = \theta_k$ to decompose the log-likelihood using Bayes' rule as

$$L_\theta(Y_N) = \mathcal{Q}(\theta, \theta_k) - \mathcal{V}(\theta, \theta_k) \tag{12}$$

where

$$\mathcal{Q}(\theta, \theta_k) \triangleq \int \log p_\theta(Y_N, X)\, p_{\theta_k}(X \mid Y_N)\, \mathrm{d}X \tag{13}$$

$$\mathcal{V}(\theta, \theta_k) \triangleq \int \log p_\theta(X \mid Y_N)\, p_{\theta_k}(X \mid Y_N)\, \mathrm{d}X. \tag{14}$$

The choice of the incomplete data X is a key design variable in the implementation of the EM-algorithm, and will be discussed in detail presently.

The resulting function $\mathcal{Q}(\theta, \theta_k)$ acts as a local (about θ_k) approximant of $L_\theta(Y_N)$. The EM algorithm seeks a maximizer of $L_\theta(Y_N)$ by computing and seeking maximizers of $\mathcal{Q}(\theta, \theta_k)$ as follows:

Algorithm 1 : Expectation Maximization Algorithm

```
1. Set $k = 0$ and initialize $\theta_0$ such that $L_{\theta_0}(Y_N)$ is finite.
2. **Expectation (E) step:** Compute
       $\mathcal{Q}(\theta, \theta_k) = \mathbf{E}_{\theta_k}\{\log p_\theta(Y_N, X) \mid Y_N\}.$    (15)
3. **Maximization (M) step:** Compute
       $\theta_{k+1} = \arg\max_\theta \mathcal{Q}(\theta, \theta_k).$    (16)
4. If not converged, update $k := k + 1$ and return to step 2.
```

The evaluation of $\mathcal{Q}(\theta, \theta_k)$ can be thought of as a smoothing step since it involves computing an expectation conditional on the whole observations sequence Y_N. In what follows, this smoothing will be approximately computed using a particle smoothing approach.

Crucially, this depends on the point θ_k around which $L_\theta(Y_N)$ is being approximated, which is fixed. The dependency of $\mathcal{Q}(\theta, \theta_k)$ on θ then arises via differentiable functional forms of the smoothed quantities, and this facilitates the maximization of $\mathcal{Q}(\theta, \theta_k)$ via gradient based search.

3.1. Convergence of the maximum-likelihood estimate

The main topic of this paper is to suggest an algorithm that finds the θ_N^{ML} that maximizes (11) for each N. That will involve a number of steps like EM (15)–(16) and particle filters to be defined below (32)–(34). As a result we cannot guarantee that we find θ_N^{ML}. This is basically the same situation as when (11) can be explicitly maximized by gradient methods and we cannot guarantee that we do not end up in a local maximum. Nevertheless it is if of interest to establish what are the properties of the sought estimate θ_N^{ML} as a function of N.

The maximum-likelihood (ML) framework employed here has historically been greatly favored due its general statistical efficiency; i.e. consistency with variability approaching the Cramér–Rao bound as the amount of available measurements N increases (Caines, 1988; Ljung, 1999).

However, these attractive theoretical properties do not apply generically, as evidenced by counter-example (Ninness, 1998). A rigorous analysis of the stochastic convergence of the ML estimate proposed here requires establishing detailed moment bounds on various signals and their derivatives together with checking detailed technical conditions on model structure parametrization. See, for example, the work (Bauer & Ninness, 1999) where the consistency of the Hammerstein–Wiener model estimates derived from a least-squares criterion is studied by application of the stochastic convergence framework developed in Pötscher and Prucha (1997).

Nevertheless, the convergence properties of a very general class of estimation methods, including the ML one proposed here, have been established in works such as Ljung (1978) and Heunis (1988). For example, the essential conditions required to apply the results of Ljung (1978) are first that the true system is exponentially stable Ljung (1978, condition (S3)).

Second, that the log-likelihood criterion (11) can be expressed as

$$L_\theta(Y_N) = h\left(\frac{1}{N}\sum_{t=1}^{N} \ell(t, \theta, \epsilon_t(\theta))\right) \tag{17}$$

where $h(\cdot)$ and $\ell(t, \theta, \cdot)$ are functions satisfying mild growth conditions (Ljung, 1978, conditions (C1–C3)) and the "prediction error" $\epsilon_t(\theta)$ is defined as

$$\epsilon_t(\theta) = y_t - \widehat{y}_{t|t-1}(\theta) \tag{18}$$

for some predictor $\widehat{y}_{t|t-1}(\theta)$ based on the model structure (1)–(3).

Third, that the predictor $\widehat{y}_{t|t-1}(\theta)$ has an exponentially decaying dependence on past data (Ljung, 1978, condition (M1)) for all parameter values $\theta \in \Theta$ with the latter being some compact set. The results of Ljung (1978) then establish the convergence

$$\lim_{N\to\infty} \widehat{\theta} \in \{\theta \in \Theta : \mathcal{L}(\theta) \geq \mathcal{L}(\beta), \forall \beta \in \Theta\} \tag{19}$$

with probability one where

$$\mathcal{L}(\theta) = \lim_{N\to\infty} \mathbf{E}\{L_\theta(Y_N)\}. \tag{20}$$

Furthermore, if there exists a set of parameter values $\Theta_\circ$ whose members are "true" in that sense that $\epsilon_t(\theta_\circ), \theta_\circ \in \Theta_\circ$ becomes an innovations process satisfying

$$\mathbf{E}\{\epsilon_t(\theta_\circ) \mid \epsilon_{t-1}(\theta_\circ), \epsilon_{t-2}(\theta_\circ), \ldots\} = 0 \tag{21}$$

then

$$\lim_{N\to\infty} \widehat{\theta} \in \Theta_\circ. \tag{22}$$

Let us comment on how to establish the conditions that guarantee this result.

Stability of the system (S3): We assume that the true system is given by (1)–(7) for a θ_0 that gives stable eigenvalues of A and A_ξ. Then we can define the output $y_s^0(t)$ that would be obtained by the true system if v_t and e_t are zero prior to time s. Clearly this would differ from the actual output $y(t)$ by an exponentially decaying (in $t - s$) amount. If the fourth order moments of v and e exist, and f_W is such that also the fourth order moment of the output exists, condition S3 of Ljung (1978) is satisfied.

Smoothness of the criterion function (C1–C3): It is well known (e.g. Lemma 5.1 with discussion in Ljung (1998)) how the ML criterion in terms of joint probabilities can be rewritten in the general form (17) by repeated application of Bayes' rule:

$$L_\theta(Y_N) = \sum_{t=1}^{N} \log p_\theta(y_t \mid Y_{t-1}) = \sum_{t=1}^{N} \ell(t, \theta, \epsilon_t(\theta)) \tag{23}$$

where $\ell(t, \theta, \epsilon_t(\theta))$ is the log of the conditional pdf of the innovations (given past data). We can thus take $h(x) = x/N$ and this $\ell(\cdot, \cdot, \cdot)$ in (2.23)–(2.24) in Ljung (1978) so that conditions C1–C3 reduce to smoothness conditions on the log of the conditional density of the innovations. Now, we do not have any closed form expression for this pdf, since the innovations are formed from e and v and also the Wiener nonlinearity. That is why we will use particle methods to handle the posterior densities, which is the main motivation for this paper. But we only need to establish condition C1, that requires the log of the innovations pdf to be differentiable wrt θ and show limited growth as a function of ϵ_t. These should be rather weak restrictions.

Smoothness and stability of the predictor function (M1): M1 requires that the predictors are differentiable wrt θ and that the influence of observations $y(s)$ in the remote past on the current prediction $\hat{y}(t|\theta)$ is exponentially decaying. Since we have no closed form expression for the predictors in this nonlinear setting, it is difficult to establish this formally. But the smoothness assumptions on nonlinearity models (8) and the linearly parameterized state space model make it reasonable that this is inherited by the predictors. Likewise, the only dependence of the past in the models follows from the exponentially stable linear model (5)–(6) so it is reasonable that the predictors must depend on past observations to an exponentially decreasing degree.

4. Computing $\mathcal{Q}(\theta, \theta_k)$

The function $\mathcal{Q}(\theta, \theta_k)$ is completely determined by the choice of the incomplete data X. In general, a sensible choice for X is a set of measurements that, while not available, would greatly simplify the estimation problem.

In previous work (Gibson & Ninness, 2005; Schön et al., 2011; Wills et al., 2009), the utility of choosing X as the time history of the full state vector of the underlying dynamics has been established. However, in this paper, the particulars of the Hammerstein Wiener structure lead to a different choice.

This involves noting that since the input u_t is assumed observed, if the noise $v_t = C_v\xi_t$ were known, then for a given fixed θ_k, the input $z_t = \mathcal{L}(w_t, \vartheta) + v_t$ to the Wiener nonlinearity f_W would also be known. Furthermore, if the state ξ_t where known, then this would allow the likelihood $L_\theta(Y_N)$ to be simply computed by noting that the density of y_t in this case is simply the density (7) of e_t evaluated at

$$\varepsilon_t \triangleq y_t - f_W(z_t, \beta) - C_\mu\xi_t \tag{24}$$

where

$$z_t = \mathcal{L}(w_t, \vartheta) + C_v\xi_t. \tag{25}$$

With this as motivation, this paper examines the incomplete data choice of

$$X \triangleq [\xi_1, \xi_2, \dots, \xi_N]. \tag{26}$$

This leads to the formulation of $\mathcal{Q}(\theta, \theta_k)$ according to the following lemma.

Lemma 4.1. *Assume that $p_\theta(\xi_1)$ does not depend on θ, but instead it is a fixed and known distribution. Then neglecting any additive constants, the choice (26) for the incomplete data implies*

$$\begin{aligned}-2\mathcal{Q}(\theta, \theta_k) &= N\log|Q| + N\log|R| + \mathrm{Tr}\left\{R^{-1}\Upsilon\right\}\\ &\quad + \mathrm{Tr}\left\{Q^{-1}\left[\Phi - \Psi A_\xi^T - A_\xi\Psi^T + A_\xi\Sigma A_\xi^T\right]\right\}\end{aligned} \tag{27}$$

with

$$\Phi \triangleq \sum_{t=1}^{N-1} \mathbf{E}_{\theta_k}\left\{\xi_{t+1}\xi_{t+1}^T \mid Y_N\right\}, \tag{28}$$

$$\Psi \triangleq \sum_{t=1}^{N-1} \mathbf{E}_{\theta_k}\left\{\xi_{t+1}\xi_t^T \mid Y_N\right\}, \tag{29}$$

$$\Sigma \triangleq \sum_{t=1}^{N-1} \mathbf{E}_{\theta_k}\left\{\xi_t\xi_t^T \mid Y_N\right\}, \tag{30}$$

$$\Upsilon \triangleq \sum_{t=1}^{N} \mathbf{E}_{\theta_k}\left\{\varepsilon_t\varepsilon_t^T \mid Y_N\right\}. \tag{31}$$

Proof. By Bayes' rule, the Markov property of the noise model (6) and the definition (13)

$$\begin{aligned}\mathcal{Q}(\theta, \theta_k) &= \mathbf{E}_{\theta_k}\{\log p_\theta(X) + \log p_\theta(Y_N|X) \mid Y_N\}\\ &= \sum_{t=1}^{N-1} \mathbf{E}_{\theta_k}\{\log p_\theta(\xi_{t+1}|\xi_t) \mid Y_N\}\\ &\quad + \mathbf{E}_{\theta_k}\{\log p_\theta(\xi_1) \mid Y_N\} + \sum_{t=1}^{N} \mathbf{E}_{\theta_k}\{\log p_\theta(y_t|\xi_t) \mid Y_N\}.\end{aligned}$$

Again using the formulation (6), the Gaussian assumptions, and neglecting additive constants (and this includes $p_\theta(\xi_1)$)

$$\begin{aligned}-2\mathcal{Q}(\theta, \theta_k) &= N\log|Q|\\ &\quad + \sum_{t=1}^{N-1} \mathbf{E}_{\theta_k}\left\{(\xi_{t+1} - A_\xi\xi_t)^T Q^{-1}(\xi_{t+1} - A_\xi\xi_t)\right\}\\ &\quad + N\log|R| + \sum_{t=1}^{N} \mathbf{E}_{\theta_k}\left\{\varepsilon_t^T R^{-1}\varepsilon_t\right\}.\end{aligned} \tag{32}$$

Using the identity that $\mathrm{Tr}\{x^T y\} = \mathrm{Tr}\{yx^T\}$ for arbitrary vectors x and y then completes the proof. □

To address the difficulty of computing the conditional expectations (28)–(31) that are required to evaluate $\mathcal{Q}(\theta, \theta_k)$ this paper will employ sequential importance resampling (SIR) methods, which are more colloquially known as "particle" techniques.

Underpinning these approaches, is the central idea of generating a user chosen number M of random realizations (particles) ξ_t^i, $i = 1, \dots, M$ from the smoothing density of interest $\xi_t^i \sim p(\xi_t \mid Y_N)$.

Generating random realizations from the smoothing density requires a preceding step of generating realizations ζ_t^i for $i = 1, \dots, M$ from the *filtering* density $p(\xi_t \mid Y_t)$. The following algorithm for achieving this has now become a benchmark, although there are many variants on it (Arulampalam, Maskell, Gordon, & Clapp, 2002; Douc, 2001; Ristic, Arulampalam, & Gordon, 2004).

The development of particle smoothing methods is much less mature. However, the recent work (Douc, Garivier, Moulines, & Olsson, 2011) has derived a new approach that is both computationally efficient, and has the great advantage of generating realizations from the complete *joint* smoothing density $p(\xi_1, \dots, \xi_N \mid Y_N)$.

This is particularly important for the work here since via (29), approximations based on realizations drawn from the joint density $p(\xi_{t+1}, \xi_t \mid Y_N)$ are required. In previous work where realizations only from the marginal $p(\xi_t \mid Y_N)$ are available, it is then necessary to approximate an extra integration step (Schön et al., 2011, Lemma 6.1) that can now be avoided.

Algorithm 2 Particle Filter

```
1: Initialize particles, $\{\tilde{\zeta}_1^i\}_{i=1}^M \sim p_\theta(\tilde{\zeta}_1)$ and set $t = 1$;
2: Compute the importance weights $\{w_t^i\}_{i=1}^M$,
       $w_t^i \triangleq w(\tilde{\zeta}_t^i) = \frac{p_\theta(y_t | \tilde{\zeta}_t^i)}{\sum_{j=1}^M p_\theta(y_t | \tilde{\zeta}_t^j)}, \qquad i = 1, \dots, M.$   (33)
3: For each $j = 1, \dots, M$ draw a new particle $\zeta_t^j$ with replacement
   (resample) according to,
       $\mathrm{P}(\zeta_t^j = \tilde{\zeta}_t^i) = w_t^i, \qquad i = 1, \dots, M.$   (34)
4: Predict the particles by drawing $M$ i.i.d. samples according to
       $\tilde{\zeta}_{t+1}^i \sim p_\theta(\tilde{\zeta}_{t+1} | \zeta_t^i), \qquad i = 1, \dots, M.$   (35)
5: If $t < N$ increment $t \mapsto t + 1$ and return to step 2, otherwise
   terminate.
```

The smoothing method developed in Douc et al. (2011) addresses a very general class of problems and initial particle filtering methods for which a central consideration is a desired target density $p(\xi_{t+1} \mid \xi_t)$ which in this paper, according to the model (6) has the Gaussian form

$$p(\xi_{t+1} \mid \xi_t) = (|2\pi Q|)^{-1/2} g(\xi_{t+1}, \xi_t, \theta), \tag{36}$$

where

$$g(\xi_{t+1}, \xi_t, \theta) \triangleq \exp\left(-\frac{1}{2} (\xi_{t+1} - A_\xi \xi_t)^T Q_\xi^{-1} (\xi_{t+1} - A_\xi \xi_t) \right). \tag{37}$$

This form, and the fact that the particle filter defined in Algorithm 2 resamples at every time step allows some important simplification of the general smoother developed in Douc et al. (2011) so that it can be expressed in the following concrete form of Algorithm 3.

Algorithm 3 Rejection Sampling Based Particle Smoother

```
1: Run the particle filter (Algorithm 2) and store all the generated
   particles $\zeta_t^i$ for $t = 1, \dots, N$ and $i = 1, \dots, M$;
2: Set $t = N$ and initialize the smoothed particles $\xi_N^i = \zeta_N^i$ for
   $i = 1, \dots, M$;
3: for $i = 1 : M$ do
4:     Draw an integer $j$ randomly according to $j \sim \mathcal{U}([1, \dots, M])$
       where the latter is the uniform distribution over the integers
       $1, \dots, M$;
5:     Draw a real number $\tau$ randomly according to $\tau \sim U([0, 1])$
       where the latter is the uniform distribution over the real
       numbers in the interval $[0, 1]$;
6:     if $\tau > g(\xi_t^i, \zeta_{t-1}^j, \theta)$ then
7:         return to step 4;
8:     end if
9:     Set $\xi_{t-1}^i = \zeta_{t-1}^j$.
10: end for
11: if $t > 1$ then
12:     Decrement $t \mapsto t - 1$. Return to step 4
13: else
14:     Terminate;
15: end if
```

This paper proposes using the realizations ξ_t^i $i = 1, \dots, M$ from the joint smoothing density $p(\xi_1, \dots, \xi_N \mid Y_N)$ generated by Algorithm 3 to approximate the components (28)–(31) as follows

$$\Phi \approx \widehat{\Phi} \triangleq \frac{1}{M} \sum_{t=1}^{N-1} \sum_{i=1}^{M} \xi_{t+1}^i (\xi_{t+1}^i)^T, \tag{38}$$

$$\Psi \approx \widehat{\Psi} \triangleq \frac{1}{M} \sum_{t=1}^{N-1} \sum_{i=1}^{M} \xi_{t+1}^i (\xi_t^i)^T, \tag{39}$$

$$\Sigma \approx \widehat{\Sigma} \triangleq \frac{1}{M} \sum_{t=1}^{N-1} \sum_{i=1}^{M} \xi_t^i (\xi_t^i)^T \tag{40}$$

$$\Upsilon \approx \widehat{\Upsilon} \triangleq \frac{1}{M} \sum_{t=1}^{N} \sum_{i=1}^{M} \varepsilon_t^i (\varepsilon_t^i)^T \tag{41}$$

$$\varepsilon_t^i \triangleq y_t - f_W(z_t^i, \beta) - C_\mu \xi_t^i, \tag{42}$$

$$z_t^i = \mathcal{L}(w_t, \vartheta) + C_\nu \xi_t^i \tag{43}$$

and therefore approximate $\mathcal{Q}(\theta, \theta_k) \approx \widehat{\mathcal{Q}}(\theta, \theta_k)$ defined as

$$\begin{aligned} -2\widehat{\mathcal{Q}}(\theta, \theta_k) &\triangleq N \log |Q| + N \log |R| + \mathrm{Tr}\left\{ R^{-1} \widehat{\Upsilon} \right\} \\ &\quad + \mathrm{Tr}\left\{ Q^{-1} \left[\widehat{\Phi} - \widehat{\Psi} A_\xi^T - A_\xi \widehat{\Psi}^T + A_\xi \widehat{\Sigma} A_\xi^T + R^{-1} \widehat{\Upsilon} \right] \right\}. \end{aligned} \tag{44}$$

This approximation $\widehat{\mathcal{Q}}(\theta, \theta_k)$ is based on the standard rationale underpinning particle filtering and smoothing methods wherein by the law of large numbers (LLN), sample averages of the random realizations (38)–(41) converge, with increasing number of particles M to the ensemble expectations (28)–(31), and therefore approximate convergence can be expected to hold in the finite M cases (38)–(41).

To formally establish that the LLN applies in this particular case is a formidable technical challenge, since the particle realizations are not independent. For certain classes of particle *filtering* methods, some results are available establishing stochastic convergence for general functions of the particle realizations (Douc & Moulines, 2008; Hu, Schön, & Ljung, 2008). Unfortunately, there are at present no such theoretical studies available for the recently developed particle smoothing method employed here. In absence of this, Section 6 following provides an empirical study to establish evidence for convergence and the utility of the LLN-based approximations (38)–(41).

5. Maximizing $\widehat{\mathcal{Q}}(\theta, \theta_k)$

The second "M-step" of the EM algorithm involves the maximization of $\mathcal{Q}(\theta, \theta_k)$ over θ. In this paper, this will be approximated by the maximization of $\widehat{\mathcal{Q}}(\theta, \theta_k)$, which may be decomposed into two separately parametrized components

$$-2\widehat{\mathcal{Q}}(\theta, \theta_k) = I_1(A_\xi, Q) + I_2(R, \eta) \tag{45}$$

where

$$\begin{aligned} I_1(A_\xi, Q) &\triangleq N \log |Q| \\ &\quad + \mathrm{Tr}\left\{ Q^{-1} \left[\widehat{\Phi} - \widehat{\Psi} A_\xi^T - A_\xi \widehat{\Psi}^T + A_\xi \widehat{\Sigma} A_\xi^T \right] \right\} \end{aligned} \tag{46}$$

$$I_2(R, \eta) \triangleq N \log |R| + \mathrm{Tr}\left\{ R^{-1} \widehat{\Upsilon} \right\} \tag{47}$$

and

$$\eta \triangleq \left[\vartheta^T, \alpha^T, \beta^T, \mathrm{vec}\{C_\nu\}^T, \mathrm{vec}\left\{ C_\mu \right\}^T \right]^T \tag{48}$$

where the vec$\{\cdot\}$ operator creates a vector from a matrix by stacking its columns on top of one another.

Maximizing $\widehat{\mathcal{Q}}(\theta, \theta_k)$ therefore involves minimizing these two components. Achieving this for $I_1(A_\xi)$ is straightforward.

Lemma 5.1. *If $\widehat{\Sigma} \succ 0$ then $I_1(A_\xi, Q)$ as a function of A_ξ is uniquely minimized by the choice*

$$A_\xi = \widehat{\Psi} \widehat{\Sigma}^{-1}. \tag{49}$$

Proof. The term inside the trace operator in (46) may be expressed as

$$\begin{aligned}&\widehat{\Phi} - \widehat{\Psi}A_\xi^T - A_\xi\widehat{\Psi}^T + A_\xi\widehat{\Sigma}A_\xi^T \\ &\quad = (A_\xi - \widehat{\Psi}\widehat{\Sigma}^{-1})\widehat{\Sigma}(A_\xi - \widehat{\Psi}\widehat{\Sigma}^{-1})^T + \widehat{\Phi} - \widehat{\Psi}\widehat{\Sigma}^{-1}\widehat{\Psi}^T. \end{aligned} \tag{50}$$

Therefore, I_1 depends as a function of A_ξ only on the first term in (50) which is non-negative, but may be set to zero by the choice (49). □

Likewise, minimizing $I_1(A_\xi, Q)$ with respect to Q and $I_2(R, \eta)$ with respect to R is also straightforward.

Lemma 5.2. *The value*

$$Q = \frac{1}{N}\left[\widehat{\Phi} - \widehat{\Psi}\widehat{\Sigma}^{-1}\widehat{\Psi}^T\right] \tag{51}$$

is a stationary point of $I_1(\widehat{\Psi}\widehat{\Sigma}^{-1}, Q)$ with respect to Q, and the value

$$R = \frac{1}{N}\widehat{\Upsilon} \tag{52}$$

is a stationary point of $I_2(R, \eta)$ with respect to R.

Proof. Beginning with $I_2(R, \eta)$, via well known results of matrix calculus (Bernstein, 2005)

$$\frac{\partial}{\partial R}N\log|R| + \frac{\partial}{\partial R}\text{Tr}\left\{R^{-1}\widehat{\Upsilon}\right\} = NR^{-1} - R^{-1}\widehat{\Upsilon}R^{-1} \tag{53}$$

which is clearly zero for the choice (52). Furthermore, via (50), $I_1(A_\xi, Q)$ evaluated at the minimizer (49) is given as

$$I_1(\widehat{\Psi}\widehat{\Sigma}^{-1}, Q) = N\log|Q| + \text{Tr}\left\{Q^{-1}\left[\widehat{\Phi} - \widehat{\Psi}\widehat{\Sigma}^{-1}\widehat{\Psi}^T\right]\right\}. \tag{54}$$

Establishing that Q given by (51) is a stationary point of this expression then proceeds via the argument (53) just used in relation to $I_2(R, \eta)$. □

Unfortunately, it is not possible to derive closed form expressions for the stationary point of $I_2(R, \eta)$ with respect to the remaining parameter vector η. As a solution, this paper suggests computing a minimizer with respect to η via a standard gradient based search update of the form

$$\eta \leftarrow \eta + \gamma\rho. \tag{55}$$

Here the vector ρ is given by the Gauss–Newton search direction (Dennis & Schnabel, 1983) defined as

$$\rho = H(\eta)^{-1}g(\eta), \tag{56}$$

where the j'th element of the (negative) gradient vector g is given by

$$g_j(\eta) \triangleq \frac{\partial I_2(R, \eta)}{\partial \eta_j} = \frac{1}{M}\sum_{t=1}^{N}\sum_{i=1}^{M}\varepsilon_t^i(\eta)\frac{\partial\varepsilon_t^i(\eta)}{\partial\eta_j} \tag{57a}$$

and the (ℓ, j)'th element of the scaling matrix H is given by

$$H_{(\ell,j)}(\eta) = \frac{1}{M}\sum_{t=1}^{N}\sum_{i=1}^{M}\frac{\partial\varepsilon_t^i(\eta)}{\partial\eta_\ell}\frac{\partial\varepsilon_t^i(\eta)}{\partial\eta_j}. \tag{58}$$

Based on this choice for ρ, it can be shown that there exists a $\gamma > 0$ so that $I_2(R, \eta + \gamma g(\eta)) > I_2(R, \eta)$, which we achieve using a backstepping line search in this paper.

To be more precise, the combination of the results of Lemmas 5.1 and 5.2 together with a gradient based search relative to η results in the following proposed Algorithm 4 for maximizing $\widehat{\mathcal{Q}}(\theta, \theta_k)$.

The utility and efficacy of this combined EM/particle smoothing approach will now be illustrated via empirical study.

Algorithm 4 M-step

Given the current parameter values θ_k and a positive scalar ϵ, perform the following:

```
Update the elements of $\theta$ affected by $A_\xi$, $Q$ and $R$ via (49), (51), (52);
Initialize $\eta$ from the appropriate elements of $\theta_k$.
while $\|g(\eta)\| < \epsilon$ do
  Compute $\rho = H(\eta)^{-1}g(\eta)$;
  Set $\gamma = 1$;
  while $I_2(\eta + \gamma\rho) < I_2(\eta)$ do
    Update $\gamma \leftarrow \gamma/2$;
  end while
  Set $\eta \leftarrow \eta + \gamma\rho$;
end while
Set the appropriate elements of $\theta$ to the terminal values of $\eta$.
Compute $R$ via (52), using the new estimates just obtained and update the appropriate elements in $\theta$.
```

6. Simulation examples

6.1. Blind estimation of Wiener model with 4'th order linear part and non-invertible nonlinearity

In this first example we consider a Wiener system in the form of Fig. 1 where the Hammerstein nonlinearity f_H and the linear dynamic block $\mathcal{L}$ are not present. This results in a *blind Wiener estimation* problem where only the output measurements are available for estimating the parameters of the state-space coloring filter (6), and the Wiener nonlinearity f_W. To that end, the process noise μ_t was generated by a passing Gaussian white noise υ_t through a 4'th order transfer function

$$\nu_t = H(q)\upsilon_t, \tag{59}$$

$$H(q) = \frac{c_1q^{-1} + \cdots + c_4q^{-4}}{1 + a_1q^{-1} + \cdots + a_4q^{-4}} \tag{60}$$

with parameter values $a = [a_1, \ldots, a_4]$, $c = [c_1, \ldots, c_4]$ given by

$$a = \begin{bmatrix}0.3676, & 0.88746, & 0.52406, & 0.55497\end{bmatrix}, \tag{61}$$

$$c = \begin{bmatrix}1, & 0.1, & -0.49, & 0.01\end{bmatrix}. \tag{62}$$

The true nonlinearity $f_{W\text{true}}$ is given by a saturation function according to

$$f_{W\text{true}}(\nu_t) = \begin{cases}0.3 & : \nu_t > 0.3\\ \nu_t & : -0.2 \le \nu_t \le 0.3\\ -0.2 & : \nu_t < -0.2.\end{cases} \tag{63}$$

In terms of the estimation model structure, the nonlinearity was modeled as a piecewise linear function with a number n_{pw} of transitions between linear sub-components. It is parametrized by a vector $\beta \in \mathbf{R}^{2(n_{\text{pw}}+1)}$ that specifies a linear base together with n_{pw} "hinge" functions $h_j(\cdot, \beta)$ (Breiman, 1993):

$$f_W(\nu_t, \beta) = \beta_{0,0} + \beta_{0,1}\nu_t + \sum_{j=1}^{n_{\text{pw}}}h_j(\nu_t, \beta), \tag{64a}$$

$$h_j(\nu_t, \beta) = \begin{cases}\beta_{j,0} + \beta_{j,1}\nu_t; & \nu_t > -\dfrac{\beta_{j,0}}{\beta_{j,1}},\\ 0; & \text{Otherwise}\end{cases} \tag{64b}$$

$$\beta = \begin{bmatrix}\beta_{0,0} & \beta_{0,1} & \beta_{1,0} & \beta_{1,1} & \cdots & \beta_{n_{\text{pw}},0} & \beta_{n_{\text{pw}},1}\end{bmatrix}. \tag{64c}$$

For the purposes of estimation, $N = 5000$ samples of the output were simulated via

$$y_t = f_{W\text{true}}(\nu_t) + e_t, \qquad \nu_t = H(q)\upsilon_t, \tag{65}$$

with the state noise source $\upsilon_t \sim \mathcal{N}(0, 0.1)$. The measurement noise was distributed according to $e_t \sim \mathcal{N}(0, 0.001)$.

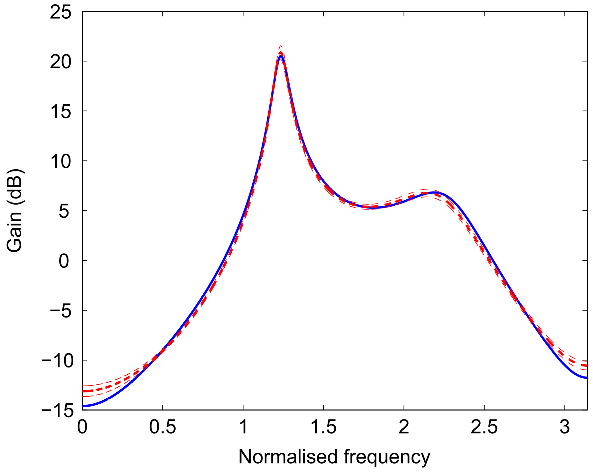

Fig. 2. Bode plot of estimated mean (thick red-dashed) and standard deviation (thin red-dashed) against the true (blue-solid) system for the example studied in Section 6.1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

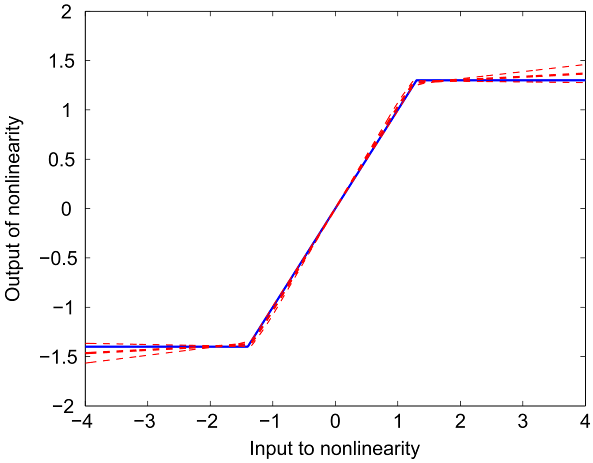

Fig. 3. Estimated mean (thick red-dashed) and standard deviation (thin red-dashed) together with the true (blue-solid) memoryless nonlinearities for the example studied in Section 6.1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The number of hinges used in the nonlinear block was chosen as $n_{\text{pw}} = 2$. Further, the parameter vector β was initialized as

$$\beta = \begin{bmatrix} 0, & 1, & -0.001, & -1, & 0.001, & 1 \end{bmatrix}, \tag{66}$$

which approximates a straight line.

The parameters of the noise filter state-space matrices (A_ξ, C_ν, Q) were initialized by using a subspace method (van Overschee & De Moor, 1996) based on the measurements $\{y_1, \dots, y_N\}$.

Using the above combination of initial parameter values, the EM method was employed to provide ML estimates based on $M = 200$ particles and using 100 iterations. The results of 100 Monte Carlo runs are shown in Figs. 2–3. For each run, different noise realizations were used according to the distributions specified above.

6.2. MIMO Hammerstein–Wiener system

As a further example, a multiple-input/multiple-output Hammerstein–Wiener system is now considered. The system has two inputs, two outputs and the linear dynamic block $\mathcal{L}(\cdot, \vartheta)$ is a 4'th order system described by (5), where the state-space matrices (A, B, C, D) conform with the transfer function

$$G \triangleq C(qI - A)^{-1}B + D = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}, \tag{67a}$$

where

$$G_{11} = \frac{1.1 - 0.99q^{-1} - 0.17q^{-2} + 0.51q^{-3} - 0.18q^{-4}}{1 - 0.77q^{-1} - 0.56q^{-2} + 0.38q^{-3} + 0.012q^{-4}},$$

$$G_{12} = \frac{0.35q^{-1} - 0.31q^{-2} - 0.24q^{-3} + 0.066q^{-4}}{1 - 0.77q^{-1} - 0.56q^{-2} + 0.38q^{-3} + 0.012q^{-4}},$$

$$G_{21} = \frac{-0.86 + 0.39q^{-1} + 0.40q^{-2} - 0.20q^{-3} + 0.012q^{-4}}{1 - 0.77q^{-1} - 0.56q^{-2} + 0.38q^{-3} + 0.012q^{-4}},$$

$$G_{22} = \frac{-0.12q^{-1} + 0.15q^{-2} + 0.12q^{-3} - 0.0033q^{-4}}{1 - 0.77q^{-1} - 0.56q^{-2} + 0.38q^{-3} + 0.012q^{-4}}.$$

The true Hammerstein nonlinearity f_H is given by

$$f_H(u_t, \alpha) = \begin{bmatrix} f_{H,1}(u_t(1), \alpha) \\ f_{H,2}(u_t(2), \alpha) \end{bmatrix} \tag{67b}$$

where $f_{H,1}$ is a saturation function, $f_{H,2}$ is a deadzone function and $u_t(i)$ is used to denote the i'th input signal. More specifically,

$$f_{H,1}(u_t(1), \alpha) = \begin{cases} \alpha_1 & : u_t(1) < \alpha_1 \\ u_t(1) & : \alpha_1 \le u_t(1) \le \alpha_2 \\ \alpha_2 & : u_t(1) > \alpha_2 \end{cases} \tag{67c}$$

$$f_{H,2}(u_t(2), \alpha) = \begin{cases} u_t(2) - \alpha_3 & : u_t(2) < \alpha_3 \\ 0 & : \alpha_3 \le u_t(2) \le \alpha_4 \\ u_t(2) - \alpha_4 & : u_t(2) > \alpha_4 \end{cases} \tag{67d}$$

with the true values for α given by

$$\alpha_1 = -0.8, \qquad \alpha_2 = 0.8, \qquad \alpha_3 = -0.9, \qquad \alpha_4 = 0.9. \tag{67e}$$

The true Wiener nonlinearity f_W is given in a similar manner by

$$f_W(z_t, \beta) = \begin{bmatrix} f_{W,1}(z_t(1), \beta) \\ f_{W,2}(z_t(2), \beta) \end{bmatrix} \tag{67f}$$

where $f_{W,1}$ is a deadzone function, $f_{W,2}$ is a saturation function and $z_t(i)$ is used to denote the i'th element of the vector signal $z_t \in \mathbf{R}^2$. More specifically,

$$f_{W,1}(z_t(1), \beta) = \begin{cases} z_t(1) - \beta_1 & : z_t(1) < \beta_1 \\ 0 & : \beta_1 \le z_t(1) \le \beta_2 \\ z_t(1) - \beta_2 & : z_t(1) > \beta_2 \end{cases} \tag{67g}$$

$$f_{W,2}(z_t(2), \beta) = \begin{cases} \beta_3 & : z_t(2) < \beta_3 \\ z_t(2) & : \beta_3 \le z_t(2) \le \beta_4 \\ \beta_4 & : z_t(2) > \beta_4 \end{cases} \tag{67h}$$

with the true values for β given by

$$\beta_1 = -0.8, \qquad \beta_2 = 0.8, \qquad \beta_3 = -0.9, \qquad \beta_4 = 0.9. \tag{67i}$$

Finally, the process noise signal μ_t was colored according to (6c) with state-space matrices given by

$$A_\xi = \begin{bmatrix} 0.2 & -0.82 \\ 1 & 0 \end{bmatrix}, \qquad C_\mu = \begin{bmatrix} 0 & -0.81 \\ 0 & -0.81 \end{bmatrix}, \qquad Q = I_2. \tag{67j}$$

In terms of the estimation model structure, we used a 4'th order state-space model for the linear dynamic system and the nonlinearities were modeled by the 4'th order piecewise linear structure described in (64a)–(64c). The colored noise was modeled using a 2'nd order state-space structure.

For the purposes of estimation, $N = 2000$ samples of the inputs and outputs were simulated using (67). The output measurements were corrupted by Gaussian noise $\nu_t \sim \mathcal{N}(0, I_2)$.

The goal is to estimate the parameter vector θ based on input and output measurements. In this case, three different algorithms were employed:

(1) A prediction error method that assumes output noise only, called the OE method. This is the approach used in the industry standard software toolbox (Ljung, 2011);

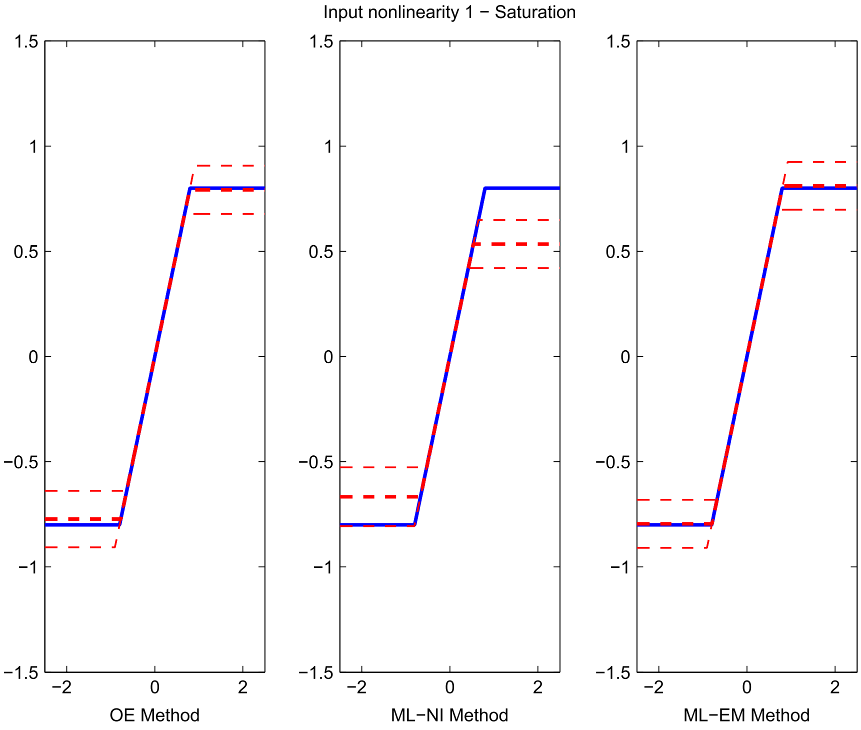

Fig. 4. Input nonlinearity-1 for the example studied in Section 6.2.

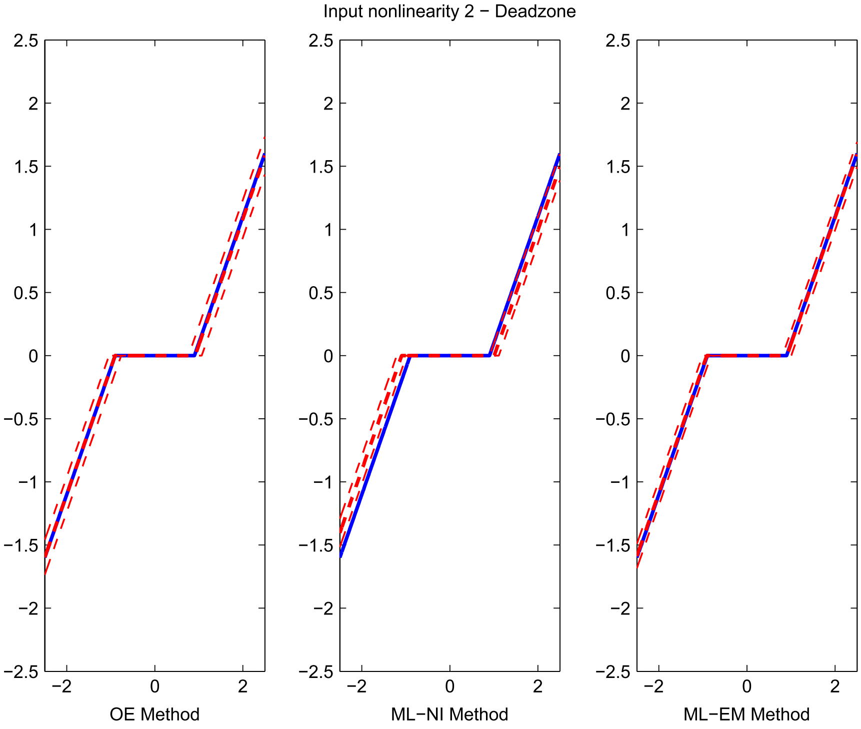

Fig. 5. Input nonlinearity-2 for the example studied in Section 6.2.

(2) A maximum-likelihood method developed in Hagenblad et al. (2008) that employs numerical integration techniques and assumes that the noise ν_t is Gaussian and independent. This will be called the ML-NI approach;
(3) The method developed in this paper, called the ML-EM approach.

It should be mentioned that the first two algorithms do not cater for estimating either of the noise filter dynamics. It is interesting nonetheless to observe their performance based on the wrong assumptions that each make about the process noise, i.e. it doesn't exist in the first case, and it is assumed white in the second.

The first two algorithms were initialized with the true parameter values in order to reduce the likelihood of capture in local minima. The EM approach was initialized at $\theta/5$ in order to demonstrate that the method performs well even when starting from poor initial estimates.

For the ML-EM method, $M = 100$ particles were used. Again the algorithm was terminated after just 100 iterations. The results

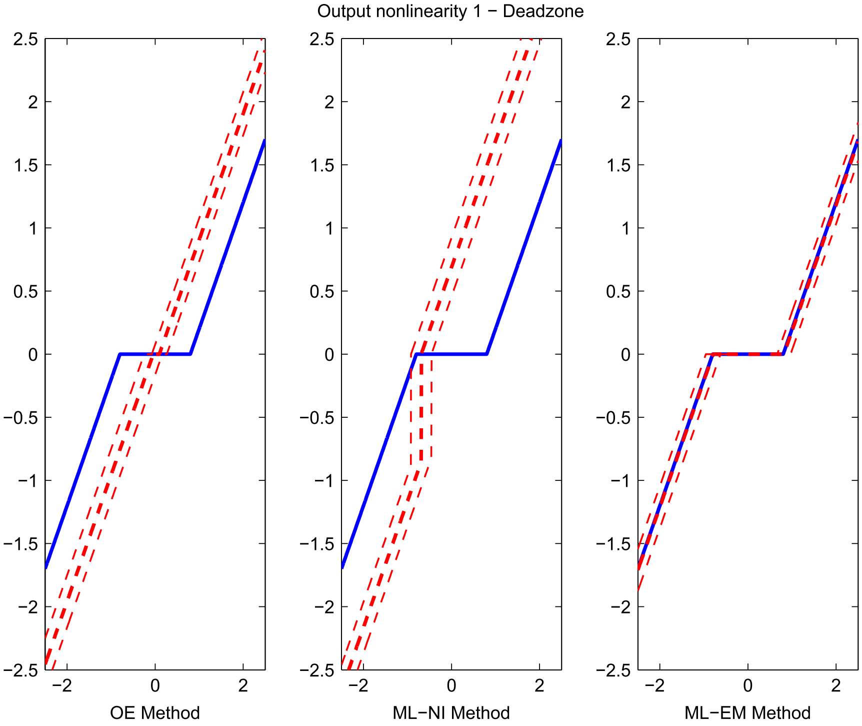

Fig. 6. Output nonlinearity-1 for the example studied in Section 6.2.

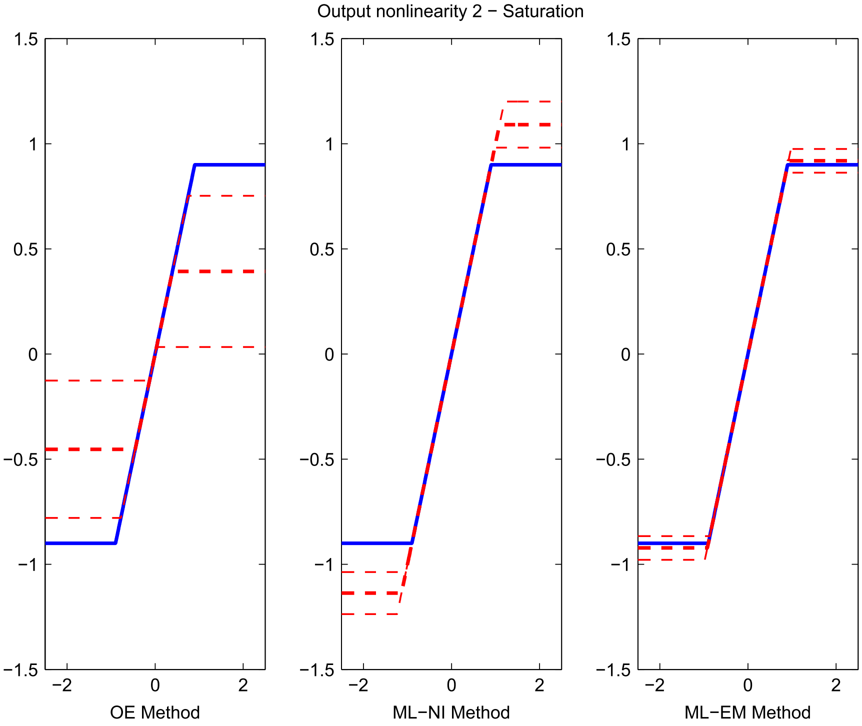

Fig. 7. Output nonlinearity-2 for the example studied in Section 6.2.

of 80 Monte Carlo runs for all algorithms are shown in Figs. 4–10. For each run, different noise realizations were used according to the distributions specified above.

These figures demonstrate the utility of the proposed algorithm in that the estimates appear to be informative, even though the initial estimates are clearly far from accurate. Note in particular that both the OE and ML-NI methods appear to produce biased estimates, while the ML-EM approach appears to be unbiased and accurate.

7. Conclusion

This paper has considered the problem of identifying parameter values for Hammerstein–Wiener systems where both colored process noise and white measurement noise are considered. It also straightforwardly captures the blind identification problem for Wiener systems as an interesting special case. The static nonlinearities associated with the Hammerstein–Wiener system are allowed to be quite general and do not need to be invertible.

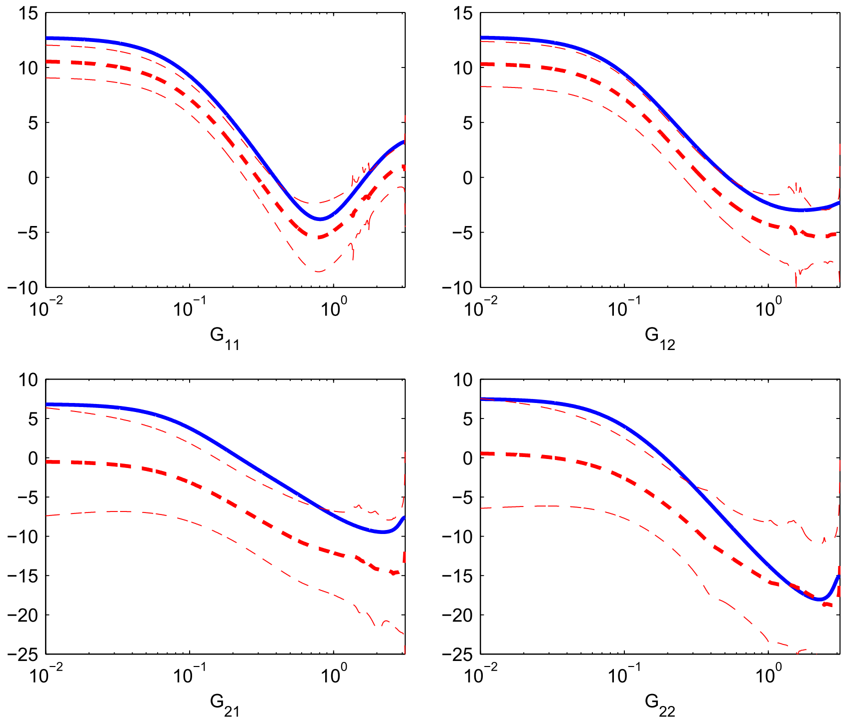

Fig. 8. Bode magnitude response using the OE method for the example studied in Section 6.2.

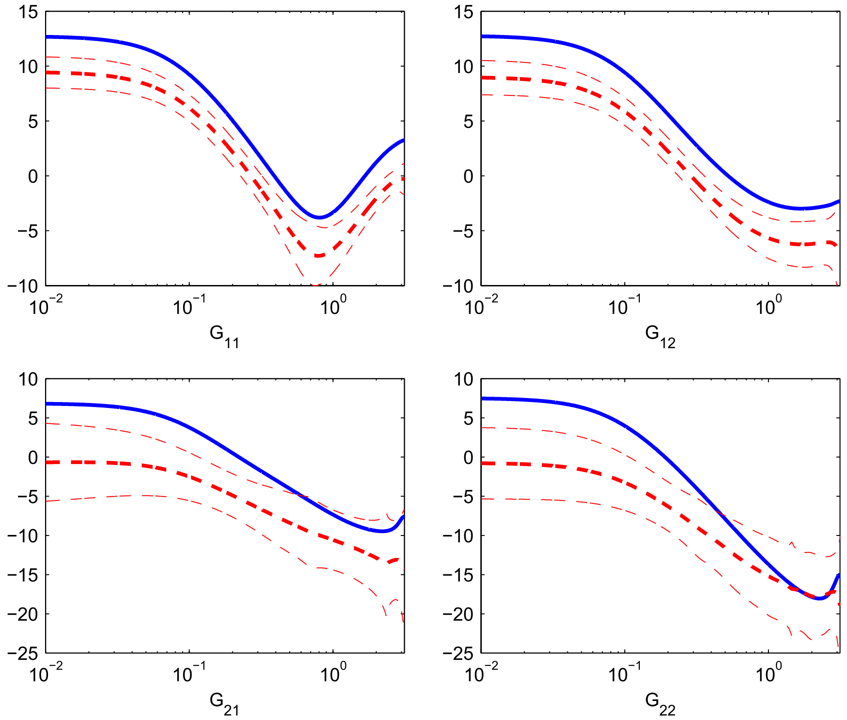

Fig. 9. Bode magnitude response using the ML-NI method for the example studied in Section 6.2.

This identification problem was specified using a maximum likelihood formulation, which depends on an underlying prediction density. The key technical difficulty in solving this problem is that the prediction density cannot be straightforwardly characterized. The impact is that the likelihood function cannot be straightforwardly evaluated, let alone maximized.

To address this, the paper employs the expectation maximization (EM) algorithm, which does not need to evaluate the likelihood nor directly maximize it. The results of this new approach were profiled on two examples that establish the utility of the new methods developed here.

Acknowledgment

The authors would like to thank Dr Jimmy Olsson for pointing us to the particle smoother introduced in Douc et al. (2011).

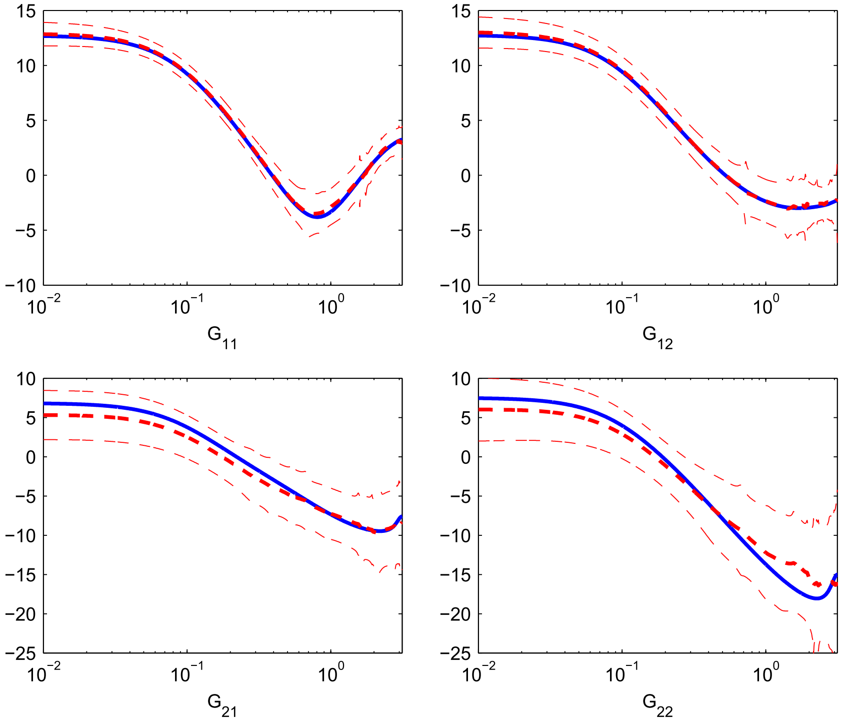

Fig. 10. Bode magnitude response using the ML-EM method for the example studied in Section 6.2.

References

Abed-Meraim, K., Qiu, W., & Hua, Y. (1997). Blind system identification. *Proceedings of the IEEE*, *85*(8), 1310–1322.

Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, *50*(2), 174–188.

Bai, E. (2002a). Identification of linear systems with hard input nonlinearities of known structure. *Automatica*, *38*(5), 853–860.

Bai, E. (2002b). A blind approach to Hammerstein–Wiener model identification. *Automatica*, *38*(6), 967–979.

Bauer, Dietmar, & Ninness, Brett (1999). Asymptotic properties of least-squares estimates of Hammerstein–Weiner model structures. Technical Report EE9947. Department of Electrical and Computer Engineering. University of Newcastle.

Bernstein, D. S. (2005). *Matrix mathematics*. Princeton University Press.

Billings, S. A., & Fakhouri, S. Y. (1982). Identification of systems containing linear dynamic and static nonlinear elements. *Automatica*, *18*(1), 15–26.

Breiman, L. (1993). Hinging hyperplanes for regression, classification, and function approximation. *IEEE Transactions on Information Theory*, *39*(3), 999–1013.

Caines, P. E. (1988). *Linear stochastic systems*. New York: John Wiley and Sons.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38.

Dennis, J. E., & Schnabel, R. B. (1983). *Numerical methods for unconstrained optimization and nonlinear equations*. Prentice Hall.

Djurić, P. M., & Godsill, S. J. (2002). Special issue on Monte Carlo methods for statistical signal processing. *IEEE Transactions on Signal Processing*, *50*(2), Guest Eds.

Doucet, A., de Freitas, N., & Gordon, N. (Eds.) (2001). *Sequential Monte Carlo methods in practice*. Springer-Verlag.

Doucet, A., & Johansen, A. (2011). A tutorial on particle filtering and smoothing: fifteen years later. In D. Crisan, & B. Rozovsky (Eds.), *The Oxford handbook of nonlinear filtering*. Oxford University Press.

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

Douc, Randal, & Moulines, Eric (2008). Limit theorems for weighted samples with applications to sequential Monte Carlo methods. *The Annals of Statistics*, *36*(5), 2344–2376.

Gibson, S. H., & Ninness, B. (2005). Robust maximum-likelihood estimation of multivariable dynamic systems. *Automatica*, *41*(10), 1667–1682.

Giri, F., & Bai, E. (Eds.) (2010). *Lecture notes in control and information sciences*: *vol. 404. Block-oriented nonlinear system identification*. Springer.

Hagenblad, A., Ljung, L., & Wills, A. G. (2008). Maximum likelihood identification of Wiener models. *Automatica*, *44*(11), 2697–2705.

Heunis, A. J. (1988). Asymptotic properties of prediction error estimators in approximate system identification. *Stochastics*, *24*, 1–43.

Hsu, K., Vincent, T., & Poolla, K. (2006). A kernel based approach to structured nonlinear system identification part i: algorithms. part ii: convergence and consistency. In *Proceedings of IFAC symposium on system identification*. Newcastle. March.

Hu, X. L., Schön, Thomas, & Ljung, Lennart (2008). A basic convergence result for particle filtering. *IEEE Transactions on Signal Processing*, *56*(4), 1337–1348.

Kalafatis, A. D., Wang, L., & Cluett, W. R. (1997). Identification of Wiener-type nonlinear systems in a noisy environment. *International Journal of Control*, *66*, 923–941.

Ljung, Lennart (1999). *System identification: theory for the user* (2nd ed.). New Jersey: Prentice-Hall, Inc.

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

Ljung, Lennart (1998). Identification for control—what is there to learn? In Y. Yamamoto, & S. Hara (Eds.), *Springer lecture notes in control and information sciences*: *vol. 241. Learning, control and hybrid systems* (pp. 207–221). Berlin: Springer Verlag.

Ljung, L. (2011). *MATLAB system identification toolbox users guide, version 9*. The Mathworks.

Ljung, Lennart, & Wills, Adrian (2010). Issues in sampling and estimating continuous-time models with stochastic disturbances. *Automatica*, http://dx.doi.org/10.1016/j.automatica.2010.02.011. Available online 15th March 2010.

McLachlan, G., & Krishnan, T. (2008). *The EM algorithm and extensions* (2nd ed.). John Wiley and Sons.

Narendra, K. S., & Gallman, P. G. (1966). In iterative method for the identification of nonlinear systems using a Hammerstein model. *IEEE Transactions on Automatic Control*, *11*(7), 546–550.

Ninness, B. M. (1998). Estimation of $1/f$ noise. *IEEE Transactions on Information Theory*, *44*(1), 32–46.

Pötscher, B., & Prucha, I. (1997). *Dynamic nonlinear econometrics models*. Berlin–Heidelberg: Springer-Verlag.

Raich, R., Zhou, G. T., & Viberg, M. (2005). Subspace based approaches for Wiener system identification. *IEEE Transactions on Automatic Control*, *50*(10), 1629–1634.

Ristic, B., Arulampalam, S., & Gordon, N. (2004). *Beyond the Kalman filter: particle filters for tracking applications*. Boston, MA, USA: Artech house.

Schetzen, M. (1980). *The Volterra and Wiener theories of nonlinear systems*. Wiley.

Schön, T. B., Wills, A. G., & Ninness, B. (2011). System identification of nonlinear state-space models. *Automatica*, *37*(1), 39–49.

Schoukens, J., Nemeth, J. G., Crama, P., Rolain, Y., & Pintelon, R. (2003). Fast approximate identification of nonlinear systems. *Automatica*, *39*(7), 1267–1274.

Vanbeylen, L. R., Pintelon, R., & Schoukens, J. (2009). Blind maximum likelihood identification of Wiener systems. *IEEE Transactions on Signal Processing*, *57*(8), 3017–3029.

van Overschee, Peter, & De Moor, Bart (1996). *Subspace identification for linear systems—theory, implementation, applications*. Kluwer Academic Publishers.

Westwick, D., & Verhaegen, M. (1996). Identifying MIMO Wiener systems using subspace model identification methods. *Signal Processing*, *52*, 235–258.

Wiener, N. (1942). Response of a nonlinear system to noise. Technical report. Radiation Lab MIT 1942. restricted, report V-16. no. 129. (p. 112). Declassified Jul 1946. Published as rep. no. PB-1-58087. US Dept. Commerce.

Wigren, T. (1993). Recursive prediction error identification using the nonlinear Wiener model. *Automatica*, *29*(4), 1011–1025.

Wills, A. G., & Ljung, L. (2010). *Lecture notes in control and information sciences*: *vol. 404. Block-oriented nonlinear system identification*. Springer, (Chapter Wiener system identification using the maximum likelihood method).

Wills, A. G., Ninness, B., & Gibson, S. H. (2009). Maximum likelihood estimation of state space models from frequency domain data. *IEEE Transactions on Automatic Control*, *54*(1), 19–33.

Wills, A.G., Schön, T.B., Ljung, L., & Ninness, B. (2011). Blind identification of Wiener models. In *Proceedings of the IFAC world congress*.

Adrian Wills was born in Orange, N.S.W. Australia and received his B.E. (Elec.) and Ph.D. degrees from The University of Newcastle, Australia (Callaghan Campus) in May 1999 and May 2003, respectively. Since then he has held a postdoctoral research position at Newcastle, where the focus of his research has been in the area of system identification.

Thomas B. Schön was born in Jönköping (Sweden) on December 25, 1977. He is an Associate Professor with the Division of Automatic Control at Linköping University (Linköping, Sweden). He received the B.Sc. degree in Business Administration and Economics in Jan. 2001, the M.Sc. degree in Applied Physics and Electrical Engineering in Sep. 2001, the Lic. Eng. degree in Automatic Control in Oct 2003 and the Ph.D. degree in Automatic Control in Feb. 2006, all from Linköping University. He has held visiting positions with the University of Cambridge (UK) and the University of Newcastle (Australia). He is a Senior member of the IEEE.

Lennart Ljung received his Ph.D. in Automatic Control from Lund Institute of Technology in 1974. Since 1976 he has been Professor of the chair of Automatic Control In Linkoping, Sweden, and is currently Director of the Strategic Research Center "Modeling, Visualization and Information Integration" (MOVIII). He has held visiting positions at Stanford and MIT and has written several books on System Identification and Estimation. He is an IEEE Fellow, an IFAC Fellow and an IFAC Advisor. He is as a member of the Royal Swedish Academy of Sciences (KVA), a member of the Royal Swedish Academy of Engineering Sciences (IVA), an Honorary Member of the Hungarian Academy of Engineering, an Honorary Professor of the Chinese Academy of Mathematics and Systems Science, and a Foreign Associate of the US National Academy of Engineering (NAE). He has received honorary doctorates from the Baltic State Technical University in St Petersburg, from Uppsala University, Sweden, from the Technical University of Troyes, France, from the Catholic University of Leuven, Belgium and from the Helsinki University of Technology, Finland. In 2002 he received the Quazza Medal from IFAC, and in 2003 he received the Hendrik W. Bode Lecture Prize from the IEEE Control Systems Society, and he was the 2007 recipient of the IEEE Control Systems Award.

Brett Ninness was born in 1963 in Singleton, Australia and received his B.E., M.E. and Ph.D degrees in Electrical Engineering from the University of Newcastle, Australia in 1986, 1991 and 1994 respectively. He has stayed with the School of Electrical Engineering and Computer Science at the University of Newcastle since 1993, where he is currently a Professor.

His research interests are in the areas of system identification and stochastic signal processing, in which he has authored approximately one hundred papers in journals and conference proceedings. He has served on the editorial boards of Automatica, IEEE Transactions on Automatic Control and is currently Editor in Chief for IEE Control Theory and Applications.

Together with Håkan Hjalmarsson he jointly organized the 14th IFAC Symposium on System Identification in Newcastle, Australia in 2006. Further details of his professional activities are available at http://sigpromu.org/brett.