A kernel-based approach to Hammerstein system identification*

Riccardo S. Risuleo * **Giulio Bottegal** * **Håkan Hjalmarsson** *

* *ACCESS Linnaeus Centre, School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden (e-mail: {risuleo; bottegal; hjalmars}@kth.se)*

Abstract: In this paper, we propose a novel algorithm for the identification of Hammerstein systems. Adopting a Bayesian approach, we model the impulse response of the unknown linear dynamic system as a realization of a zero-mean Gaussian process. The covariance matrix (or kernel) of this process is given by the recently introduced stable-spline kernel, which encodes information on the stability and regularity of the impulse response. The static non-linearity of the model is identified using an Empirical Bayes approach, i.e. by maximizing the output marginal likelihood, which is obtained by integrating out the unknown impulse response. The related optimization problem is solved adopting a novel iterative scheme based on the Expectation-Maximization (EM) method, where each iteration consists in a simple sequence of update rules. Numerical experiments show that the proposed method compares favorably with a standard algorithm for Hammerstein system identification.

arXiv:1412.4055v1 [cs.SY] 12 Dec 2014

1. INTRODUCTION

The Hammerstein system is a cascaded system composed of a static nonlinearity followed by a linear dynamic system (see e.g. Ljung [1999]). Hammerstein system identification has become object of research apparently since the Sixties (Narendra and Gallman [1966]). Due to the wide spectrum of applications (see e.g. Hunter and Korenberg [1986], Westwick and Kearney [2001], Bai et al. [2009]), Hammerstein system identification has gained popularity through the years and a wide range of methods has been developed (see Rangan et al. [1995], Bai [1998], Bai and Li [2004], and references therein).

Several identification approaches have been proposed. For instance, Greblicki and Pawlak [1986] exploits kernel regression arguments, Westwick and Kearney [2001] uses a separable least squares approach, Greblicki [2002] focuses on stochastic system identification of Hammerstein models, while Goethals et al. [2005] proposes a subspace approach. Research on this topic is still rather active (see Schoukens et al. [2011], Han and De Callafon [2012]).

In this paper, we propose a novel method for Hammerstein system identification. Following recent developments in identification of linear dynamic systems (Chen et al. [2012], Pillonetto et al. [2014]), we adopt a kernel-based identification approach for the linear part of the Hammerstein model. To this end, we model the impulse response of the unknown linear system as a realization of a Gaussian process. The covariance matrix (or kernel) of this process has a specific structure given by the *stable spline kernel* (see Pillonetto and De Nicolao [2010], Pillonetto et al. [2011], Bottegal and Pillonetto [2013]). This structure induces properties such as BIBO stability and smoothness in the Gaussian process realizations and depends on a *shaping parameter* which regulates the exponential decay of the generated impulse responses.

In the context of Hammerstein system identification, we can define an effective estimator of the linear dynamic block using Bayesian arguments by exploiting the kernel-based framework. Such an estimator is function of the static nonlinearity, as well as the kernel shaping parameter and the noise variance. A crucial point of the proposed approach is the estimation of these quantities. Exploiting a Bayesian interpretation of kernel-based methods (Maritz and Lwin [1989]), we perform this estimation step by maximizing the marginal likelihood of the output measurements, which is obtained by integrating out the unknown impulse response. This approach has been shown to be effective in kernel-based linear system identification (Pillonetto and Chiuso [2014]). However, when applied to Hammerstein system identification, the related optimization problem becomes more involved due to the presence of the unknown static nonlinearity. To overcome this difficulty, we propose a novel iterative solution scheme based on the Expectation-Maximization method proposed by Dempster et al. [1977]. We show that the resulting Hammerstein system identification algorithm has a rather low computational burden. Remarkably, the proposed method does not need any parameter to be set nor requires the user to select the model order of the linear system. This in contrast with standard parametric methods, where, when little is known about the system, one as to estimate the optimal model order using complexity criteria or cross validation (see e.g. Ljung [1999]).

The method used in this paper is also used in Bottegal et al. [2015] in the context of blind system-identification.

The structure of the paper is as follows. In the next section, we formulate the Hammerstein system identification problem. In Section 3, we describe the model adopted for

* This work was supported by the European Research Council under the advanced grant LEARN, contract 267381 and by the Swedish Research Council under contract 621-2009-4017

the linear system and the static nonlinearity. In Section 4, we introduce the proposed algorithm, which is tested in Section 5. Some conclusions end the paper.

2. PROBLEM FORMULATION

We consider a single input single output discrete-time system described by the following time-domain relations (see Figure 1)

$$\begin{aligned} w_t &= f(u_t) \\ y_t &= \sum_{k=1}^{\infty} g_k w_{t-k} + e_t \,. \end{aligned} \tag{1}$$

In the above equation, $f(\cdot)$ represents a (static) nonlinear function transforming the measurable input u_t into the unavailable signal w_t, which in turn feeds a linear time-invariant (LTI) strictly causal system described by the impulse response g_t. The output measurements of the system y_t are corrupted by white Gaussian noise, denoted by e_t, which has unknown variance σ^2. For simplicity, we assume null initial conditions.

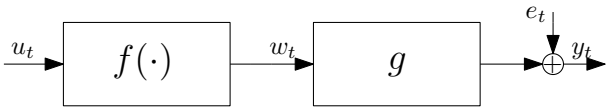

Fig. 1. Block scheme of the Hammerstein system.

We assume that N input-output samples are collected, and denote them by $\{u_t\}_{t=0}^{N-1}$, $\{y_t\}_{t=1}^{N}$. For notational convenience, we also assume null initial conditions. Then, the system identification problem we discuss in this paper is the problem of estimating n samples of the impulse response say, $\{\hat{g}_t\}_{t=1}^{n}$ (where n is large enough to capture the system dynamics), as well as the static nonlinearity $f(\cdot)$.

Remark 1. The identification method we propose in this paper is not affected by the choice of n. Furthermore, it can be derived also in the continuous-time setting, using the same arguments as in Pillonetto and De Nicolao [2010]. However, for ease of exposition, here we focus only on the discrete-time case.

2.1 Non-uniqueness of the identified system

It is well-known (see e.g. Bai and Li [2004]) that the two components of a Hammerstein system can be determined up to a scaling factor. In fact, for every $\alpha \in \mathbb{R}$, the pair $(\alpha g_t, \frac{1}{\alpha} f(\cdot))$, describes the input-output relation equally well. We will address this issue in the modelling of the impulse response by fixing the kernel scaling parameter (see Subsection 3.3).

3. MODELING AND IDENTIFICATION OF HAMMERSTEIN SYSTEMS

In this section, we first introduce the models adopted for the input static nonlinearity and the LTI system. Then, we describe the proposed system identification method.

3.1 Notation

Let us introduce the vector-based notation

$$u \triangleq \begin{bmatrix} u_0 \\ \vdots \\ u_{N-1} \end{bmatrix}, \; y \triangleq \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \; g \triangleq \begin{bmatrix} g_1 \\ \vdots \\ g_n \end{bmatrix}, \; e \triangleq \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}.$$

Furthermore, we define also the operator $\mathbf{T}_n(\cdot)$ that, given a vector of length N, maps it to an $N \times n$ Toeplitz matrix, e.g.

$$\mathbf{T}_n(w) = \begin{bmatrix} w_0 & 0 & & \dots & 0 \\ w_1 & w_0 & 0 & \dots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ w_{N-2} & w_{N-3} & \dots & w_{N-n+1} & 0 \\ w_{N-1} & w_{N-2} & \dots & \dots & w_{N-n} \end{bmatrix} \in \mathbb{R}^{N \times n}. \tag{2}$$

We shall reserve the symbol W for $\mathbf{T}_n(w)$. Then, the input-output relation of the LTI system can be written as

$$y = Wg + e \,. \tag{3}$$

3.2 The input static nonlinearity

Following Bai [1998], Bai and Li [2004], we assume that the input static nonlinearity belongs to a p-dimensional space of functions and thus can be described using a linear combination of known basis functions $\{\phi_i\}_{i=1}^{p}$. Hence, we can write

$$w_t = f(u_t) = \sum_{i=1}^{p} c_i \phi_i(u_t) \,, \tag{4}$$

where the coefficients c_i are unknown. The problem of estimating $f(\cdot)$ is thus equivalent to the problem of determining such coefficients. By introducing the following matrix

$$F(u) \triangleq \begin{bmatrix} \phi_1(u_0) & \dots & \phi_p(u_0) \\ \vdots & \vdots & \vdots \\ \phi_1(u_{N-1}) & \dots & \phi_p(u_{N-1}) \end{bmatrix} \tag{5}$$

we can write

$$w = F(u)c \,, \tag{6}$$

where $c \triangleq [c_1 \; \dots \; c_p]^T$.

3.3 Kernel-based modeling of the LTI system

In this paper, we adopt the kernel-based identification approach for LTI systems, proposed in Pillonetto and De Nicolao [2010], Pillonetto et al. [2011]. To this end, following a Gaussian process regression approach (Rasmussen and Williams [2006]), we assume that the impulse response of the system is a realization of a zero-mean Gaussian process, namely

$$g \sim \mathcal{N}(0, \lambda K_\beta) \,. \tag{7}$$

The matrix K_β, which is also known as *kernel*, is a covariance matrix parameterized by a shaping parameter β, and $\lambda \geq 0$ is a scaling factor. In the context of system identification, the family of the *stable spline kernels* (Pillonetto and De Nicolao [2010], Pillonetto et al. [2011]) constitutes a valid choice, since they promote BIBO stable and smooth realizations. Specifically, we employ the *first-order stable spline kernel* (or *TC kernel* in Chen et al. [2012]) given by

$$\{K_\beta\}_{i,j} \triangleq \beta^{\max(i,j)} \,, \quad 0 \leq \beta < 1 \tag{8}$$

where β determines the decaying velocity of the generated impulse responses. As λ regulates the amplitude of the impulse response g, we can we can arbitrarily set $\lambda = 1$, to cope with the identifiability issue described in Section 2.1.

3.4 Estimation of the LTI system

In this section we derive the system identification strategy that arises naturally when kernel-based methods are employed. The estimator we will obtain is function of the vector

$$\theta \triangleq \begin{bmatrix} c^T & \sigma^2 & \beta \end{bmatrix} \quad \in \mathbb{R}^{p+2}, \tag{9}$$

which we shall call *hyperparameter vector*. Since the noise is assumed Gaussian, the joint distribution of the vectors y and g is again Gaussian, and parametrized by θ. Thus

$$p\left(\begin{bmatrix} y \\ g \end{bmatrix}; \theta\right) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_y & \Sigma_{yg} \\ \Sigma_{gy} & K_\beta \end{bmatrix}\right), \tag{10}$$

where $\Sigma_{yg} = \Sigma_{gy}^T = WK_\beta$ and $\Sigma_y = WK_\beta W^T + \sigma^2 I$. It follows that the posterior distribution of g given y is also Gaussian, i.e.

$$p(g|y;\, \theta) = \mathcal{N}(Cy,\, P), \tag{11}$$

where

$$P = \left(\frac{W^T W}{\sigma^2} + K_\beta^{-1}\right)^{-1} \quad , \quad C = P\frac{W^T}{\sigma^2}. \tag{12}$$

The hyperparameter vector θ needs to be estimated from the available data, and we will address this point in the next section.

Given a value of θ, from (11), we find the impulse response estimator as the minimum mean squared error estimator (Anderson and Moore [1979])

$$\hat{g} = \mathbb{E}[g|y;\, \theta] = Cy. \tag{13}$$

4. EMPIRICAL BAYES ESTIMATES OF THE PARAMETERS

In this section we deal with the estimation of the hyperparameter vector θ. Exploiting the Bayesian framework introduced in the previous section, we adopt an Empirical Bayes approach (Maritz and Lwin [1989]) for this task. The hyperparameter vector is obtained by maximizing the marginal likelihood of the output, i.e.

$$\begin{aligned} \hat{\theta} &= \arg\max_\theta \log p(y;\theta) \\ &= \arg\min_\theta \log\det \Sigma_y + y^T \Sigma_y^{-1} y. \end{aligned} \tag{14}$$

4.1 Solution via the EM method

Problem (14) is non-convex and nonlinear, and involves $p+2$ decision variables. For its solution, below we propose a scheme based on the EM method. To this end, we introduce the complete-data log-likelihood

$$L(y,\, g;\theta) \triangleq \log p(y,\, g;\theta), \tag{15}$$

where g plays the role of *latent variable*. The EM method solves (14) by iteratively marginalizing out g from (15). This operation is performed by iterating the following steps:

(E-step) At the k-th iteration, using the estimate $\hat{\theta}^{(k)}$, compute

$$\mathcal{Q}(\theta,\, \hat{\theta}^{(k)}) \triangleq \mathbb{E}_{p(g;y,\,\hat{\theta}^{(k)})}\left[L(y,\, g;\theta)\right]; \tag{16}$$

(M-step) update the the estimate solving

$$\hat{\theta}^{(k+1)} = \arg\max_\theta \mathcal{Q}(\theta,\, \hat{\theta}^{(k+1)}). \tag{17}$$

Such a procedure converges to a maximum (not necessarily global) of (14) (see e.g. McLachlan and Krishnan [2007]).

Let us assume that the estimate $\hat{\theta}^{(k)}$ of the hyperparameter vector has been computed at the k-th iteration of the EM method. Using (12), we can compute the quantities $\hat{P}^{(k)}$ and $\hat{m}_g^{(k)}$, namely the posterior mean and variance of g given y. The following theorem illustrates how to compute $\hat{\theta}^{(k+1)}$.

Theorem 2. Assume that $\hat{\theta}^{(k)}$ is available. Then, the updated estimate

$$\hat{\theta}^{(k+1)} = \begin{bmatrix} \hat{c}^{(k+1)T} & \hat{\sigma}^{2,(k+1)} & \hat{\beta}^{(k+1)} \end{bmatrix} \tag{18}$$

is obtained by means of the following steps:

- The coefficients of the nonlinear block are given by

$$\hat{c}^{(k+1)} = (\hat{A}^{(k)})^{-1}\hat{b}^{(k)}, \tag{19}$$

where

$$\begin{aligned} \hat{A}^{(k)} &= F(u)^T \mathbf{R}^T \big(\hat{P}^{(k)} + \hat{m}_g^{(k)} \hat{m}_g^{(k)T}\big)\mathbf{R}F(u), \\ \hat{b}^{(k)} &= F(u)^T \mathbf{T}_N\big(\hat{m}_g^{(k)}\big) y, \end{aligned} \tag{20}$$

where $\mathbf{R} \in \mathbb{R}^{Nn \times N}$ is the (unique) matrix such that, for all $u \in \mathbb{R}^N$, we have

$$\mathbf{R}u = \mathrm{vec}\big(\mathbf{T}_n(u)\big); \tag{21}$$

- The noise variance is updated using

$$\begin{aligned} \hat{\sigma}^{2,(k+1)} = \frac{1}{N}\Big(&\|y - \hat{W}^{(k+1)}\hat{m}_g^{(k)}\|_2^2 \\ &+ \mathrm{Tr}\{\hat{W}^{(k+1)}\hat{P}^{(k)}\hat{W}^{(k+1)T}\}\Big) \end{aligned} \tag{22}$$

where $\hat{W}^{(k+1)} = \mathbf{T}_n\big(F(u)\hat{c}^{(k+1)}\big)$ results by plugging the new estimates $\hat{c}^{(k+1)}$ in (6);

- The updated Kernel shaping parameter is solution of

$$\hat{\beta}^{(k+1)} = \arg\max_\beta Q_\beta(\beta, \hat{\theta}^{(k)}), \tag{23}$$

where

$$Q_\beta(\beta, \hat{\theta}^{(k)}) \triangleq \log\det K_\beta + \mathrm{Tr}\Big[K_\beta^{-1}\big(\hat{P}^{(k)} + \hat{m}_g^{(k)}\hat{m}_g^{(k)T}\big)\Big]. \tag{24}$$

Therefore, the solution of (14) can be retrieved in a simple and quick way. In fact, Theorem 2 states that, given an estimate of the hyperparameter vector, the updated values of the coefficients of the nonlinear block are obtained solving a system of linear equations. Then, the new estimate of the noise variance can also be computed using a closed-form expression. Finally, the new value of the kernel shaping parameter is retrieved by solving a simple optimization problem. Although such a problem does not seem to admit a closed-form solution, we note that it is a one dimensional problem in the domain (0, 1]. Hence, it can be quickly solved by pointwise evaluation.

Below, we give our novel Kernel based method for the identification of Hammerstein systems.

Algorithm: Kernel-based Hammerstein System Identification

Input: $\{u_t\}_{t=0}^{N-1}$, $\{y_t\}_{t=1}^{N}$

Output: $\{\hat{g}\}_{t=1}^{n}$, $\hat{f}(\cdot)$

(1) Initialization: randomly set $\hat{\theta}^{(0)}$

(2) Repeat until convergence:

(a) **E-step:** update $\hat{P}^{(k)}$, $\hat{C}^{(k)}$ from (12) and $\hat{m}_g^{(k)}$ from (13);

(b) **M-step:** update the parameters:

- $\hat{c}^{(k+1)}$ from (19);
- $\hat{\sigma}^{(k+1)}$ from (22),
- $\hat{\beta}^{(k+1)}$ from (23)

(3) Compute $\{\hat{g}\}_{t=1}^{n}$ from (13) and $\hat{f}(\cdot) = \sum_{i=1}^{p} \hat{c}_i \phi_i(\cdot)$;

Remark 3. The choice of random initial is motivated by the fact that, after several numerical experiments we have noticed that the algorithm is capable of reaching the global maximum of (14) independently of the initial conditions.

5. NUMERICAL EXPERIMENTS

In order to assess the performance of the proposed Hammerstein system identification scheme, we run a set of Monte Carlo experiments. Specifically, we perform 8 different identification experiments, each consisting of 100 independent Monte Carlo runs. Depending on the experiment, we generate systems of order ν, where $\nu = 4$, 8, 10, 20. At each Monte Carlo run, a system is generated by picking ν random poles and zeros. The poles and zeros are located in the set $\{z \in \mathbb{C} \text{ s.t. } 0.4 \leq z \leq 0.93\}$. The input nonlinearity is chosen to be a polynomial of sixth order, so that $\phi_i(u) = u^{i-1}$, $i = 1, \ldots, 7$. The roots of the polynomial are in random locations within the interval $[-2, 2]$. The input to the system is white noise, with uniform distribution in the same interval. We generate $N = 500$ input/output samples for any Monte Carlo run. The output is corrupted by Gaussian white noise whose variance is chosen so to obtain a certain signal to noise ratio, according to

$$\sigma^2 = \frac{\text{Var}\{Wg\}}{\text{SNR}}, \tag{25}$$

where SNR is either 10 or 1, depending on the experiment. The features of the 8 experiments are summarized in Table 1. The goal of the experiments is to estimate $n = 100$ samples of the impulse response of the LTI system and the $p = 7$ coefficients of the nonlinear block. In order to obtain uniqueness in the decompositions, we impose $\|g\|_2 = 1$, and we assume the sign of the first nonzero element of g to be known.

Experiment #	1	2	3	4	5	6	7	8
LTI System Order	4	8	10	20	4	8	10	20
SNR	10	10	10	10	1	1	1	1

Table 1. Features of the 8 Monte Carlo experiments performed to test the proposed method.

We compare the following two estimators:

- **KB-H** This is the proposed kernel-based Hammerstein system identification method, which estimates the prior shaping parameter β, the nonlinear function and the noise variance through marginal likelihood maximization and the EM method. The convergence criterion for the EM method is $\|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}\|_2 < 10^{-3}$.
- **NLHW** This is the MATLAB function `nlhw` that uses the prediction error method to identify the linear block in the system (see Ljung et al. [2009] for details). To get the best performance from this method, we equip it with an oracle that knows the true order of the LTI system generating the measurements.

We use two metrics to evaluate the performance of the estimators. We consider the fitting score of the system impulse response

$$FIT_{g,i} = 1 - \frac{\|g_i - \hat{g}_i\|_2}{\|g_i - \bar{g}_i\|_2}, \tag{26}$$

where g_i is the system generated at the i-th run of each experiment, $\hat{g}_i$ its estimate and $\bar{g}_i$ its mean. A similar fitting score is defined for the nonlinearity, namely

$$FIT_{f,i} = 1 - \frac{\|f_i(u) - \hat{f}_i(u)\|_2}{\|f_i(u) - \overline{f_i(u)}\|_2}. \tag{27}$$

Figure 2 shows one Monte Carlo realization with LTI system order equal to 10 and SNR = 10, while Figure 3 shows the results of the outcomes of the 8 experiments. The box-plots compare the results of KB and NLHW for the considered model orders. We can see that, for low model orders, the estimator NLHW equipped with the true model order outperforms the proposed method. For higher model orders, however, the proposed method KB-H gives substantially better performance than NLHW. This because KB-H is not affected by the increasing complexity (model order) of the system, and the fitting score remains approximately constant. In addition, we can notice that the estimation of the nonlinear block computed with KB-H always provides a higher accuracy than NLHW.

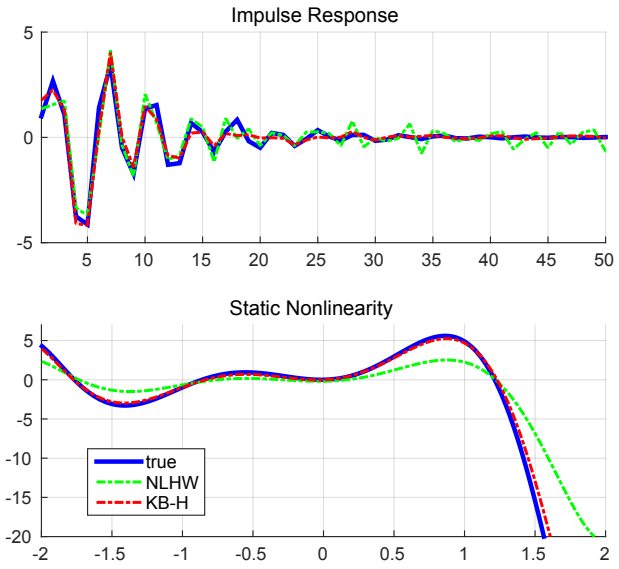

Fig. 2. Realizations of one Monte Carlo run with LTI system order 10 and SNR = 10.

6. CONCLUSIONS

In this work, we have proposed a novel kernel-based approach to the identification of Hammerstein dynamic

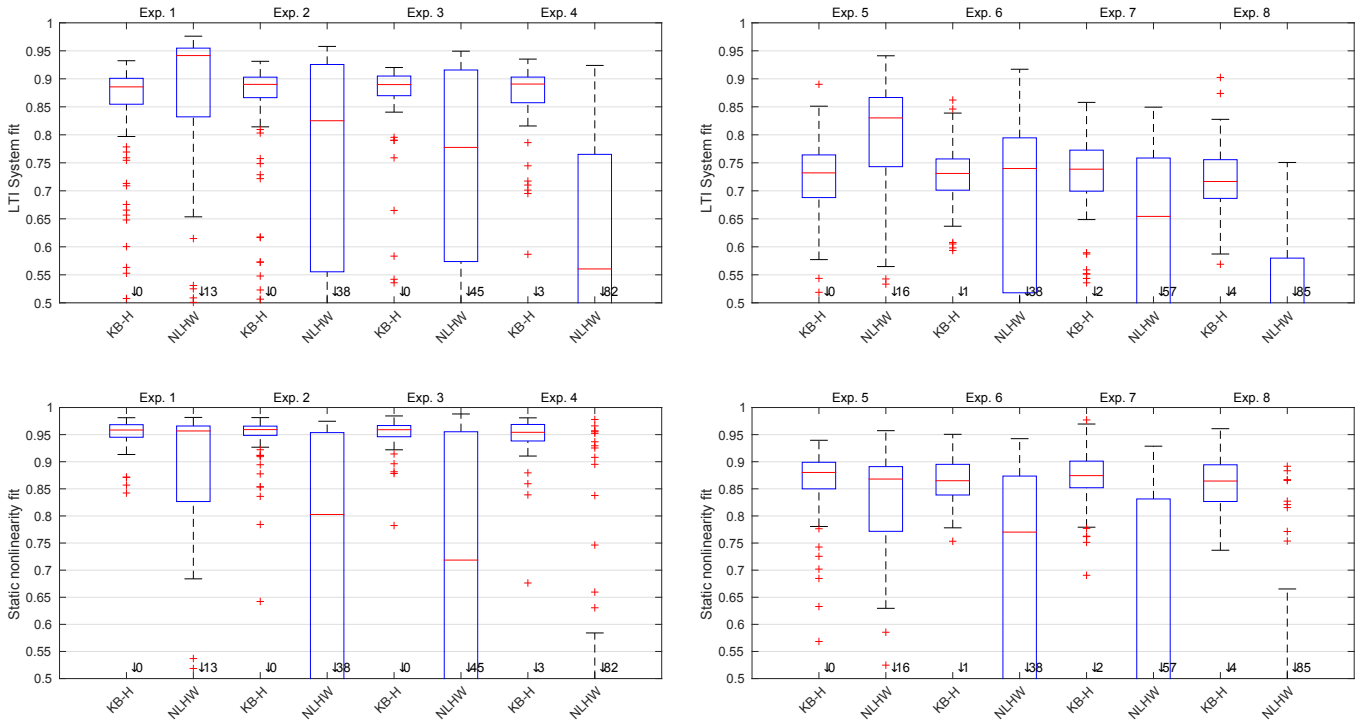

Fig. 3. Results of the 8 Monte Carlo experiments summarized in Table 1.

systems. To model the impulse response we have adopted a Gaussian regression approach and employed the stable spline kernel. The identification of the input nonlinearity, together with the kernel hyperparameter and the noise variance has been performed using an empirical Bayes approach. The related marginal likelihood maximization has been carried out resorting to the EM method. We have shown that this approach leads to an iterative scheme consisting of a set of simple update rules, which allow for fast computation. The effectiveness of the proposed method has been tested by means of several numerical experiments. When compared with standard state-of-the-art algorithms, the proposed method has shown a better fitting capacity in both the nonlinear block and the LTI system impulse response.

We are currently working on the extension of the algorithm to a wider class of system models. Furthermore, nonparametric descriptions of the nonlinear function will be considered in order to obtain a completely parameter-free identification method.

Appendix A. PROOF OF THEOREM 2

The proof runs along the same arguments as the proof of Theorem 1 in Bottegal et al. [2015] and is included here for the sake of self-completeness.

$$p(y,\, g;\, \theta) = p(y|g;\, \theta)p(g;\, \theta)\,. \tag{A.1}$$

Hence we can write the complete-data log-likelihood as

$$L(y,\, g;\, \theta) = \log p(y|g;\, \theta) + \log p(g;\, \theta) \tag{A.2}$$

and so

$$\begin{aligned} L(y,\, g;\, \theta) = &-\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left\|y - Wg\right\|^2 \\ &- \frac{1}{2}\log\det K_\beta - \frac{1}{2}g^T K_\beta^{-1} gl \\ = &-\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(y^T y + g^T W^T W g - 2y^T W g\right) \\ &- \frac{1}{2}\log\det K_\beta - \frac{1}{2}g^T K_\beta^{-1} g\,. \end{aligned}$$

We now proceed by taking the expectation of this expression with respect to the random variable $g|y;\, \hat\theta^{(k)}$. We obtain the following components

$$(a) : \mathbb{E}\left[-\frac{N}{2}\log\sigma^2\right] = -\frac{N}{2}\log\sigma^2$$

$$(b) : \mathbb{E}\left[-\frac{1}{2\sigma^2}y^T y\right] = -\frac{1}{2\sigma^2}y^T y$$

$$(c) : \mathbb{E}\left[-\frac{1}{2\sigma^2}g^T W^T W g\right] = \operatorname{Tr}\left[-\frac{1}{2\sigma^2}W^T W(\hat P^{(k)} + \hat m_g^{(k)}\hat m_g^{(k)T})\right]$$

$$(d) : \mathbb{E}\left[\frac{1}{\sigma^2}y^T W g\right] = \frac{1}{\sigma^2}y^T W \hat m_g^{(k)}$$

$$(e) : \mathbb{E}\left[-\frac{1}{2}\log\det K_\beta\right] = -\frac{1}{2}\log\det K_\beta$$

$$(f) : \mathbb{E}\left[-\frac{1}{2}g^T K_\beta^{-1} g\right] = -\frac{1}{2}\operatorname{Tr}\left[K_\beta^{-1}(\hat P^{(k)} + \hat m_g^{(k)}\hat m_g^{(k)T})\right]$$

It follows that $Q(\theta,\, \hat\theta^{(k)})$ is the summation of the elements obtained above. By inspecting the structure of $Q(\theta,\, \hat\theta^{(k)})$, it can be seen that such a function splits in two independent terms, namely

$$\mathcal{Q}(\theta,\, \hat\theta^{(k)}) = \mathcal{Q}_1(c,\, \sigma^2,\, \hat\theta^{(k)}) + \mathcal{Q}_\beta(\beta,\, \hat\theta^{(k)})\,, \tag{A.3}$$

where

$$\mathcal{Q}_1(c,\, \sigma^2,\, \hat\theta^{(k)}) = (a) + (b) + (c) + (d) \tag{A.4}$$

is function of c and σ^2, while

$$\mathcal{Q}_\beta(\beta,\, \hat\theta^{(k)}) = (e) + (f) \tag{A.5}$$

depends only on β and corresponds to (24). We now address the optimization of (A.4). To this end we write

$$\begin{aligned}\mathcal{Q}_1(c, \sigma^2, \hat{\theta}^{(k)}) &= \frac{1}{\sigma^2}\mathcal{Q}_c(c, \hat{\theta}^{(k)}) + \mathcal{Q}_{\sigma^2}(\sigma^2, \hat{\theta}^{(k)}) \qquad \text{(A.6)}\\ &= \frac{1}{\sigma^2}\Big(\mathrm{Tr}\left[-\frac{1}{2}W^T W(\hat{P}^{(k)} + \hat{m}_g^{(k)}\hat{m}_g^{kT})\right]\\ &\quad + y^T W\hat{m}_g^{(k)}\Big) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}y^T y\end{aligned}$$

This means that the optimization of $\mathcal{Q}_1$ can be carried out first with respect to c, optimizing only the term $\mathcal{Q}_c$, which is independent of σ^2 and can be written in a quadratic form

$$\mathcal{Q}_c(c, \hat{\theta}^{(k)}) = -\frac{1}{2}c^T\hat{A}^{(k)}c + \hat{b}^{(k)T}c\,. \qquad \text{(A.7)}$$

To this end, first note that, for all $v_1 \in \mathbb{R}^n$, $v_2 \in \mathbb{R}^m$:

$$\mathbf{T}_m(v_1)v_2 = \mathbf{T}_n(v_2)v_1\,. \qquad \text{(A.8)}$$

Recalling (21), we can write

$$\begin{aligned}&\mathrm{Tr}\left[W^T W(\hat{P}^{(k)}+\hat{m}_g^{(k)}\hat{m}_g^{(k)T})\right]\\ &= \mathrm{vec}(W)^T((\hat{P}^{(k)}+\hat{m}_g^{(k)}\hat{m}_g^{(k)T})\otimes I_N)\mathrm{vec}(W)\\ &= -\frac{1}{2}w^T\mathbf{R}^T\left((\hat{P}^{(k)} + \hat{m}_g^{(k)}\hat{m}_g^{(k)T})\otimes I_N\right)\mathbf{R}w\\ &= -\frac{1}{2}c^T F(u)^T\mathbf{R}^T\left((\hat{P}^{(k)} + \hat{m}_g^{(k)}\hat{m}_g^{(k)T})\otimes I_N\right)\mathbf{R}F(u)c\,,\end{aligned}$$

where the matrix in the middle corresponds to $\hat{A}^{(k)}$ defined in (20). For the linear term we find

$$y^T W\hat{g}^{(k)} = y^T\mathbf{T}_N(\hat{m}_g^{(k)})w = y^T\mathbf{T}_N(\hat{m}_g^{(k)})F(u)c\,, \quad \text{(A.9)}$$

so that the term $\hat{b}^{(k)T}$ in (20) is retrieved and the maximizer $\hat{c}^{(k+1)}$ is as in (19). Plugging back $\hat{c}^{(k+1)}$ into (A.4) and maximizing with respect to σ^2 we easily find $\hat{\sigma}^{2,(k+1)}$ corresponding to (22). This concludes the proof.

REFERENCES

Anderson, B.D.O. and Moore, J.B. (1979). *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, N.J., USA.

Bai, E. (1998). An optimal two-stage identification algorithm for Hammerstein–Wiener nonlinear systems. *Automatica*, 34(3), 333–338.

Bai, E., Cai, Z., Dudley-Javorosk, S., and Shields, R. (2009). Identification of a modified Wiener–Hammerstein system and its application in electrically stimulated paralyzed skeletal muscle modeling. *Automatica*, 45(3), 736–743.

Bai, E. and Li, D. (2004). Convergence of the iterative Hammerstein system identification algorithm. *IEEE Trans. on Automatic Control*, 49(11), 1929–1940.

Bottegal, G. and Pillonetto, G. (2013). Regularized spectrum estimation using stable spline kernels. *Automatica*, 49(11), 3199–3209.

Bottegal, G., Risuleo, R.S., and Hjalmarsson, H. (2015). Blind system identification using kernel-based methods. In *Proc. of the 16th IFAC Symp. on System Identification (submitted)*.

Chen, T., Ohlsson, H., and Ljung, L. (2012). On the estimation of transfer functions, regularizations and Gaussian processes - revisited. *Automatica*, 48(8), 1525–1535.

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

Goethals, I., Pelckmans, K., Suykens, J., and De Moor, B. (2005). Subspace identification of Hammerstein systems using least squares support vector machines. *IEEE Trans. on Automatic Control*, 50(10), 1509–1519.

Greblicki, W. (2002). Stochastic approximation in nonparametric identification of Hammerstein systems. *IEEE Trans. on Automatic Control*, 47(11), 1800–1810.

Greblicki, W. and Pawlak, M. (1986). Identification of discrete Hammerstein systems using kernel regression estimates. *IEEE Trans. on Automatic Control*, 31(1), 74–77.

Han, Y. and De Callafon, R.A. (2012). Hammerstein system identification using nuclear norm minimization. *Automatica*, 48(9), 2189–2193.

Hunter, I. and Korenberg, M. (1986). The identification of nonlinear biological systems: Wiener and Hammerstein cascade models. *Biological cybernetics*, 55(2-3), 135–144.

Ljung, L. (1999). *System Identification, Theory for the User*. Prentice Hall.

Ljung, L., Singh, R., Zhang, Q., Lindskog, P., and Ioudiski, A. (2009). Developments in the MathWorks System Identification Toolbox. In *Proc. of the 15th IFAC Symp. on System Identification*.

Maritz, J.S. and Lwin, T. (1989). *Empirical Bayes Method*. Chapman and Hall.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Narendra, K. and Gallman, P. (1966). An iterative method for the identification of nonlinear systems using a Hammerstein model. *IEEE Trans. on Automatic Control*, 11(3), 546–550.

Pillonetto, G. and Chiuso, A. (2014). Tuning complexity in kernel-based linear system identification: The robustness of the marginal likelihood estimator. In *Proc. of the European Control Conference (ECC)*, 2386–2391.

Pillonetto, G., Chiuso, A., and De Nicolao, G. (2011). Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica*, 47(2), 291–305.

Pillonetto, G. and De Nicolao, G. (2010). A new kernel-based approach for linear system identification. *Automatica*, 46(1), 81–93.

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., and Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3), 657–682.

Rangan, S., Wolodkin, G., and Poolla, K. (1995). New results for Hammerstein system identification. In *Proc. of the 34th IEEE Conference on Decision and Control*.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Schoukens, M., Pintelon, R., and Rolain, Y. (2011). Parametric identification of parallel Hammerstein systems. *IEEE Trans. on Instrumentation and Measurement*, 60(12), 3931–3938.

Westwick, D.T. and Kearney, R.E. (2001). Separable least squares identification of nonlinear Hammerstein models: Application to stretch reflex dynamics. *Annals of Biomedical Engineering*, 29(8), 707–718.